International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

Comparison of Various Tools for Data Mining

Mrs. Parminder Kaur[1.]
Research Scholar, Sri Venkateshwara University,
Gajraula, U.P.

Dr. Qamar Parvez Rana[2]
Course Director, Jamia Hamdard,
New Delhi

Abstract :- **The concept of Data Mining has emerged to meet th e requirement of quick and accurate information support for decision making process. The idea is to extract the data from the database for the operational use. Data Mining is analysis of data to identify relationship between different data elements or entities. The process of data mining can also involve correlation or association between two or more data elements, entities or events. They allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.In this research we have focused on comparison of various Data Mining tools which are helpful and marked as the important field of data mining Technologies. As we are aware that many national and Multinational companies and small or large organizations are operated in different places of the different countries. Each operation may generate big data or unstructured. This type of big data is available in the form of bytes which has drastically changed in various areas. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields.**

Keywords are: Data Mining, Decision Making, Entities, Big Data

I. INTRODUCTION

Data mining tools are component and method that allow end-user to mine useful information from unstructured data. Some common uses of data mining are education, pattern finding, research, marketing and fraud detection. During this era, IT emerged to support methods for developing data mining tools within the data mining industry, some standards has been established to define the rules for "how to use data mining tool?". Special group on knowledge discovery and data mining determines what processes are used and how they are used?. Data Mining software use mathematical and statistical techniques to analyze data to reveal hidden pattern and trends. Data Mining is a collaborative tool which comprises of database systems, machine learning, statistic, visualization information science and other discipline. The Data Mining tools include a neutral network component that facilitates classification, prediction and profiling. Data Mining can often help in making prediction about future events. Data Mining can make use of the valuable historical data and provide valuable input.

a. Changes In Data Access

The first change occurred in the area of basic data collection. Before user made the transition from ledgers and other paper-based records to computer-based systems, one had to wait for another to put the pieces together to know how well the organization was performing or how current performance periods compared with previous periods. As organization started collecting and saving basic data in computers, they were able to start answering detailed questions quicker and with more ease.

b.Changes In Data Mining Techniques

Where there has been greater empowerment and integration, particularly over the past 30 years also have impacted data mining techniques. The introduction of microcomputers and networks, and the evolution of middleware, protocols, and other methodologies that enable data to be moved seamlessly among programs and other machines, allowed companies to link certain data questions together

Most Data Mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools.

- *Traditional Data Mining Tool:-* Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using Online Analytical Processing or a similar technology.

- *Dashboards:-* Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen often in the form of a chart or table enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

- *Text-mining Tools.* The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an

easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

Steps for Mining the Data

With enterprises multi-dimensional geographic locations, multi-database mining is becoming important for effective and informed decision making. The following data mining techniques will help you optimize your mining:

Step 1: Handling of unstructured data

Unstructured data affects accuracy and effective data mining. The following techniques are effective for working with such kind of data.

1. The ISOM-DH model useful to estimate the missing data and visualize the handled high-dimensional data by using independent component analysis (ICA) and self-organizing maps (SOM).

Another technique is based on the strategies built using parametric and non-parametric imputation methods or Genetic algorithms have to be applied to develop a framework.

2. Network approaches based on multi-task learning used for pattern classification, with missing inputs, can be compared with representative procedures used for handling incomplete data on two well-known data sets.

Step 2: Provide effective data mining algorithms

Expertise is required for the use, implementation, maintenance, and performance-effective data mining application. These technique may help:

1. Execution of data mining algorithms.
2. Grid-enable data mining applications without any intervention on the application side.
3. Opt for scalable data mining.
4. Remove barriers

Step 3: Mining of big-databases

Use combine set architectural with database systems. Such data mining techniques could include:

1. Encapsulation of the data mining algorithm.
2. Caching the data, then mining.
3. Tight-coupling with user-defined functions.
4. SQL implementations for DBMS.

Step 4: Handling of data types

It's difficult to develop a system for interactive mining of multiple-level knowledge in large databases and data warehouses. This requires tight coupling of online analytical processing with a wide spectrum of data mining functions including characterization, association, classification, prediction, and clustering. The system should facilitate query-based, interactive mining of multidimensional databases by implementing a set of advance data mining includes:

- Multidimensional analysis
- Data mining refined knowledge
- Meta-mining, and data and knowledge visualization
- Assessing data mining results
- Analyzing graph databases
- sub-graph histogram representation

Step 5: Handling Heterogeneous environment

Heterogeneous database systems are popular one in information industry in 2011. Data warehouses must support data extraction from multiple databases to keep up with the trend.

II. LITERATURE SURVEY-

Data mining is not all about the tools or database software that you are using. You can perform data mining with comparatively modest database systems and simple tools, including creating and writing your own, or using off the shelf software packages. Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques.

For example, IBM SPSS, which has its roots in statistical and survey analysis, can build effective predictive models by looking at past trends and building accurate forecasts. IBM InfoSphere Warehouse provides data sourcing, preprocessing, mining, and analysis information in a single package, which allows you to take information from the source database straight to the final report output.

Now an entirely new range of tools and systems available, including combined data storage and processing systems. You can mine data with a various different data sets, including, traditional SQL databases, raw text data, key/value stores, and document databases. Clustered databases, such as Hadoop, Cassandra, CouchDB, and Couchbase Server, store and provide access to data in such a way that it does not match the traditional table structure. With Hadoop's entirely raw data processing it can be complex to identify and extract the content before you start to process and correlate the it.

Besides these tools, other applications and programs may be used for data mining purposes. For instance, audit interrogation tools can be used to highlight fraud, data anomalies, and patterns. An example of this has been published by the United Kingdom's Treasury office in the *2002–2003 Fraud Report: Anti-fraud Advice and Guidance*, which discusses how to discover fraud using an audit interrogation tool. Additional examples of using audit interrogation tools to identify fraud are found in David G. Coderre's 1999 book, *Fraud Detection*.

In addition, internal auditors can use spreadsheets to undertake simple data mining exercises or to produce summary tables. Some of the desktop, notebook, and server computers that run operating systems such as Windows, Linux, and Macintosh can be imported directly into Microsoft Excel. Using pivotal tables in the spreadsheet,

auditors can review complex data in a simplified format and drill down where necessary to find the underlining assumptions or information.

When evaluating data mining strategies, companies may decide to acquire several tools for specific purposes, rather than purchasing one tool that meets all needs. Although acquiring several tools is not a mainstream approach, a company may choose to do so if, for example, it installs a dashboard to keep managers informed on business matters, a full data-mining suite to capture and build data for its marketing and sales arms, and an interrogation tool so auditors can identify fraud activity.

Data Mining Research

It is a still developing technology. The fact shows that data is growing at a very rapid rate, but many of data has once been stored and have never been used. This data is collected from different sources , if it is processed properly it can provide immense hidden knowledge, which can be used for future used. So, there is an eminent need for developing proper mechanism of processing these abandon data and extracting knowledge. The common issue for all data mining application and techniques includes the detection, interpretation and prediction of qualitative and quantitative pattern in data. To achieve such goal, one should employ appropriate machinery learning, artificial intelligence, statistics, and query processing.

Evolution of Data Mining

Progress	Technology	Features	Review
Data gathering(1960)	Tapes, computer	Use of static data	Problem in large storage
Data process(1980)	RDBMS,ODBC, SQL	Dynamic data at record level	Easy access
Data warehousing & Decision Making Support(1990)	OLAP, Multi-Dimensional	Dynamic data at multiple level	Easy to understand
Data Mining (Running)	Advanced Algorithm, Big database	Proactive information	Successful in various field

Some free data Mining Tool available for Users:

Data Mining Product	Main Features
SAS Enterprise Miner	• It comes from statistics; • Easy to use graphical interface; • Rich set of algorithms including algorithms for data mining: decision trees, neural networks, regression, association, etc. • Ability to analyze text
SPSS	• It comes from statistics; • Includes among others, decision tree data mining algorithms (Answer Tree); • Allows users to perform data cleansing and data transformation.
IBM Intelligent Miner	• It comes from the database field; • Features advanced visualization tools and data presentation; • Compatible with PMML language (Predictive Modeling Markup Language) for exporting the data models found; • Can work with DB2 database management system.
Microsoft SQL Server 2005	• It comes from the database field of activity; • Among others, it offers algorithms for decision trees, prediction and clustering; • Implements the OLE DB standard for Data Mining, which defines a data mining language similar to SQL; • Features an easy to use API interface for facilitating the integration of data mining facilities into the user applications
Microsoft SQL Server 2005	• It comes from the database field of activity; • Among others, it offers algorithms for decision trees, prediction and clustering; • Implements the OLE DB standard for Data Mining, which defines a data mining language similar to SQL; • Features an easy to use API interface for facilitating the integration of data mining facilities into the user applications
Oracle Data Mining (from Oracle 10g)	• It comes from the database; • It started with algorithms such as association and Naive Bayes (version 9i) and with the 10g version it includes a great variety of algorithms; • Integrates Java Data Mining API, a Java package for including the data mining facilities into the user's applications.
Angoss Knowledge STUDIO	• Presents algorithms for building decision trees, cluster analysis (grouping) and predictive models; • Allows users to exploit data in different forms; • Offers powerful visualization tools of the results that make it very user friendly; • It is compatible with other databases such as Microsoft SQL Server, and can interact with them at datamining level.
KXEN	• Has algorithms for regression, time series analysis, classification, etc. • Implements procedures for working with OLAP data cubes; • It can retrieve data from spreadsheet programs like Microsoft Excel.

Some Stand-alone Application

1.TANAGRA is a DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. It is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms.

TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license. The main purpose of Tanagra project is

to give researchers and students an easy-to-use **data mining software**, allowing them to easily add their own data mining methods, to compare their performances and in direction of novice developers, consists in **diffusing a possible methodology for building this kind of software**. They should take advantage of free access to source code, to look how this sort of software is built, the problems to avoid, the main steps of the project, and which tools and code libraries to use for. In this way, Tanagra can be considered as a pedagogical tool for learning programming techniques.

2.RapidMiner is unquestionably the world-leading open-source system for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. Thousands of applications of RapidMiner in more than 40 countries give their users a competitive edge.

3.RapidMiner as a powerful engine for analytical ETL, data analysis, and predictive reporting, the new business analytics server. RapidAnalytics is the key product for all business critical data analysis tasks and a milestone for business analytics.

4.Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes

5.PSPP is a program for statistical analysis of sampled data. It has a graphical user interface and conventional command-line interface. It is written in C, uses GNU Scientific Library for its mathematical routines, and plotutils for generating graphs. It is a Free replacement for the proprietary program SPSS (from IBM) predict with confidence what will happen next so that you can make smarter decisions, solve problems and improve outcomes.

6.KNIME is a user-friendly graphical workbench for the entire analysis process: data access, data transformation, initial investigation, powerful predictive analytics, visualisation and reporting. The open integration platform provides over 1000 modules (nodes)

7.Orange is an Open source data visualization and analysis for novice and experts. Data mining through visual programming or Python scripting. Components for machine learning. Add-ons for bioinformatics and text mining. Packed with features for data analytics.

8.Apache Mahout is an Apache project to produce free implementations of distributed or otherwise scalable machine learning algorithms on the Hadoop platform.
Currently Mahout supports mainly four use cases: Recommendation mining takes users' behavior and from that tries to find items users might like. Clustering takes e.g. text documents and groups them into groups of topically related documents. Classification learns from exisiting categorized documents what documents of a specific category look like and is able to assign unlabelled documents to the (hopefully) correct category. Frequent itemset mining takes a set of item groups (terms in a query session, shopping cart content) and identifies, which individual items usually appear together.

9.jHepWork (or "jWork") is an environment for scientific computation, data analysis and data visualization designed for scientists, engineers and students. The program incorporates many open-source software packages into a coherent interface using the concept of scripting, rather than only-GUI or macro-based concept.
jHepWork can be used everywhere where an analysis of large numerical data volumes, data mining, statistical analysis and mathematics are essential (natural sciences, engineering, modeling and analysis of financial markets).

10.Rattle
Rattle (the R Analytical Tool To Learn Easily) presents statistical and visual summaries of data, transforms data into forms that can be readily modelled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets.
It is a free and open source data mining toolkit written in the statistical language R using the Gnome graphical interface. It runs under GNU/Linux, Macintosh OS X, and MS/Windows. Rattle is being used in business, government, research and for teaching data mining in Australia and internationally.

Data Mining Trends.
Here is the list of trends in data mining that reflects pursuit of the challenges such as construction of integrated and interactive data mining environments, design of data mining languages:

- Application Exploration
- Scalable and Interactive data mining methods
- Integration of data mining with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language
- Visual Data Mining
- New methods for mining complex types of data
- Biological data mining
- Data mining and software engineering
- Web mining
- Distributed Data mining
- Real time data mining
- Multi Database data mining
- privacy protection and Information Security in data mining

III. SCOPE OF RESEARCH

a. Provide powerful techniques for better interpretation of these data that exceeds the human's ability for comprehension and making decision in a better way.
b. Reveal the best tools for dealing with the task that helps in decision making, this research has conducted a comparative study between a number of some of the data mining and knowledge discovery tools and software packages.
c. Data Mining is fundamentally application-oriented area motivated by business and scientific needs to make sense of mountains of data.
d. Support or do some task(s) by human beings in an organizational environment both having their desires related to DMS.
e. The majority of data mining techniques can deal with different data types. The research provides an explanation and comparison study on some of the most common data mining tool in use today in day to day life and business predictions. Though there are a number of other techniques and many variations of the methods described, one of the techniques from the mentioned group is almost always used in real world deployments of data mining systems.

IV. REQUIREMENTS OF THE RESEARCH WORK

Research Design :

- Theoretical approach: theory creating
  - Hypothesis, etc.
- Constructive approach
  - Comparison of DM tool
- Theoretical approach: theory testing and evaluation
  - Artificial, benchmark, real-world data

Data Collection strategy (Primary & Secondary methods): The study in the research paper will be interdisciplinary the data will be drawn from Secondary as well as from Primary source. The Secondary data will comprise of various references which already exist in the published form such as research papers, articles and books related to inflation. The Primary sources that will constitute the major part of the study will be based on my paper present in seminar on topic "Data Mining For Business Intelligence by using Web-Focus".

V. CONCLUSION

In this paper we briefly reviewed the various data mining tools and their applications from its inception to the future. This review puts focus evolution and trends of data mining. This paper provides a new perspective of a researcher regarding applications of data mining in social welfare.

VI. REFERENCES/BIBLIOGRAPHY:

[1] Pareek Astha, IIS University, Efficient data mining techniques for application in elementary educational system with special reference to out of school children, 2013.
[2] Berry, M. J. A., & Linnof, G, Data mining Techniques, New York: Wiley, (1997).
[3] Bouckaert, Remco R., Frank, Eibe, Hall, Mark A., Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project". Journal of Machine Learning Research 11: 2533–2541. "the original title, "Practical machine learning", was changed ... The term "data mining" was [added] primarily for marketing reasons."
[4] D. Andersson, H Fries: Data Mining Maturity. A Quantitative Study of Large Companies in Sweden, Jonkoping University, Master's Thesis in Informatics, 2008;
[5] Data Mining: Parctical Machine learning tools and techniques, 3rd Edition(January 20, 2011).
[6]Searchbusinessintelligence.techtarget.in/tip/5-data-mining-techniques-for-optimal-results
[7] Ricco RAKOTOMALALA, "TANAGRA: a free software for research and academic purposes", in Proceedings of EGC'2005, RNTI-E-3, vol. 2, pp.697-702, 2005. (in French)
[8] Heikki, Mannila, "Data mining: machine learning, statistics, and databases", Statistics and Scientific Data Management , pp. 2-9. 1996.
[9]Fayadd, U., Piatesky -Shapiro, G., and Smyth, P.From Data Mining To Knowledge Discovery in Databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–26256097–6. MIT1996.
[10]Piatetsky-Shapiro, Gregory, "The Data-Mining Industry Coming of Age", in IEEE Intelligent Systems, vol. 14, issue 6, Nov 1999. Doi.10.1109/5254.809566
[11]Salmin, Sultana et al., "Ubiquitous Secretary: A Ubiquitous Computing Application Based on Web Services Architecture", International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 4, October, 2009
[12]Hsu J., "Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century", in The Proceedings of the 19th Annual Conference for Information Systems Educators (ISECON 2002) ISSN: 1542-7382.
[13]Shonali Krishnaswamy, "Towards Situation awareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application", Proceedings of Conference on Intelligent Vehicles and Road Infrastructure 2005, pages-16, 17.
[14]S. Mitra, S. K. Pal, and P. Mitra. "Data mining in soft computing framework: A survey", IEEE Trans. Neural Networks, vol. 13, pp. 3 -14., 2006[16].