ISSN: 2456-0006

Volume 1, Issue 1, April 2016

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Open Source Data Mining Tools

Dhara Parmar[1]
JESITMR,Nashik

Department of Computer Engineering
Savitibai Phule Pune University
Nashik, India
parmard777@gmail.com

Abstract: ***Now-a-days, open source tools are used in today's generation, as everyone around knows that they are good and have the potential to become the best because they are the results, motivation and passion and is not just a monthly pay-cheque. Data mining is a widely used new concept the industry, also the commercial tools and methods are expensive, and thus, second alternative is open-source tools, at the moment. The open-source tools is not having the facilities for customer support, timely updates or any fancy advertisements and offer and used. They are meant for developers and others who cannot afford a hole in their pockets, but they still want to their jobs to be done. Finding the most appropriate tool is essential. What is the complexity degree? The amount of data will be handled propely? What kind of data would we deal with? Whether the tool can do open data visualization? The paper gives the comprehensive and theoretical analysis of open source tools such as data mining. This describes the technical specification, features, and specialization for each selected tool along with its applications.***

Keywords: Data mining, Data Mining Tools, open source Tools.

I.INTRODUCTION

The history depend on the deeply rooted of the machine learning, artificial intelligence concepts, and also DB along with its statistics data mining was earlier coined. Data science was strongly associated and involved in classification of data and also the manipulation is done by applying various concepts statically. The important phase is data mining in knowledge discovery which includes the application like discovery, analytical methods on which data is to be produced specifically. The available data is everywhere. It's also used for predicting the future. The statistical approach is also used. The data mining is also an extension of data analysis which is traditionally and statistical approach is that which incorporates techniques like analytical drawing from a disciplines ranges. The widespread availability of huge is described also the complex information of data sets, and the ability of extracting the useful hidden knowledge in that data and also to act of that knowledge which has been increasingly in today's world. Hence the data mining is going to analyze the large data sets which were observational to find relationships which were unsuspected and also to summarize the data which is understandable and useful for the data owner. [1]. The detailed, data mining the approach to research and analysis [2]. It is exploration and analysis of large quantities of data in order to discover meaningful patterns and rules [3].

Different researchers and also the practitioners are using the data mining as their synonym for their knowledge discovery but the data mining is just another one step for the knowledge discovery processing. The techniques is following an automated processes of knowledge discovery (KDD) which includes data cleaning, data integration, data selection, data transformation, data mining and knowledge representation [5].

Types of Data mining are as follows:

- Flat files: Flat files are actually making the most common data sources in making data mining algorithms,specially in the research level. These simple data files are included in the text or in binary format with a structure type known as the data mining algorithm which is applied. The data in which these files the transactions and the time-series data, and also the scientific measurements, etc.

- Relational Databases: It consists of various data sets of tables which contains the values for the entity attributes and values of attributes in the entity relationships databases. Tables consists of columns and rows, an in the columns it represents the attributes while the rows represents tuples.

• Transaction Databases: It is the set of records which represents the transactions, when each of the time stamp, the identifier and also includes the set of items. Associating with the transactions files which will be descriptive data in the items. For ex., the case of video store, and the rentals table.

• Multimedia Databases: Multimedia databases also includes videos, images, and text media contents. They are used to store on extended object-oriented databases concepts, or simply on a file system. Multimedia is then characterized by its high dimensions, by making the data mining by making its more challenging. Then from multimedia repositories that may requires more computer visions, computer graphics concepts, and image interpretation, also NLP methodologies.

• Spatial Databases: Spatial databases are the databases in which addition is to usual data, the store geographical information includes maps, and also the global or regional positioning is includes. It also includes he new challenges for the data mining algorithms.

• World Wide Web: The WWW is the most used heterogeneous and dynamic repository which is available widely. A large number of authors and the publishers are also continuously contributing to their growth and the metamorphosis, and the massive number of users is included while accessing its resources on daily bases. Data in the WWW is organized in their inter-connected documents specified. These documents can be in text format, audio, video, raw data, and even some applications.

• Time-Series Databases: Time-series databases contain the time related data which includes the stock market data or even the logged activities. These databases are usually having the continuous flow of new data coming in markets, which sometimes reflects the need for the challenging real-time analysis. Data mining is also such type of databases which commonly includes the study of trends and even the correlations between the evolutions of different variables, as well as predictions of the trends and the movements of those variables in the time.

II. DATA MINGING TOOLS

Applications which are ranging from the markets and advertising of services and products, AI research, biological sciences, and crime investigations to those higher-level government intelligences. Because of this the widespread use and the complexity involving in building and data mining applications, and also the large numbers of Data mining tools are being developed over the decades. Each tool has its their own advantages and disadvantages. [6] With the data mining concepts, the group of tools that is having been developed by a research community and data analysis enthusiasts; they are offered free of charge using one of the existing open-source licenses. An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences Data mining provides many mining techniques to extract data from databases. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions. The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult. [7] The top open source tools available for data mining are briefed as below.

A. **RapidMiner** (formerly known as YALE)

It is written in the Java Programming language. This tools offer advanced analytic through template based framework. A bonus to the users hardly has to write any code. It is offered as a service, rather than a piece of local software. These tools hold top position on the list of data mining tool.
In the addition to data mining, Rapid Miner also provides functionalities like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation and deployment. What makes it even more powerful? It provides learning schemes, models and algorithms from WEKA and scripts. Rapid Miner comes under the AGPL open source license and it can be downloaded from Source, Where it is rated the number one business analytics software.

B.**Weka-:**
The original non-Java version of the WEKA primarily was developed for analyzing data for the agricultural domain. With the Java-based version, the tool is sophisticated and used in many different applications or ways including visualizations and algorithms for data analysis and predictive modeling. It's free for the GNU General Public License, which is compared to Rapid Miner because users can customize it however they please it. It support several standard data mining tasks including data preprocessing, clustering, classification, regression, visualization and the feature selection. WEKA would be the more powerful with the addition of sequence modeling. Which is currently is not included.

C. **R-Programming-:**
What if I will tell you that Project R a GNU project is written in R itself? It is a primarily written in C and FORTRAN. And a lot of its modules are written in the R itself. It is a free software programming language and software environment for statistical computing and graphic. The R language is widely using among the data miners for developing statistical software and data

analysis. Ease of use and extensibility has raised popularity substantially in the recent years. Data mining provides statistical and graphical techniques including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

D**. Orange-:**
Python is popular because it's simple and easy to learn yet powerful. Hence, when it comes to the looking for a tool for your work and you are a Python developer. It looks no further than Orange a Python-based, powerful and open source tools for both novice and expert. You will fall in love with this tool visual programming and Python scripting. It also has a component for machine learning add ons for bioinformatics and text mining. It is packed with features for data analytics.

E. **KNIME-:**
Data preprocessing has three different components extraction, transformation and loading. KNIME does all three. It will give you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analysis, reporting and integration platform. KNIME also integrate various components for machine learning and data mining through the modular data pipelining concept and has caught the eye of business intelligence and financial data analysis. Written in the Java and based on Eclipse KNIME is easy to extend and to add plugins. Additional functionalities that can be added on go. There is a Plenty of data integration modules are already included in the core version.

F.**NLTK-:**
When it comes to the language processing task, nothing can beat the NLTK. NLTK provide a pool of language processing tools including the various data mining, machine learning, data scraping, sentiment analysis and other various language processing task. All you need to do is to install NLTK pull packages for your favorite task and you are ready to go. Because it is written in Python you can build applications on top of it by customizing it for small tasks.

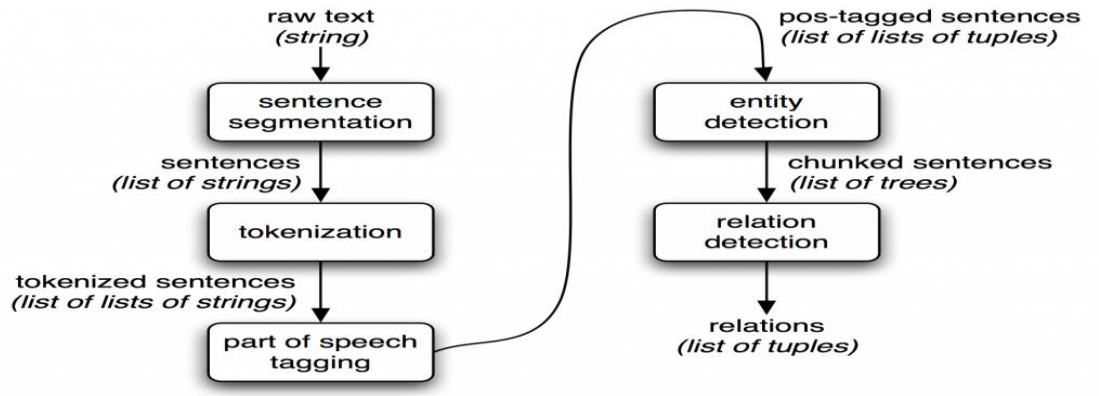

Fig-1-Nltk Work Process

III. COMPARATIVE STUDY OF TOOLS

The best available open source data mining tools were chosen and analytical study was made by taking into

Sr. No	Tool Name	Release Date	License Operating	Operating System	Language	Website
1.	RAPID MINER /6.0	2006 21November,2013	AGPL Proprietary	Cross platform	Language Independent	www.rapidminer.com
2.	WEKA	April,2014/3.7.11	GNU General Public License	Cross Platform	Java	www.cs.waikato.ac.nz/~ml/weka
3.	R	April,2014/3.1.0	GNU General Public License	Cross Platform	C, Fortran and R	www.rproject.org
4.	ORANGE	May,2013/2.7	GNU General Public License	Cross Platform	Python C++,C	www.orange.biolab.si
5.	KNIME	December,2013/2.9	GNU General Public License	Linux ,OS X, Windows	Windows Java	www.knime.org
6.	Nltk	March,2014/3.2	Apache	Windows, Linux	Python	www.nltk.org

Table-1: Technical Overview of best six data mining open source tools

CONCLUSION

The study has presented the specific details along with the description of various open source data mining tools enlisted area of specialization. With the recent Meeting of various developers concerning the uses of tool in various fields one can expect more enhanced environment along with the more technical improvement. The work can be a helping hand to provide an idea in future to develop an application with more efficiency, availability and reliability i.e. a tool can be designed that can be extended to more fields rather than supporting a specific area. The efforts may be increased and the development may be a complex procedure but it can result in an efficient product.

REFERENES

1. Hand David, Mannila Heikki, Smyth Padhraic.: "Principles of data mining", Prentice hall India, pp.1, 2004.

2. Sethi I. K., "Layered Neural Net Design Through Decision Trees, Circuits,and Systems", IEEE International Symposium,1990.

3. Meheta M., Aggarwall R., Rissamen I. : "SLIQ:A fast Scalable Classifier for Data Mining", In Proc. International Conference Extending data base Technology(EDBI), Avignon, France, March 1996.

4 . Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.),Advances in Knowledge Discovery and Data Mining, AAAI Press, Cambridge, 1996.

5. Kalpana Rangra Dr. K. L. Bansal: "Comparative Study of Data Mining Tools" International Journal of Advanced Research in Computer Science and Software Engineering,ISSN:2277-128x,pp216-223;

6. Witten, I.H., Frank, E.: "Data Mining: Practical machine Learning tools and techniques", 2nd addition,Morgan Kaufmann, San Francisco(2005).

7. Ralf Mikut and Markus Reischl Wiley Interdisciplinary Reviews: "Data Mining and Knowledge Discovery" Volume 1, Issue 5, pages 431–443, September/October 2011.

8. Alcala-Fdez, J.,L., del Jesus, M.J., Ventura, s., Garrell, J.M, Otero, J., Romero,C., bacardit, j., Rivas, V.M., Fernandez, J.C., Herrera., F., : "KEEL: A software tool to Assess Evolutionary Algorithms to Data mining Problems", Soft computing 13:3,pp 307-318(2009).

9. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler,T. "YALE: Rapid Prototyping for Complex Data Mining tasks", in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-06), pp. 935-940, 2006.

10. http://orange.biolab.si/features

11. https://github.com

12. http://www.r-project.org

13. http://www.knime.org

14. http://rapidminer.com