SURVEY ON CURRENT TREND AND TOOLS OF DATAMINING

R. Preethi[1] *Dr.S. Sheeja*[2]

ABSTRACT

The paper surveys different aspects of datamining tools used for research. Data mining helps to gain knowledge from big databases, data warehouses, and data marts realms. Different and current areas of datamining also discussed. This paper also includes the popular and critical features of every tool listed here. Datamining is an essential and evolving research area and used by the biologists to statisticians and computer scientists as well.

Keywords: *Data Mining, KDD, Data Mining Tools*

1. INTRODUCTION

Data mining can be described as the collection from experimental results of unique patterns and patterns or as a method used to process data. We know Data mining as the exploration of information. Removal, or "MINING," basically means information from large amounts of data. Data mining is used because of the explosive growth of data, i.e., from tera bytes to peta bytes. We drown in data but the thirst for information! Data mining's alternate names are data archeology, data dredging, data harvesting, data analytics, etc. Data mining techniques are used to identify the patterns hidden or new to store the data. We know that data mining can take advantage of every sector, such as industry, agriculture, marketing, etc. Data mining techniques are various, such as clustering, classification, etc. Data mining methods and techniques can be applied to data to build a new ecosystem to improve existing data output and help create new data predictions. [1]

Data Mining is the method of analyzing data from different perspectives and turning it into useful information. It often seeks interesting, non-obvious information from a broad data set. It is an knowledge discovery method that helps us to interpret the quality of the data in a specific unsuspected manner. Data mining involves trying to collect, transferring and loading data into the data warehouse network, storing and managing data in a database design system, providing business analysts and IT practitioners with access to data, analyzing data using software applications, presenting data in a graphical format, such as a chart or map. [2]

2. DATAMINING

The field of analysis and knowledge discovery refers to a new, important research area that includes applications in critical technology, engineering, medicine, business and learning. Data mining tries to establish and maintain essential processes of induction that enable the extraction of useful information and knowledge from large quantities of data[3]. Data mining uses statistical analysis to identify data patterns. Popularly known as knowledge discovery, it is the non-trivial extraction of data from secret, previously unknown and potentially useful knowledge in databases. Although data mining and information discovery are often seen as synonyms in databases, data processing is part of the information exploration phase. Some businesses integrate data mining, such as analytics, pattern recognition, and other necessary tools. Data mining can be used to detect patterns and connections which would be difficult to find otherwise. This method is normal for many businesses, as it lets them understand their customers better while making sensible marketing decisions[5]. In short, data mining may be tantamount to finding a haystack. We live in a golden age of knowledge and the greatest challenge is not only obtaining information, but digging through it to find connections and evidence that have not been found before.

Around the world, there are hundreds of libraries, data marts,

[1]Research Scholar, Department of Computer Science
Karpagam Academy of Higher Education, Coimbatore.

[1]Professor, Department of Computer Science
Karpagam Academy of Higher Education, Coimbatore.

data warehouses. If the data are not analyzed to identify the delightful patterns, then the data would become the tombs of data. Data miners are scanning data in the sea for the pearl. A data-mining program can create many models. A tiny fraction of the models is usually entertaining. The fascinating mean here is accessible, accurate, and novel. However, the exciting secret trends in the sea of data can be almost impossible to detect without the help of data mining software. Data mining is carried out in seven phases. These include data cleaning, data integration, data collection, data transformation, data mining, presentation of information, and pattern evolution [6].The database technology had advanced from rudimentary file processing to the development of software and applications for data mining .The data can be obtained from various apps like science and engineering, management, corporate houses, government administration, and environmental control. From spatial, time-related, text, biological, digital, web, and legacy databases, interesting data patterns can be mined. Data mining helps make decision-making simpler. The job of data mining involves the discovery of idea definitions, correlation, grouping, prediction, clustering, trend analysis, analysis of variations, and analysis of similarities. Data mining in extensive data provides researchers and developers with different demands and challenges.

3. TYPES OF TOOLS

The marketplace offers different types of data hashing algorithm, each with its benefits and drawbacks. Most data mining tools can be categorized into one of three sections: known data mining tools, dashboards and text analysis instruments.

3.1 Traditional Data Mining Tools:

Traditional data mining systems, using many complex techniques and techniques, help organizations develop data habits and trends. Several of these tools are designed to track data and view models on the web, while others gather knowledge that occurs independently of a database. Most are used in both Windows and UNIX models, while several are specialized only in one file system. Additionally, although some may concentrate on one form of database, others may handle any data using advanced analytics storage or similar systems.

3.2 Dashboards:

Revised to monitor server data in computers, dashboards show data adjustments as well as on-screen notifications — mostly in the form of a chart or table — that enable the user to see how the business is doing. statistical information may also be comparative, enabling users to see how things have changed (e.g., revenue growth from the same period last year). This approach makes dashboards simple to get and appealing to executives wanting a performance analysis of the product.

3.3 Text-mining Tools:

Because of the ability to gather data from multiple text sources — including Microsoft Word and Acrobat PDF documents to plain text files — the third form of data mining program is often called a text processing tool; for examples. These tools locate information and convert the processed data into a format that is consistent with the data base, making it simple and convenient for users to access data while having to open different applications. Scanned material can be disorganized or ordered (i.e. information is scattered almost randomly across the document, including emails, web pages, audio and video data) Capturing these observations can provide a lot of knowledge for companies to recognize trends, ideas and attitudes. [7]

4. FAMOUS TOOLS OF DATA MINING

Here we mentioned and described the most popular data mining tools and the top 15 currently on the market.

4.1. Rapid Miner

This tool consists of Java programming language and provides graduate-level analyses via its template-based

layout. Users hardly have to do any coding. Apart from data mining activities, Rapid Miner can handle various tasks such as statistical modeling, predictive analytics, and visualization. Rapid- Miner provides WEKA and R script training schemes, models, and algorithms that make it more efficient. This open-source is licensed under the open-source license AGPL and is available for download from Source Forge. It is one of the best technologies in business analytics. All the activities related to data mining are bundled into one package.

4.2. Orange

Orange, a Scripting-based, versatile, and open-source tool used by data mining users to extract information. It has strong visual programming attached to it and Python scripting.

When incorporating add ons, it can be used for machine learning as well as for bio-informatics and text mining. It is filled with data analytics tools. Orange has specialist bio-informatics add-ons, such as Bioorange.

4.3. WEKA

Machine learning algorithms developed at Waikato University in New Zealand also recognized as Waikato Area. This is ideally suited for the data processing and mathematical modeling. That includes algorithms and techniques for machine learning visualization. Weka was initially built in a non-Java version to analyze the agricultural data. The Java version was later developed, and became a powerful tool for various applications in data mining, such as statistical modeling and data analysis. Under the Gnu lesser general public license, this software is free, which is a significant advantage over Rapid Miner. Since it is available under the GNU General Public License, this is a significant advantage over its competitors such as Rapid Miner. Users can adapt it to their needs. Weka finances the majority of the data mining work. They're grouping, clustering, regression, extraction of elements, visualization, etc. Its user interface makes the data mining process a better-sophisticated tool. So, Weka has become one of the most popular open source software for data mining. Weka has a GUI which allows easy access to all of its features. Weka facilitates critical data mining which includes data extraction, examination, interpretation, correlation, etc. This implies that the knowledge is in flat file form. Weka can provide links to SQL databases via a server connection, and further evaluate the data / scores returned by the query.

4.4. KNIME

KNIME is capable of performing three main tasks in data preprocessing. They are extraction,
Transformation, and loading. The data processing is done by allowing the assembly of nodes. It is an integration platform with robust data analytics and reporting. KNIME used a modular data pipelining concept for machine learning, and information is used for business intelligence as well as financial datamining. KNIME is easily extendible and can be added as a plug-in for specific jobs. This open-sourceis also written in Java and based on Eclipse. The core version consists of various data integration modules. Its research area includes not only pharmaceutical research but also business data, financial intelligence, and CRM customerdata.

4.5. R-Programming

Project R, which is a GNU project, is written in C, FORTRAN, and R Language. R language is used for writing lots of modules of the software itself. R programming software is free, and it is also used for statistical computing and graphics. Data miners used R for developing statistical packages and analyzing the data. In recent years the popularity of R had increased because of its ease of use and extensibility. R provides different analytical techniques that include linear and nonlinear modeling, data mining processes, i.e., classification, clustering, time series analysis, and others.

4.6. Sisense

Sisense is extremely useful and the BI software is ideally

adapted for business reporting purposes. The company of the same name is created by the 'Sisense.' It has a fantastic ability for small-scale / large-scale organizations to handle and process data. It allows the combination of data from multiple sources to build a common database and also extends data to generate rich reports that are exchanged with units for reporting. Sisense has been called the best BI software in 2016 and still holds a strong position. The Sisense creates highly visual files. It is specifically designed for the non-technical users. It lets both drag & drop as well as widgets. In the basis of a company's strategy to create documents in the form of graphs, line charts, diagrams, etc., different widgets can be chosen. Reports can be further dug by simply pressing on the information and reading regular data.

4.7. SQL Server Data Tools

SSDT is a hierarchical, concise architecture covering all development modes for the Visual Studio IDE dataset. BIDS was intel's former data management tool, and Software Company for data processing. Using SSDT transactions, programmers may use the SQL creation technology to construct, manage, test and refactor data sources. A user can either work with an application server, or work directly with a special database, including on-site or off-site facilities. People can use visual studio software to build databases such as IntelliSense, code navigation tools, and programmer support through C #, basic visual, etc. SSDT allows Table Developer to construct new tables and review tables in similar databases and similar databases as well. BIDS replaced the SSDT BI, which did not follow Visual Studio2010, and substituted BIDS.es.

4.8. Apache Mahout

Apache Mahout is a technology created by the Apache Foundation that supports computer vision algorithms as their primary objective. This targets the clustering, sorting, and sorting of cooperative data. Mahout is written in Java programming language and offers JAVA libraries, such as linear algebra and statistics, for mathematical operations. Mahout keeps expanding as the calculations implemented within Apache Mahout continue to grow. Mahout formulas often impose a degree above Hadoop by mapping parameters / raising them. In short, Mahout has major features Extensible selection of sample, pre-made algorithms, a creative math framework, and mathematical calculations for enhancement of GPU results.

4.9. Oracle Data Mining

Oracle data science, an aspect of Oracle Advanced Analytics, offers excellent data mining techniques for data detection, prediction, correlation and advanced analysis that allow researchers to evaluate trends, make smarter choices, draw best clients, recognize cross-selling opportunities and prevent corruption. Within ODM, the models developed to maximize the Oracle database's possible advantages. SQL's data mining function will dig out data from database tables, views, and data types. An updated version of the Oracle SQL Server is the Oracle data miner app. It provides users with the facility to drag & drop data directly into the database, allowing users to analyze more.

4.10. Rattle

The rattle is a GUI-focused, data mining application that uses the programming language for R stats. By giving significant usability in data mining, Rattle shows the statistical strength of R. Although Rattle does have a robust and well-developed user experience, and a built-in log code tab that creates duplicate content for any GUI action. Rattle has script-checking external rights, use it for unique purposes, and expand the file unrestrictedly.

4.11. DataMelt

Data Melt is a computing and visualization platform, also known as DMelt, that supports integrated statistical testing and representation methods. It's built mainly for artists, researchers & graduates. DMelt is an application published in JAVA for multi-platforms.It can run on any JVM-compatible (Java Virtual Machine) device. The

collections of Science & Math are included. Scientific libraries: Draw 2D/3D plots. Curve matching libraries, architectures, etc. To produce random data. DataMelt may be used to analyze large amounts of data, data mining, and measurement. It is commonly used in the research, natural sciences & engineering of the financial sector.

4.12. IBM Cognos

IBM Cognos BI is a clinical variable run by the IBM for monitoring, predictive analysis, rating carding, etc. It includes sub-components which meet the criteria of Cognos Link, Query Studio, Document Studio, Processing Studio, Event Studio and Advanced Workflow. Cognos Connection: An online scoreboard / studies page for collecting and summing up information. Query Studio: Includes queries regarding data formatting & schematic creation. Studio Report: The generation of governing papers. Studio Analysis: Recognizing & identifying trends for the management of large volumes of data. Event Studio: Notification system to keep the activities synchronized. Workspace Advanced: Safe platform for creating custom &user-friendly reports.

4.13. IBM SPSS Modeler

IBM SPSS is an IBM-owned configuration tool used for machine learning & data analysis to create collective models. It was first produced by SPSS Inc. and later acquired by IBM. SPSS Modeler has a visualization framework that lets users deal with data mining methods without scripting. This removes the unwanted uncertainties encountered during data upgrades and makes predictive models simple to use. IBM SPSS arrives in two versions, focused on IBM SPSS Modeler Professional and IBM SPSS Modeler Premium capabilities, which include external text analytics, entity data analysis, etc.

4.14. SAS Data Mining

SAS is an analytics & data management platform developed by SAS Organization. SAS can collect, alter; optimize reports and figures from different sources. It supplies the non-technical users with a graphical UI. SAS data miner allows the user to evaluate big data and offers reliable information for effective decision making. SAS has a highly scalable decentralized design for storing the data. It is particularly suited for data mining, text quarrying & optimizing.

4.15. Teradata

Teradata is often termed the Teradata index. It is an industrial data warehouse that includes software for information management along with machine learning technology. It needed for automated analysis. Teradata is used for insight into enterprise data such as sales, product placement, customer needs, etc. It could also discriminate across ' hot ' and ' cold ' content; meaningless widely used data is placed in a slow processing division. Teradata works on the ' sharing nothing ' paradigm as it has its capabilities for storage and computation.

4.16. Board

The Committee is often concerned with the Project Toolkit. It's knowledge of the market, perspectives and a transition management plan. It is the most suitable tool for companies looking to make better options. Board collects data from all fields as well as optimizes the data for finding in this study into the selected form. Board does have the most appealing and robust design among all sector BI applications. The Board provides meaningful interpretation, process management, and production scheduling facilities.

5. CONCLUSION

This paper clarified the key concepts of data mining and the popular and most used data mining techniques. And it also clarified the types of data mining tools and the fundamental aspects of the data mining tools described. It gives the basic idea and features of the tools available and also helps to choose the tools that are suitable for a particular area of the research and mining process. Future enhancement

of this work can be done with a complete comparative study with sample data with these tools.

Reference

[1] http://research.ijcaonline.org/volume74/number5/pxc3889673.pdf

[2] http://www.ijarcsse.com/docs/papers/Volume_3/3_March2013/V3I3-0162.pdf

[3] MH Dunham, "Data Mining: Introductory and Advanced Topics," Prentice-Hall, 2002.

[4] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics," Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[5] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996.

[6] Article: Exforsys Inc "Data Mining applications" Published on 26th Jul 2006 Source: http://www.exforsys.com/tutorials/data-mining/data-mining-applications.html

[7] Jiawei Hanand Micheline Kamber, DataMining: Concepts and Techniques, Morgan Kaufmann Publishers, 2003.

[8] Y.Ramamohan, K. Vasantharao, C. KalyanaChakravarti, A.S.K.Ratnam "A Study of Data Mining Tools in Knowledge Discovery Process" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012