ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

A Process for Implementation of Data Preprocessing In Weka & Datapreparator Tool

[1] Vani .N, [2] Dr. Veeragangadhar Swamy .T.M
[1]Assistant Professor, [2] Professor
[1][2] Dept of CSE, R.Y.M.E.C, V.T.U, Ballari, Karnataka

Abstract - **Data Preprocessing is important task in today technology to convert raw data into knowledgeable format from web log file which helps identify individual user and session, path completion and removes inconsistent data, noisy data and irrelevant data with the help of data cleaning process. We propose tools for data cleaning process such as Weka and DataPreparator to execute cleaning data which results efficient analysis of raw data in to useful information and improves quality of data. In this paper, we use Weka and Data Preparator tool for data cleaning, visualization, integration data and produces the needy data for mining.**

Keywords: Data preprocessing, Weka, Data Preparator.

1. INTRODUCTION

The Data Preprocessing is most interesting methodology to remove noise, irrelevant data from web server log and helpful for many application such as search engines, Online education and Business Intelligence which helps for increasing resource utilization, reduces search time and making business decisions to predict user behavior based on statistics with the help of mined patterns.
The data preprocessing stage transforms raw data into efficient data representation which removes irrelevant and inconsistent data. The data in internet generally consists of Incomplete, Noisy, Inconsistent data.

a)Incomplete: data consists of missing attribute values, missing certain attribute or aggregate data.
b)Noisy: data consists of errors.
c)Inconsistent: contains mismatch data, lack of similarity in names and codes. The Weka and Data Preparator is tool used for preprocessing data from web log file to clean and remove noisy data, inconsistent data and gives cleaned data required for mining application as result.

Motivation: since, cleaning of web log data is very much essential for getting accurate patterns needed for web mining to predict user behaviors. Using modernized tools for cleaning will help data cleaning process, which is not addressed in literature survey made us to take up this work.

Contribution: In this work, the following contributions are made:
1. Implementation of data preprocessing using Weka and analyzing the results.
2. Implementation of data preprocessing using Data Preparator tool and analyzing the results.

Organization of the paper:
This paper is organized as fallows. Section 2 describes literature survey. The section 3 describes data preprocessing, section 4 describes about implementation of Weka and data Preparator and section 5 describes about conclusion.

2. LITERATURE SURVEY

- P.Sukumar [3] focuses on structure of web log file and describes about algorithm for parsing web server log, cleans data and filters information and uniquely identifies user and session which results to cleaned log file including details of bandwidth and hits.
- Dr. Sanjay Kumar Dwivedi [6] describes about different phases of data preprocessing stages and identifies contents of log file and performs cleaning data, identification of user and session.
- Vijayashri Losarwar [9] describes about analysis matrix for preprocessing. the data collection is done from server side, client side, proxy side and performs cleaning data and identifies user.
- Wasvand Chandrama [10] focuses about categorization in text, report generation and results in identifying user with IP Address, cleaning data, style removing and categorizing data.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

- Jothi Venkateswaran.C [11] describes records in web log and algorithm for cleaning data,identifying user, session, path completion and results about analyzing data in web log file.
- Theint Theint Aye [13] proposing algorithm for field extraction, data cleaning. It increases speed in extraction of data.
- Surbhi Anand [14] describes an algorithm for data fields extraction, algorithm for storing data, cleaning data which reduces log file size and removes unnecessary entries in the file and helps in reducing size of file.

3. DATA PREPROCESSING

It is the first stage in web usage mining. Where user data is collected and stored in log file of web. The log file has date, time, IP Address and message. The data cleaning process gets data from web log file and process the data for cleaning purpose.

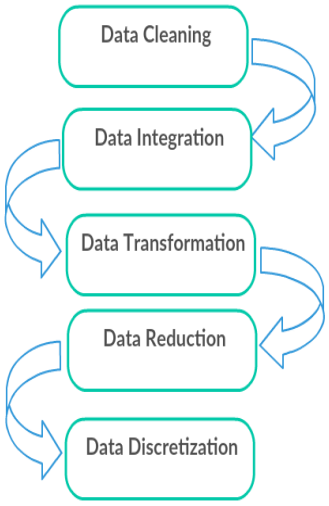

Fig: 1 Data Preprocessing Phases

There are 5 phases in Data pre-processing process.
Data Cleaning: It performs cleansing and scrubbing which is required for the source system which contains "dirty data" that must be cleaned. It fills missing data values, smooth data, removes noisy data.
There are steps in data cleaning process such as:
i) Parsing: In parsing, single data elements will be identified from source file and sends these data in to target file.
ii) Correcting: it uses different data algorithms to correct single data.
a) Integrating data: data collected from various sources will be combined and sent to data store.

There are issues in data integration such as:
• Integration of schema.
• Detecting data & resolving data value conflicts.
• Redundant data occur when integration of multiple database.

a) Data transformation:
Transformation process deals with rectifying and inconsistency techniques and helps to remove noise using smoothing technique, identifying useful information using aggregation technique, generalizing & normalizes data helps to find patterns and constructs attribute to know valuable information. Data transformation transforms data into useful pattern.

b) Data Reduction:
This technique helps to reduce data size and finds data sets and maintains quality of data. it removes irrelevant information.

c) Data Discretization:
It divides data attributes in to intervals which help to reduce values. It uses split and merge techniques. It replaces actual value of data.

4. IMPLEMENTATION

a. Data preprocessing in Weka
It is a portable tool and implemented in java. This software is useful for machine learning research studies. Weka includes the features like explorer, knowledge flow, experimenter, simple CLI. It performs preprocessing, classifying, clustering, associating, selecting attributes and visualizing results.

Steps to be fallowed:
1)Install weka tool which is open source freely available.
2)Select open file at left most top button and select sample file.
3)Choose discretize –the first last precision 6 method, which is helpful for preprocessing.
4)Select apply button to apply method for preprocessing.
5)To view results, select Visualize All button to view performance graph.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

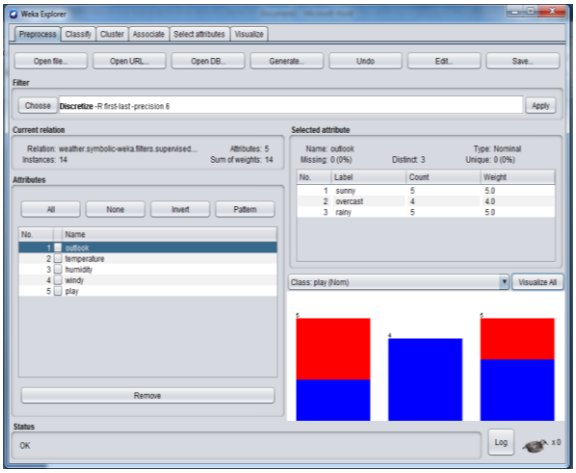

Fig 2: Data cleaning process in Weka

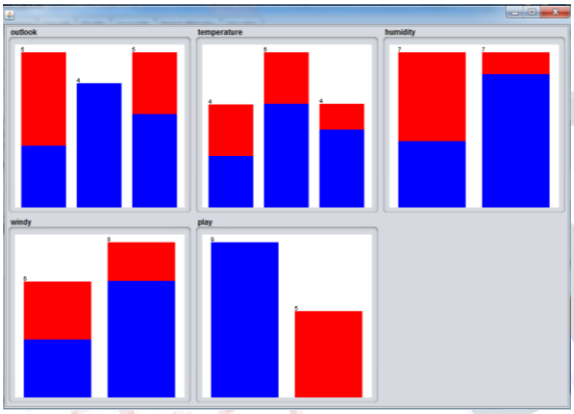

Fig 3: Performance graph of Data cleaning process in Weka

Results and discusions:

• The blue color code indicates count of climate temperature value.

• The red color code indicates weight assigned for each each paramter.

It results cleamed log file and visualizes result for each pareameter of windy,sunny and soon.

b. Data preprocessing in DataPreparator: It is open source software which specially used for data preprocessing. It contains sample web log file to perform preprocessing activity.

It contains 4 main categories such as:

i) ATTRIBUTES
ii) RECORDS
iii) UTILS
iv) OUTPUT
v) EXIT

i) ATTRIBUTES: it performs eight functions based on user requirement such as:

•Discretize
•Handle Missing
•Handle Outliers
•Numerate
•Reduce Labels
•Scale
•Select Attribute
•Delete/Move

ii) RECORDS: there are two types of records available such as:

•Sample record
•Select record

iii) UTILS: it has 2 fields such as:

•File Utils
•Integrate data

iv) OUTPUT: It shows results various types of formats such as:

•Statistics
•Tables
•File
•Database
•Excel
•Visualize

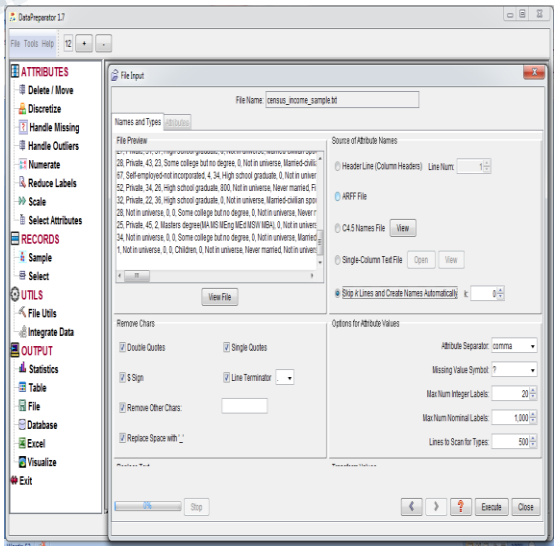

Fig 4: Input File in DataPreprocessor

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

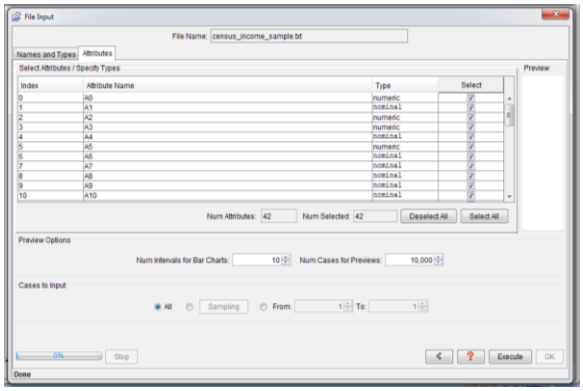

Fig 5: Identifies Attributes in data cleaning process

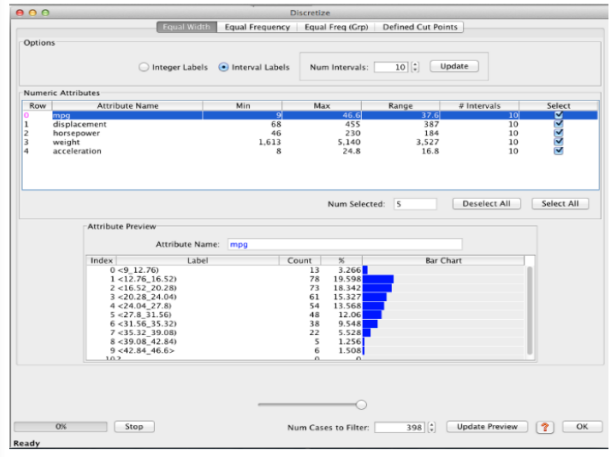

Fig 6:Results of data preparator

Result and discussion:it generates bar graph and gives cleaned log file as result.

5. CONCLUSION

Data preprocessing techniques is useful in machine learning and data mining applications. We use two tools for implementation of data preprocessing phase such as Weka and DataPreparator. Weka performs data cleaning efficiently and gives visualization effect as result and removes noisy data and inconsistent data. DataPreparator reads sample training data from web log file and removes single quotes, double quotes, missing values to remove irrelevant data. The results of Weka and data preparator process web log file and gives cleaned log file with the reduction of data and transform to the required format. Weka & DataPreprocessor tools can be used for fast and efficient process related to data cleaning, data reduction, integration, discretization and transformation of the data preprocessing phase with visualization of preprocessed data in the web mining process.

REFERENCES

1. Sahaj Chavda, Saurabh Jain, Nikunj Panchal, Manisha Valera,(2017) Recent Trends and Novel Approaches in Web Usage Mining on International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 04, e-ISSN: 2395 -0056.

2. Surbhi Sharma, Dinesh Soni, Dr. Arvind K Sharma(2017) Explorative Study of Web Data Mining Techniques and Tools: A Review on International Journal of Computer Science And Technology, IJCST Vol. 8, Issue 1, ISSN : 2229-4333.

3. P.Sukumar,L.Robert, S.Yuvaraj(2016) Review on Modern Data Preprocessing Techniques in Web usage mining (CSITSS) Oct2016,DOI:10.1109/CSITSS.2016.7779441

4. Tanya Bhattacharya, Arunima jaiswal,Vaibhav nagpal (2016) Web usage mining and text mining in the environment of web personalized for ontology development recommender system: International Conference in Reliability, Infocom Technologies and Optimization(ICRITO),DOI:10.1109/ICRITO.2016.7784930

5. Suharjito, Diana, Herianto (2016) Implementation of classification technique in web usage mining of Banking company: International Conference in 2016 International Seminar on Intelligent Technology and ItsApplications(ISITIA),DOI:10.1109/ISITIA.2016.7828660.

6. Sanjay Kumar Dwivedi, Bhupesh Rawat(2015) A review paper on data preprocessing: A critical phase in web usage mining process In: International Conference on Green Computing and Internet of Things (ICGCIoT),DOI: 10.1109/ICGCIoT.2015.7380517.

7. Zhang Huiying, Liang Wei, An Intelligent Algorithm of Data Pre-processing in Web Usage Mining, 5Ih World Congress an Intelligent Control and Automation.

8. R. M. Suresh; R. Padmajavalli, An Overview of Data Preprocessing in Data and Web Usage Mining,1st International Conference on

DigitalInformationManagement,DOI:10.1109/ICDIM.2007369352.

9. Vijayashri Losarwar, Dr. Madhuri Joshi, Data Preprocessing in Web Usage Mining in International Conference on Artificial Intelligence and Embedded Systems (ICAIES), Singapore.

10. Wasvand Chandrama, Prof. P.R.Devale, Prof. Ravindra Murumkar (2014) Data Preprocessing Method of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.

11. Jyothi Venkateswaran C.Sudhamathy G. (2015) Pre Processing of Web Logs – An Improved Approach For E-Commerce Websites in International Journal of Engineering and Technology (IJET)..

12. Satya Prakash Singh Meenu (2016) Web Usage Mining Tools & Techniques: A Survey in International Journal of Advance research , Ideas and Innovations in Technology IJARIIT.

13. Theint Theint Aye, Web Log Cleaning for Mining of Web Usage Patterns in 3rd International Conference on Computer Research and Development, DOI: 10.1109/ICCRD.2011.5764181.

14. Surbhi Anand, Rinkle Rani Aggarwal, An Efficient Algorithm for Data Cleaning of Log File using File Extensions in International Journal of Computer Applications (0975 – 888) Volume 48– No.8, June 2012.

15. Arshi Shamsi, Rahul Nayak, Pankaj Pratap Singh, Mahesh Kumar Tiwari, Web Usage Mining by Data Preprocessing in International Journal of Computer Science And Technology.

16. Rohit P R, Sushant Kumar, Prakasha.S, G.T.Raju (2013) Improving Browsing Experience Through Query Optimization Personalized Result Restructuring in International Journal of Advanced Computational Engineering and Networking, ISSN: 2320-2106.

17. Nawal Sael,Abdelaziz Marzak, Hicham Behja(2013) Web Usage Mining Data Preprocessing And Multilevel Analysis On Moodle In ACS International conference on computer system and applications,DOI:10.1109/AICCSA.2013.6616427.

18. Prathamesh S Tugaonkar,Vidya Chitre(2016) Survey on recent methodologies used for recommender system In: International Conference on Electrical, Electronics,andOptimizationTechniques(ICEEOT),DOI:10.1109/ICEEOT.2016.7754991.

19. Bhupendra Kumar Malviya; Jitendra Agrawal (2015) A Study on Web Usage Mining Theory and Applications In: Fifth International Conference on CommunicationSystemsandNetworkTechnologies,DOI:10.1109/CSNT.2015.247

20. Neha Goel ,Banasthali Vidyapith (2015) Preprocessing web logs: A critical phase in web usage mining In: IEEE , 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA),DOI:10.1109/ICACEA.2015.7164776

20. Monika Dhandi, Rajesh Kumar Chakrawarti (2016) A comprehensive study of web usage mining In: Symposium on Colossal Data Analysis and Networking (CDAN), DOI: 10.1109/CDAN.2016.7570889.

21. V.Anitha, Dr. P. Isakki (2016) A survey on Predicting User Behavior based on web server log files in a web usage mining: International Conference on Computing Technologies and Intelligent Data Engineering(ICCTIDE),DOI:10.1109/ICCTIDE.2016.7725340.

22. Ravi Khatri, Daya Gupta(2015) An Efficient Periodic Web Content Recommendation based on web usage mining in international conference in Recent Trends in Information Systems (ReTIS) .