International Journal of Advance Research in Science and Engineering
Volume No. 10, Issue No. 10, October 2021
www.ijarse.com
IJARSE
ISSN 2319 - 8354

AN OPEN SOURCE TOOLS FOR DATA MINING

Fatima[1], Dr. Md. Jawed Ikbal Khan[2]

[1]Research Scholar, Magadh University Bodh Gaya, Gaya, Bihar, India.

[2]Associate Professor, Deptt. of Mathematics, Mirza Ghalib College, Gaya, Bihar, India.

ABSTRACT

The concept of Data Mining has emerged to meet the 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. 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.

Keywords: Data mining, Data mining tools, Rapid miner, Weka, Rattle(R).

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge.

KDD process consists of iterative sequence methods as follows:

1. Selection: Selecting data relevant to the analysis task from the database

2. Preprocessing: Removing noise and inconsistent data; combining multiple data sources

3. Transformation: Transforming data into appropriate forms to perform data mining

4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; extracting data patterns

5. Interpretation/Evaluation: Interpreting the patterns into knowledge by removing redundant or irrelevant patterns; translating the useful patterns into terms that human understandable.

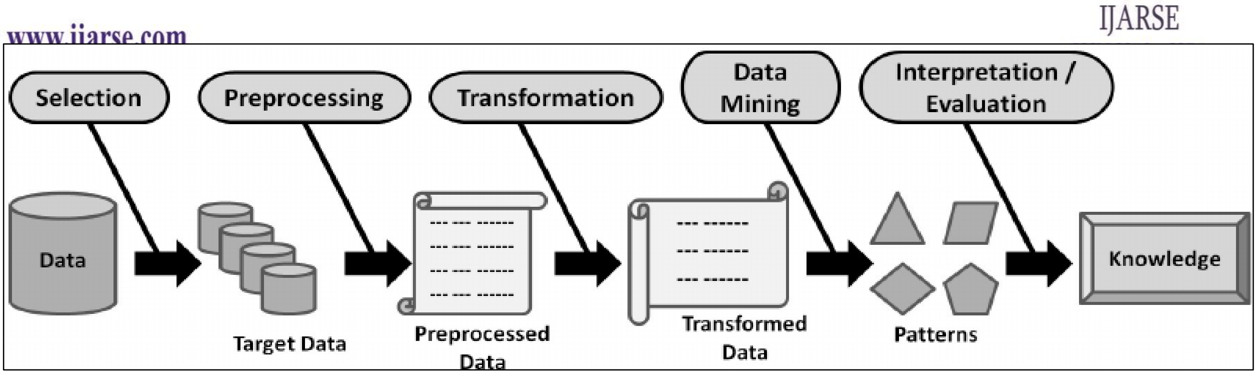


Figure 1. Knowledge Discovery Database Life Cycle

1.1 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:

1.1.1. 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.

1.1.2. 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.

1.1.3. 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.

II. LITRETURE REVIEW

Data mining is a process of discovering knowledge from data warehouse. This knowledge can be classified in different rules and patterns that can help user/organization to analyze collective data and predicted decision processes [1]. Centralized database of any organization is known as Data warehouse, where all data is stored in a single huge database. Data mining is a method that is used by organization to get useful information from raw data. Software’s are implemented to look for needed patterns in huge amount of data (data warehouse) that can help business to learn about their customers, predict behavior and improve marketing strategies. Web mining is actually an area of data mining related to the information available on internet. It is a concept of extracting informative data available on web pages over the internet [2]. Users use different search engines to fetch their required data from the internet, that informative and user needed data is discovered through mining technique called Web Mining. Different tools and algorithms are used for extraction of data from web pages that includes web documents, images etc. Web mining is rapidly becoming very important due to size of text documents increasing over the internet and finding relevant patterns, knowledge and informative data is very hard and time consuming if it is done manually. Structure (Hyperlinks), Usage (visited pages, data use), content (text document, pages) are included in information gathered through Web mining [3], [4]. Term World Wide Web is related to the combination of web documents, videos, audios etc. Some processes included in web mining are: Information Retrieval is a process of retrieving relevant and useful information over the web. Information retrieval has more focuses on selection of relevant data from large collection of database and discovering new knowledge from large quantity of data to response user query.IR steps includes searching, filtering and matching [4], [5]. Information extraction is an automatic process of extracting analyzed data (structured). IE is a task that work same like information retrieval but more focuses on extracting relevant facts [5]. Machine Learning is support process that helps in mining data from web. Machine learning can improve the web search by knowing user behavior (interest). Different machine learning methods are used in search engine to provide intelligent web service. It is much more efficient than traditional approach i.e. information retrieval. It is a process that has ability to learn user behavior and enhance the performance on specific task. Nanpoulos et al. [6] proposed a method for discovering access patterns from web logs based on a new type of ssociation patterns. They handle the order between page accesses, and allow gaps in sequences. They use a candidate generation algorithm that requires multiple scans of the database. Their pruning strategy assumes that the site structure is known. Srikant and Agrawal [7] presented an algorithm for finding generalized sequential patterns that allows user-specified window-size and user-defined taxonomy over items in the database. This algorithm required multiple scans of the database to generate candidates. Parthasarathy et al. [8] introduced a mining technique given incremental updates and user interaction. This technique avoids re-executing the whole mining algorithm on the entire data set. A special data structure called incremental sequence lattice and a vertical layout format for the database are used to store items in the database associated with customer transaction identifiers. Their performance study has shown that the incremental mining is more efficient than re-computing frequent sequence mining process from scratch. However, the limitation of their approach, as they point out, is the resulting high memory utilization as well as the need to keep an intermediate vertical database layout which has the same size as the original database.Liu et al. [9] proposed a clustering method based on a mixture of Markov models to cluster users and capture the sequential relationships hidden in user web navigation histories. The performance of this method is higher than the traditional Markov models, the association rules, or clustering methods. Fu et al. [10] were one of the early researchers to propose the idea of generalizing web data and integrated this with a clustering algorithm to extract web access patterns. A page hierarchy is used to generalize sessions by replacing actual pageclicks with their general URLs. For example, a page like /programs/ugrad/cs/ is replaced by /programs/ugrad/or /programs/ depending on a pre-determined generalization level. This level, which is critical to both the efficiency and effectiveness of the approach, is a user-specified parameter. Yang et al. [11] presented an application of web log mining that combines caching and perfecting to improve the performance of internet systems. In this work, association rules are mined from web logs using an algorithm called *Path Model Construction* [12] and then used to improve the GDSF caching replacement algorithm. These association rules assumes order and adjacency information among page references

III. TECHNIQUES AND TOOLS

1. TANAGRA is 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. Rapid Miner as a powerful engine for analytical ETL, data analysis, and predictive reporting, the new business analytics server. Rapid Analytics is the key product for all business critical data analysis tasks and a milestone for business analytics.

3. 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.

4. 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.

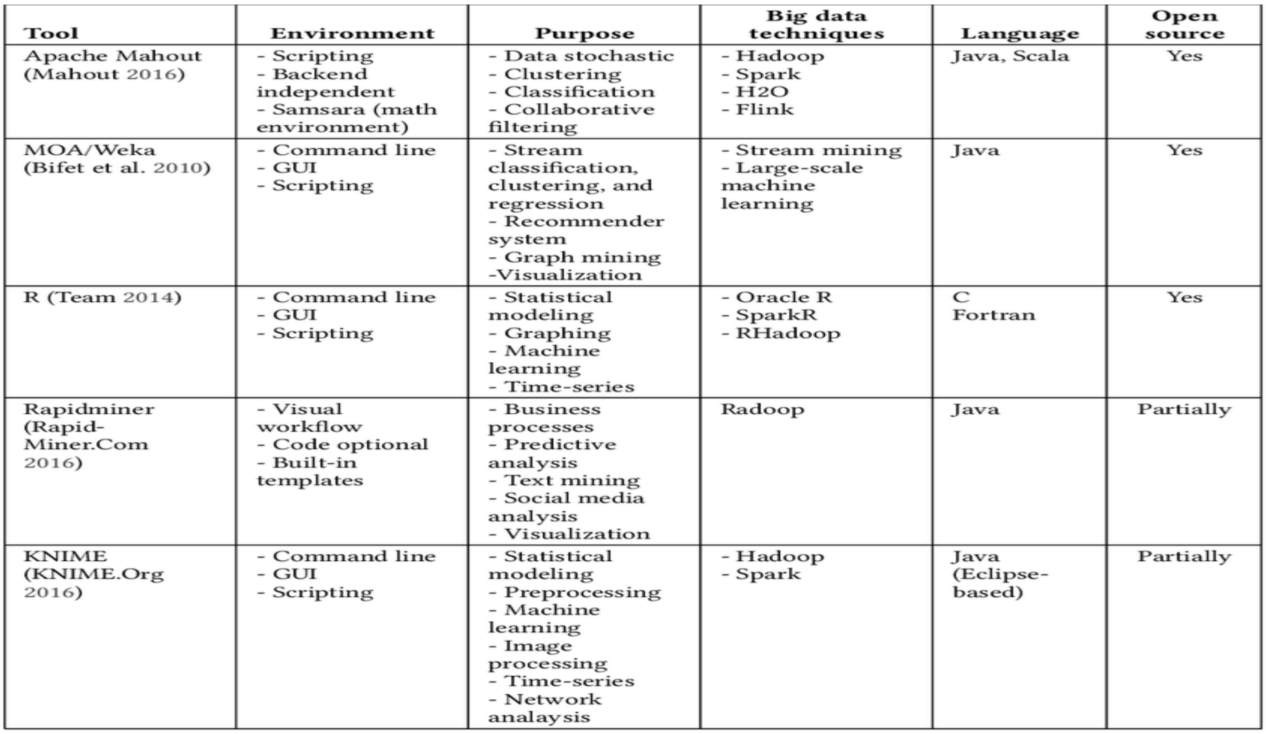
5. KNIME is a user-friendly graphical workbench for the entire analysis process: data access, data transformation, initial investigation, powerful predictive analytics, visualization and reporting. The open integration platform provides over 1000 modules (nodes).

6. 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.

7. 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 existing 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.

8. 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).

9. Rattle (the R Analytical Tool to Learn Easily) presents statistical and visual summaries of data, transforms data into forms that can be readily modeled, 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.



Tool	Environment	Purpose	Big data techniques	Language	Open source
Apache Mahout (Mahout 2016)	- Scripting - Backend independent - Samsara (math environment)	- Data stochastic - Clustering - Classification - Collaborative filtering	- Hadoop - Spark - H2O - Flink	Java, Scala	Yes
MOA/Weka (Bifet et al. 2010)	- Command line - GUI - Scripting	- Stream classification, clustering, and regression - Recommender system - Graph mining -Visualization	- Stream mining - Large-scale machine learning	Java	Yes
R (Team 2014)	- Command line - GUI - Scripting	- Statistical modeling - Graphing - Machine learning - Time-series	- Oracle R - SparkR - RHadoop	C Fortran	Yes
Rapidminer (Rapid-Miner.Com 2016)	- Visual workflow - Code optional - Built-in templates	- Business processes - Predictive analysis - Text mining - Social media analysis - Visualization	Radoop	Java	Partially
KNIME (KNIME.Org 2016)	- Command line - GUI - Scripting	- Statistical modeling - Preprocessing - Machine learning - Image processing - Time-series - Network analaysis	- Hadoop - Spark	Java (Eclipse-based)	Partially

Table 1. Comparatively Techniques for Data Mining Tools

IV. CONCLUSION

Data mining is a concept that helps to find information which is needed from large data warehouses by using different techniques. It is also used to analyze past data and improve future strategies. In this paper we described three important types of data mining that can help in finding informative data. Each type has different algorithms, tools and techniques that are used for data retrieval. Various tools and techniques for each type are described. All techniques may have some advantages and disadvantages but drawbacks can be improved by further studies.

REFERENCE

[1] Claus Pahl and Dave Donnellan, "Data Mining Technology for the Evaluation of Web-based Teaching and Learning Systems," 7th Int. Conference on E-Learning in Business, Government and Higher Education, October 2002.

[2] Anurag Kumar and Ravi Kumar Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 12, pp. 1543-1547, December 2016.

[3] Simranjeet Kaur and Kiranbir Kaur, "Web Mining and Data Mining: A Comparative Approach," International Journal of Novel Research in Computer Science and Software Engineering, vol. 2, no. 1, pp. 36-42, January - April 2015.

[4] R. Malarvizhi and K Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 2940-2945, Augest 2013.

[5] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey," SIGKDD Explorations, vol. 2, no. 1, pp. 1-15, July 2000.

[6] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos., *Effective prediction of web-user accesses: A data mining approach*, In WEBKDD Workshop, San Francisco, CA, Aug. 2001.

[7] R. Srikant and R. Agrawal, *Mining Sequential Patterns:Generalizations and Performance Improvements*, In EDBT,1996.

[8] S. Parthasarathy, M. J. Zaki, M. Ogihara, and S. Dwarkadas, *Incremental and interactive sequence mining*, In *CIKM*,1999.

[9] Liu., Huang. X, Personalized Recommendation with Adaptive Mixture of Markov Models, The American Society for Information Science and Technology,2007

[10] Y. Fu, K. Sandhu, and M. Shih., *A generalization-based approach to clustering of web usage sessions*. In Intl. WEBKDD Workshop,1999.

[11] Q. Yang, H. H. Zhang, and I. T. Y. Li, *Mining web logs for prediction models in WWW caching and prefetching*, In KDD,2001

[12] Y. Fu, K. Sandhu, and M. Shih., *A generalization-based approach to clustering of web usage sessions*. In Intl. WEBKDD Workshop, 1999.