Volume 8, Issue 6
December 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

A Study on Effective Online Data Mining Tools

[1]Mrs. K. Thilakavalli, [2]Dr. R. Umadevi, [3]Mr. D. Rajagopal
[1]Assistant Professor, Department of Computer Science and Computer Applications, Vivekanandha College for Women, Tiruchengode, Namakkal Dt, India.
[2]Professor & Head, Department of Computer Science and Computer Applications, Vivekanandha College for Women, Tiruchengode, Namakkal Dt, India.
[3]Assistant Professor, PG and Research Department of Computer Science and Computer Applications, Vivekanandha College of Arts and Sciences for Women (Autonomous), Tiruchengode, Namakkal Dt, India.
E-mail: [1]thilaksathya2782007@gmail.com , [2]mail2deviuma@gmail.com , [3]sakthiraj2782007@gmail.com

ABSTRACT

Data mining is the technique of extraction of predictive facts from massive facts hundreds and it strongly associated with facts technological know-how that entails manipulation and class of facts by making use of statistical and mathematical. Statistics managing and coping with of massive facts units, there are numerous facts mining equipment are to be had which improve data high quality from past to provide. Data mining tools are supplied for preprocessing records, feeding it into a diffusion of mastering schemes, and analyzing the resulting classifiers and their performance, ideas. Nowadays, big portions of records are being accrued. Seeking knowledge from huge statistics is one of the maximum desired attributes of data mining. In the analysis of examining information, some preliminary suspicion about the data is known, but data mining may want to support roughly the records in a more in-depth expertise. Recently manual data analysis has been small, but it creates a bottleneck to analyze large data.

Keywords: *Data Mining, Large Data Set, Data mining tools, Data Analysis.*

1. INTRODUCTION

Data mining is the method of extraction of predictive records from huge statistics hundreds. It is able to also be defined as a method of reading information from extraordinary perspectives and summarizing it into beneficial information. With a extensive records deeply rooted in machine learning, artificial intelligence, database in conjunction with statistics data mining changed into coined very early. Data mining is strongly related to statistics technology that involves manipulation and category of records with the aid of making use of statistical and mathematical. Data managing and dealing with of large information units, there are various data mining tools are to be had which improve records first-class from past to present. Data mining tools are furnished for preprocessing statistics, feeding it into a variety of gaining knowledge of schemes, and analyzing the resulting classifiers and their performance, standards.

Web data mining is mostly a technique of data mining, which is utilized for serving internet-based totally packages by means of the usage of net information over World Wide Web. It is far a method of retrieving information over the World Wide Web that incorporates web-based documents, hyper documents, internet links to numerous web pages and different resources over the internet. Web data mining is a term used for a technique, through which diverse web sources are used for collecting the useful data that makes it smooth for an character or a enterprise for using those resources and facts of their best interest. It evolves three fundamental strategies such as structure mining, content mining, and usage mining.

2. NEED FOR DATA MINING

Today World Wide Web (WWW) has become a complex universe as it updates regularly. WWW is largely a source of huge quantity of statistics that offers all of the considered necessary resources of data mining. Nowadays, massive quantities of records are being accrued. It is said that the amount of data collected is nearly doubled each year. One of the most desired characteristics of data mining is the quest for information from large data. In two ways, the data could be massive. There can usually be a big gap between the stored information and the details that could be extracted from the data. The transition will not take place automatically, this is where data mining takes place.

Furthermore, the web mining can be classified into three main categories, which are as follows:

a) Web Structure Mining
That is particularly used for describing the structure of the contents of the website. It may be defined in terms of graph, in which net pages are its nodes and its links are its edges [4]. It indicates the internet links from one net web page to some other consequently and able to say it shows the relationship some of the web and users. The point of interest of this is structural precise of net pages and web sites. Web structure mining specially works on- hyperlink mining, inner structure mining and URL mining [3].

b) Web Content MiningThis technique is used for extracting the specified content from the diverse web pages available and its contents, which include, image, audio, video, text, and many others. The number one assets of web, which are mined, are character internet pages. Internet content material mining is especially associated with text mining due to the fact maximum of the net content material is inside the shape of text [2]. Therefore, internet

content material mining needs its own packages of textual content mining and lots of different awesome methods. It in particular specializes in- web text mining and internet multimedia mining [3].

c) Web Usage Mining

This method is used to define the mode through which customers can engage with the servers or can access to be had internet pages. It consists of facts generated by consumer server transaction from one or extra web localities [4]. Its foremost objective is to locating the usage styles from packages primarily based on net. It includes three phases: preprocessing, discovery of usage sample and analysis of the sample. Server logs utilize this approach of mining and it geared toward getting useful users.

3. CATEGORIES OF DATA THAT CAN BE MINED

Flat files: Flat files, particularly at the research level, are the most popular source of data for data mining algorithms. Flat files are simple text or binary data files with a structure that is understood to be implemented by the data-mining algorithm.

Relational Databases: In a nutshell, a relational database consists of a set of values.

Data Warehouses: A data warehouse as a storehouse is a database of data collected (often heterogeneous) from multiple data sources and planned to be used as a whole under the same hierarchical schema. A data warehouse enables data to be processed under the same roof from different sources.

Transaction Databases: A transaction database is a set of transaction records, each with a time stamp, an identifier, and a set of items. It could also be descriptive data for the items associated with the transaction files. For example, the rentals table for the video store.

Multimedia Databases: The combination of video, pictures, audio and text media is included in a multimedia server. They can be stored either on stretched object-related or object-oriented databases, or simply on a file system.

Spatial Databases: Spatial databases are databases that store geographic information such as maps, as well as global or regional positioning, in addition to normal data. These spatial databases pose new challenges to the algorithms of data mining.

World Wide Web: The World Wide Web of data mining is the most heterogeneous database. In interconnected data, a data is organized in the World Wide Web. Such data could include text, audio, video, raw data, and even presentations. Theoretically, the WWW consists of three important components:

I. The content of the Web, which includes documents available.
II. The structure of the Web, which covers the hyperlinks and the relationships between documents.
III. The usage of the web, describing how and when the resources are accessed.

Time-Series Databases: Such database systems provide time-related data such as market data or activity registered. Usually, these databases have a continuous flow of new data coming in, sometimes causing the need for a stimulating real-time analysis.

4. TYPES OF DATA MINING TOOLS

There are three main categories of data mining tools. Modern methods for data mining, application-based applications / commercial technology and data mining tools on the internet. Description of each is as follows:

a) Traditional data mining tools
Some mining programs are work as traditional way to collect and analyze data, which used by various companies for decision-making process of large data sets. Windows and UNIX versions support majority of these. However, sometimes handling with only one database type.

b) Application based tools
An application that displays the sector, data performance-oriented gui. Historical data are used as a guide in this, and the current trends are reviewed to see the market changes. Application-based technologies are therefore simple to use and assist in administrative work and provide business quality services.

c) Web based data mining tools
Because of its ability to extract different types of text from any written materials, this category of software is called text-mining tool. Additionally, help to scan and convert data into a selected format that is compatible with any tools.

5. POPULAR WEB DATA MINING TECHNIQUES

a) Classification
It is one of the most commonly used data mining technique. It consists of a set of predefined examples for developing a new model, which can easily classify massive amount of data records. As its name suggests it is basically a group of items that belong to a particular category on the basis of their common features [6]. The primary aim of this technique is to assign an accurate class to the previously unseen records.

b) Association Rule Mining
It is a basic technique of web data mining that is used for associating relationships among a set of variables and its data items. It consists of two parts antecedent and consequent, an antecedent is the data item and consequent is data item found in combination of antecedent [7]. It is a technique of analyzing data for if/then patterns.

c) Artificial Neural Network
Artificial Neural Network is one of the data mining techniques that is based on the works perform by the brain or a particular task perform by the brain [8]. It is the interconnected group of nodes with a vast network of neurons in a brain. This technique is used in web data mining for gathering information from the web in the form of neural networks which may be linear or non-linear and utilizing this required information for one or other purpose of the end user.

d) Clustering
Clustering is also one of the popular techniques of data mining which is based on concept of hierarchy model which groups together those items which are having similar features. It is believed that making group of similar items into a cluster is very helpful for retrieving the relevant information easily and quickly and allows the users to focus their search in the right direction [9]. The cluster of similar items makes it more appropriate data gathering system.

6. WEB DATA MINING TOOLS

There are various Web data mining tools as open source softwares, which are freely available for mining of web data. These tools have been used to gather correct and perfect information by using weblog data. In this section, some of the useful and popular web data mining tools are explored and discussed here.

a) WEKA
Written in Java, WEKA (Waikato Knowledge Analysis Environment) is a well-known machine learning technology tool[6]. WEKA supports a variety of traditional data mining activities, including preprocessing, clustering, classification, regression, visualization, and selection of features.

The strategies are based on the assumption that the data is accessible as a single flat file or relationship, where each data point is marked by a fixed number of attributes. WEKA provides access to Java Database Connectivity (JDBC) SQL servers and can process the result returned through a database query. The main user interface is the Explorer, but it is possible to access the same features from the command line or via the information flow interface depending on the component.

Common Features:

- WEKA is the open source tool based on Java which is a group of machine learning algorithms
- Robust in machine learning techniques.
- WEKA is greatest suited for mining association rules

Advantages:

- WEKA loads data file in layouts of ARFF, CSV, and C4.5, binary. Though it is open source, Free, Extensible, Can be integrated into other java packages.

b) Tanagra
Tanagra is free Data Mining software for academic and research purposes [2,3]. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning, and databases area. It runs under almost Windows Systems, in any case it has been tested under Windows 98, 2000, XP, Vista, and Windows 7/8.1.

c) Orange
It is a component-based data mining and machine learning software tool that provides a friendly but efficient, quick and scalable front-end visual programming for explorative data analysis and visualization, as well as Python bindings and scripting libraries. This provides a complete set of components for pre-processing data, scoring and sorting functionality, modeling, model analysis, and techniques for exploration. It is written in C++ and Python[1, 3] and its user interface is based on a system cross-platform.

Common Features

- Orange tool contains a set of modules for data preprocessing, feature scoring and filtering, modeling, model evaluation, and assessment techniques.
- It is also very useful for analytical process, which have user-friendly visual programming or python scripting.
- Specially, these tools have utilities for Bioinformatics Add-On and Text Mining Add-On.

Advantages

- It is an open source data mining package build on Python, NumPy, enclosed C, C++, and Qt.
- Orange is written in python therefore is easier for most programmers to learn.
- It has superior debugger.

d) Rapid Miner
Formerly called YALE (Yet Another Learning Environment) is an environment for machine learning and data mining experiments that is used for both research and data mining tasks in the real world[4, 8]. This enables experiments to be made up of a vast number of arbitrarily nest-capable operators that are specified in XML files and rendered with Rapid Miner's graphical user interface. Rapid Miner offers over 500 operators for all major machine learning applications, integrating learning schemes and WEKA learning environment attribute evaluators. It is available for data analysis as a stand-alone tool and as a data-mining engine that can be built into your own goods.

Common Features:

- Rapid Miner has a collection of functionality, is elegant, and has good connectivity.

- Rapid Miner consists of many learning algorithms from WEKA.
- Compact and whole package.
- It simply reads and writes Excel files and diverse databases.

Advantages

- Rapid Miner has over 1,500 methods for data integration, data transformation, analysis and, modeling as well as visualization – no other solution on the market offers more procedures and therefore more possibilities of defining the optimum analysis processes.
- Rapid Miner deals several procedures, mainly in the part of attribute selection and for outlier detection, which no other clarification offers.

e) KNIME

KNIME (Konstanz Data Miner) is an open-source data integration and collection, research and discovery system that is user-friendly, intelligible and comprehensive[5,11]. It offers users the ability to visually create data flows or pipelines, perform some or all of the analytical steps selectively, and then analyze the tests, models, and interactive views. KNIME is written in Java and is based on Eclipse and uses its extension system to support plug-in with additional features. Users can add modules for text, image, and time series processing and incorporating various other open source projects such as R programming language, WEKA, and LibSVM etc. via plug-in.

Common Features:

- KNIME, pronounced "naim", is a nicely designed data mining tool that runs inside the IBM's Eclipse development environment.
- The KNIME base version already incorporates over 150 processing nodes
- KNIME is easy to extend and to add plug-in. Additional functionalities can be added

Advantages:

- It integrates all analysis modules of the well known. WEKA data mining environment and additional plug-in allow R-scripts to be run, offering access to a vast library of statistical routines.

f) Screen-Scraper

This is a method that is used to extract data from websites and uses that information in other database-like ways which facilitates data mining through the World Wide Web. This involves web data mining consisting of database scanning, which communicates with the technology available to meet the requirements. One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet [10]

g) Web Info Extractor

This tool is helpful in mining web data, extracting web content, and monitoring content update. Thorny template rules are not required to be defined. For mining web data and for content retrieval it is a very powerful tool. Some of the features [5] are as follows:

- No need to learn boring and complex template rules and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database.
- Monitor Web pages and extract new content when update.
- Can deal with text, image, and other link file.
- Can deal with Web page in all language.
- Running multi-task at the same time.
- Support recursive task definition.

h) Automation Anywhere

Automation anywhere is a tool used for data extraction used for retrieving web data, screen scrape from Web pages. It is also used for Web mining. Its main features[5] are as follows:

- Unique SMART Automation Technology for fast automation of complex tasks.
- Record keyboard and mouse or use point and click wizards to create automated tasks quickly.

i) Web Content Extractor

It is a powerful and easy to use data extraction tool for web scraping, data mining or data extraction from the internet [2]. Some of the features are:

- It helps to collect the market figures, product pricing data, or real estate data.
- It helps users to extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- It assists users in automate extraction of auction information from auction sites.
- It assists to Journalists extract news and articles from news sites.
- It helps people seeking a job extract job postings from online job websites. Find a new job faster and with minimum inconveniences
- It Extracts the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites[5].

j) Web Log Expert Tool

Web Log Expert is a web analyzer Data mining tool that is fast and powerful[7]. This software tool helps to expose vital information about the use of a website such as: user behavior, access stats, and website routes, apps for users, and more.

k) Absolute Log Analyzer Tool
Absolute Log Analyzer is a client-based software tool for the study of internet traffic[7]. First, you need to add log files to the report and then show the data. Apart from the graphical user interface (GUI), Absolute Log Analyzer also has a Command Line Interface (CLI).

l) R
R is also an open source statistical analysis software developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, based on C and FORTRAN programming language. R (Revolution) is a language and computer platform for numerical computation and graphics free software programmingIt is widely used in the production of statistical technology and analysis among statisticians and data miners. The R's Strength is the easy way to produce well-designed quality plots, including mathematical symbols and formulas.

Common Features

- R is a well-supported, open source, command line driven, statistics package.
- It consists up to hundreds of extra "packages" freely available, which provide all sorts of data mining, machine learning, and statistical techniques.
- R provides less support to data mining and WEKA, it algorithms as compared to Rapid Miner does implement a few data mining algorithm.

Advantages

- R is more transparent since the Orange are wrapped C++ Classes.
- Programming in R really is very different, we are working on a higher abstraction level, but we do lose control over the details.
- It has the ability to make a working machine learning program in just 40 lines of code.

7. VARIOUS BENEFITS OF INTERACTIVE DATA MINING

a) Mining different kinds of knowledge from database Need of different user is not same and different user may be interested in different kind of knowledge. Hence, it is necessary to cover broad range of knowledge discovery.
b) Interactive mining of knowledge at multiple levels of abstraction- Mining process need to be interactive because it allows user to focus the search for pattern. In interactive system user are providing their feedback, which is valuable to the system.
c) Adaptive and effective communication between user and system. User views, preferences, strategies play important role in user and system interactivity.

8. CLUSTERING TECHNIQUE ALGORITHM

a. Incremental k-means Algorithm: Incremental k-means is a widely used clustering algorithm in various applications. K-means value algorithm is a efficient algorithm to resolve clustering issues, this algorithm is relatively simple and fast. For large data collection, this algorithm is relatively flexible and high efficient, because the Complexity is O (ntk). Among which, n is the times of iteration, k is the number of cluster, t is the times of iteration.
b. Cobweb: (COBWEB using the modified category utility) Cobweb is incremental system for hierarchical conceptual clustering, it generate hierarchical clustering where clusters are described by probabilistically. Cobweb uses heuristic evaluation measure called category utility to guide construction of tree in order to get the highest category utility.

9. CLASSIFICATION TECHNIQUES

a. C 5.0: The important task of classification process is to classify new and unseen sample correctly. C5.0 is a classifier, which gives efficient classification in less time compare to other classifier. Memory usage is less in generating decision tree.
b. Bayesian algorithm: Bayesian networks are a powerful probabilistic representation, and their use for classification has received considerable attention. Bayesian algorithms predict the class depending on the probability of belonging to that class.

10. ASSOCIATION RULE MINING TECHNIQUE

a. Predictive Apriori Association Rule mining algorithm: In predictive Apriori association rule algorithm, support & confidence is combined into a single measure called predictive "Accuracy". This predictive accuracy is used to generate the Apriori association rule. In Weka, this algorithm generates "n" best association rule based on "n" specified by the user.
b. Apriori Hybrid: this algorithm is combination of Apriori and AprioriTid. This combination is formed to remove disadvantages of mentioned algorithm so ultimately its performance is better than those.

11. CONCLUSION

Nowadays, large quantities of data are being accumulated. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. In Exploratory Data Analysis, some initial knowledge is known about the data, but Data Mining could help in a more in-depth knowledge about the data. Through this study concluded that, it creates a bottleneck for large data analysis and gaining knowledge about data mining tools.

REFERENCES

[1] Mahesh Borhade, Preeti Mulay, " Online Interactive Data Mining Tool", 2nd International Symposium on Big Data and Cloud Computing, procedia Computer Science 50(2015), PP: 335-340.

[2] Surbhi Sharma, Dinesh Soni, Arvind K Sharma, "Explorative Study of Web Data Mining Techniques and Tools: A Review", International Journal of Computer Science and Technology, Vol.8, Iss. 1, Jan-March 2017, PP: 43-47.

[3] Komathi, Ramya, Shanmugapriya, Sarmila, "A Novel Comparative Study on Data Mining Tools", International Journal of Innovative Research in Computer and Communication Engineering", Vol. 4, Iss. 11, November 2016, PP:19118-19122.

[4] Anil Sharma, Balrajpreet Kaur, "A Research Review on Comparative Analysis of Data mining Tools, Techniques and Parameters", International Journal of Advanced Research in Computer Science, Vol. 8, No. 7, July-August 2017, PP:523-527.

[5] Rajni Jindal and Malaya Dutta Borah, "A Survey on Educational Data Mining and Research Trends", International Journal of Database Management Systems, Vol. 5, No.3, June 2013, PP:53-73.

[6] Hashmi, Ahmad, "Big Data mining: Tools & Algorithms", IJES, Vol.4, Iss. 1, 2016, PP:36-40.

[7] Prithvi Bisht, Neeraj Negi, Preeti Mishra, Pushpanjali Chauhan, "A Comparative Study on Various Data Mining Tools for Intrusion Detection", International Journal of Scientific & Engineering Research, Vol. 9, Iss. 5, May 2018, PP: 1-8.

[8] Sarumathi, Shanthi, Vidhya, Sharmila, "A Review: Comparative Study of Diverse Collection of Data Mining Tools", World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering, Vol. 8, No. 6, 2014, PP:1028-1033.

[9] Sridevi, Kanagaraj, "A Survey of Data Mining Techniques and Tools", International Journal of Advanced Research in Computer and Communication Engineering", Vol. 6, Iss. 9, September 2017, PP: 158-161.

[10] Kulwinder Kaur, Shivani Dhiman, "Review of Data mining with Weka Tool", International Journal of Computer Science and Engineering, Vol. 4, Iss. 8, 2016, PP: 41-44.

[11] Varun Malik, Sanjay Singla, "A Performance Analysis of Web Usage Mining Tools", International Journal of Research in Advent Technology, Vol. 7, No. 1, January 2019, PP: 291-296.

AUTHOR PROFILES

Mrs. K. Thilakavalli, She has completed her Bachelor of Physics degree in Bharathiar University in the year of 2006. She completed her Master of Computer Applications degree in Anna University in the year 2009. She completed her Master of Philosophy in PRIST University in the year of 2010. She has 8 Years 6 months experience in the field of Teaching. She published 15 International Journal papers, three International Conference papers, and two National Conference papers. Her research area of Interest is Image Processing, Computer Networks (Wireless & Wired), Mobile Computing, Data Mining. Currently she is working as an Assistant Professor in the Department of Computer Applications in Vivekanandha College for Women, Tiruchengode.

Dr. R. Umadevi, her research area of Interest is Computer Networks (Wireless & Wired), Image Processing, Mobile Computing, Data Mining. Currently she is working as a Professor & Head, Department of Computer Science and Computer Applications in Vivekanandha College for Women, Tiruchengode.

Mr. D. Rajagopal, he has completed his Bachelor of Computer Science degree and completed his Master of Computer Applications degree in Periyar University in the year 2003 and 2006 respectively. He completed his Master of Philosophy in PRIST University in the year of 2012. He has 3 Years and 10 Months Experience in the field of Software Developing and 11 Years and 6 months Experience in the field of Teaching. He published 19 International Journal papers and a National Conference paper. He delivered more than 10 seminar & training programs for different Academic Institutions. His research area of Interest is Computer Networks (Wireless & Wired), Image Processing, Mobile Computing, Data Mining, Software Programming (OOPS). He is the life time member of MISTE, MIAENG. Currently he is working as an assistant professor in the department of computer science, Vivekanandha College of Arts and Sciences for women, Tiruchengode.