A Comparative Analysis on Data Pre-processing Tools

Dr. S.Abdul Rahaman
Assistant Professor
King Faisal University
Kingdom of Saudi Arabia

Abstract **-** Data pre-process is one of the activities in Data mining. It will cover 80% of the processes involved in knowledge extraction. The study has applied a systematic approach and found WEKA, R- Programming, and Microsoft Excel as familiar tools for pre-processing of data. These tools have better functionality to study the relationship between data. Microsoft Excel is not an efficient data analytics tool, but able to a large volume of data without any complexity. It is a flexible tool to perform manual pre-process of complex data. The aim of the study is to compare the tools and reveal their merits and demerits. WEBKB dataset and weblogs were used in the study to evaluate the performance of tools.

Keywords **-** Microsoft Excel, Data Pre-process, WEKA, R-Programmer, Data Pre – process

I. INTRODUCTION

Data Mining (DM) is a set of techniques to extract knowledge from large volume of Data[1][2][3]. Data Pre – process (DP) is the backbone of DM techniques. Many researchers have proved that a better DP can lead to a better knowledge extraction. Data cleansing, Data Integration, Data discretization, and Data Transformation are the major activities of DM. DP is a set of Data processing activities that help DM technique to produce useful patterns from data corpus.

A. Data Cleansing - It is a process of removing unwanted data from data corpus. Removal of outliers will help to fix the boundary to dataset[4]. Replacement of missing value is also an important task in data cleansing activity[5][6]. Decision tree and Bayesian rules are handy for the replacement of missing values. The final step in data cleansing is the correction of inconsistent data. Careful studies on data provide solution to correct inconsistent data.

B. Data Integration - It is a process of combining data from different sources and store as a single set. It will give confidence to user that the data is reliable. It will make a global schema to store heterogeneous data[7][8]. The mapping unit will map heterogeneous data with global schema.

C. Data Transformation - It is an important phase in DP. Data has to be transformed into a computable format. DM technique have different logic to process data[9][10]. Data normalization, aggregation, and generalization are the activities involved in data transformation.

D. Data Reduction - It is a process of reducing the size of dataset without losing valuable information. Techniques such as Binning, Clustering, and Aggregation are use to minimize the size.

E. Data Discretization - It is a subset of Data Reduction. It will use nominal values in the place of numerical attributes. Data will be more generalized and too specific with the help of Data Reduction.

The objective of the study is to compare the familiar pre processing tools and evaluate the performance using datasets[11][12][13].

The structure of the study is as follows: The following section will provide details about the selection procedures for DP tools. The section 3 will give details about Microsoft Excel, R – programming, and Waikato Environment for Knowledge Analysis (WEKA) tools. Experiments and results will be discussed in section 4. Finally section 5 will conclude the study.

II. SELECTION PROCEDURES FOR DP TOOLS

Systematic approach was followed in the study to select the familiar DP tools. Keywords such as " Data Pre - process", " Data Mining", "Knowledge Extraction", "Pattern Recognition", and "Data Cleansing" are used to collect literatures on Data pre – process.

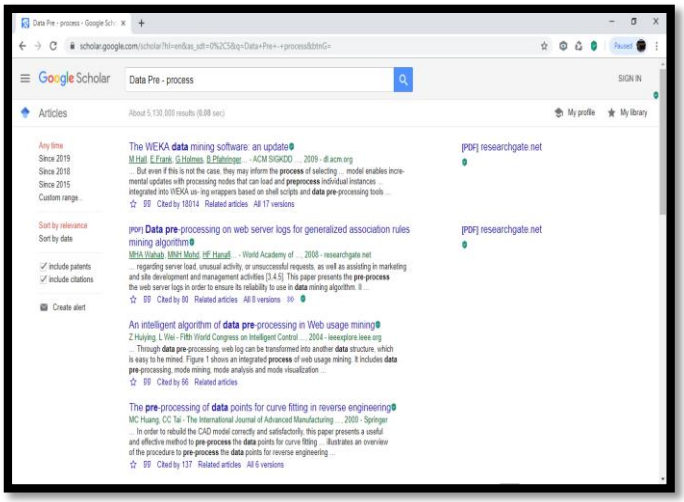

Figure 1: Collection Procedure – Google Scholar

IEEE Explore[i], Springer, and Google Scholar were the prime portals for the collection of research materials related to keywords. A total of 30 research articles were collected from the portals. The period of publication for the research articles were from year 2013 to 2018. The research also gave importance for number of citations for the research. If a research fell before the mentioned period and having high number of citations then the research was included in the

study. After applying these criteria, 10 articles were excluded from the initial collection and 5 articles were included into the collection and made a reasonable total of 25 articles. Table 1 shows the details of research articles after applying criteria.

Table 1: Details of research articles

S.No.	Portals	Citation	No. of Articles
1	Google Scholar	> 12	7
2	IEEE Explore	>20	10
3	Springer	>20	8

III. DP TOOLS

A. Microsoft Excel (Excel) - It is a handy tool to work with data. The application is offering many useful features to pre-process data. User can write code to control data[15][16][17]. The code will be stored as macro. The macro can be triggered at any time and level. Visual Basic for Application (VBA) is the language used for developing macro.

Statistical Package for Social Science (SPSS) is a most powerful tool comparing to Excel. User cannot employ SPSS to pre – process data. Excel will offer many functions to format data into desirable form.

It is a basic level to pre – process complex data for Machine Learning (ML) algorithms. Many advanced tools have existed but Excel is more preferable for data scientist. A new stable version of Excel was released on February 12, 2019.

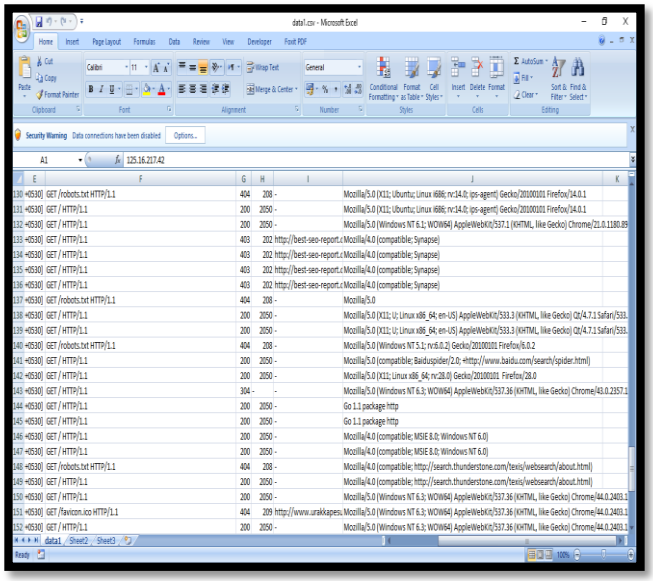

Figure 2: Microsoft Excel – Data Pre-process

B. R – Programming - R is a statistical computing tool. It is widely used for extraction of knowledge from data. It is a freeware, which is under GNU General Public Licence[ii]. It has secured 15th rank in the assessment of popular programming language.

The programming environment is providing a command line interface to execute functions in R. User has to learn the functionalities and commands to excel in R. It will support all kinds of programming such as procedural and object oriented with generic functions[19][20][21][22]. User can[iii] create their own packages in Java, C and C++. The fig. 3 will show the R – studio and fig. 4 shows the R – programming environment. The latest version of R is version 3.5.2. (Eggshell Igloo).

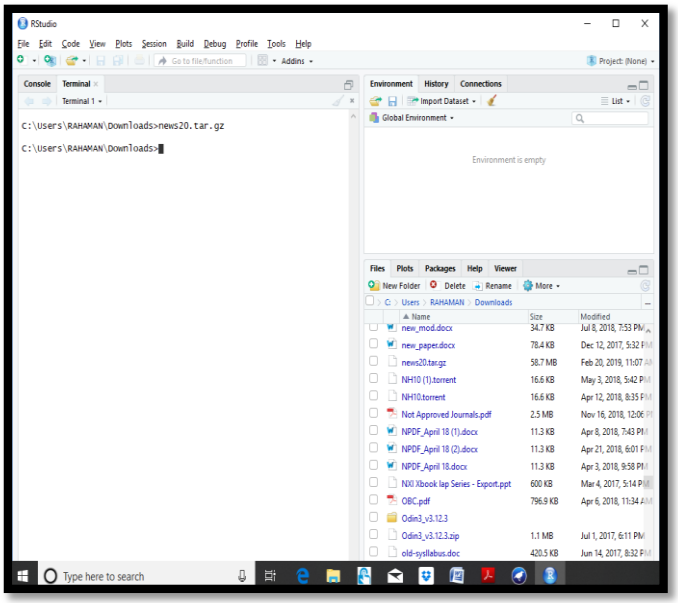

Figure 3: R-Studio

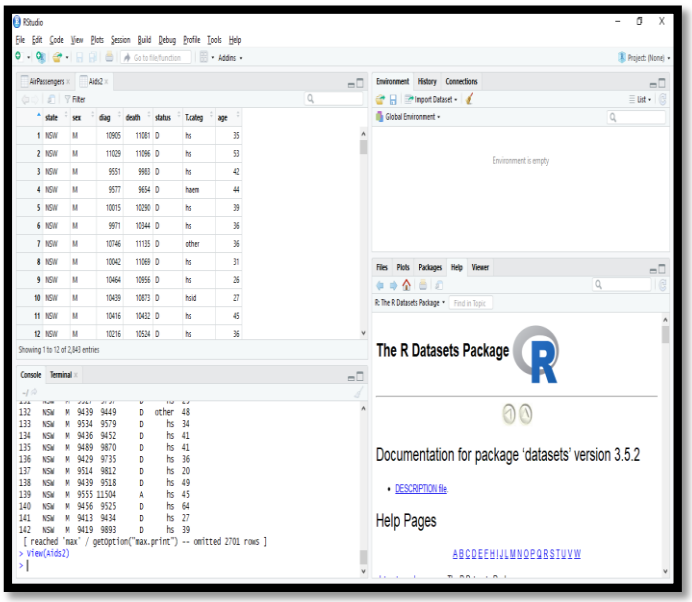

Figure 4: R-Programming Environment

C. WEKA - It is a freeware, developed by University of Waikato. It is offering a graphical interface for users to analyze data[23][24][25]. WEKA model was implemented in Java and thus operate on different platforms.

ML techniques like clustering, classification, and regression can be implemented in WEKA model without any difficulties[26][27]. Dataset has to be loaded as a single file into WEKA for the analysis. Visualization model will help

to understand the relationship between data. DT can be performed by WEKA. The latest version is 3.8.3 released on September 4, 2018. Fig. 5 shows the interface of WEKA.

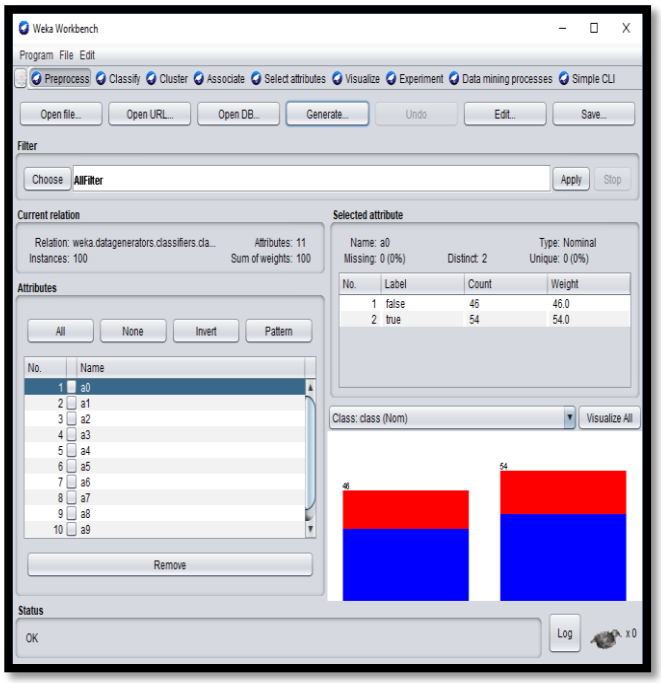

Figure 5: WEKA interface

IV. EXPERIMENTS AND RESULTS

In order to show the ability of DP tools, a variety of datasets were used in the study. World
Web Knowledge base (WEBKB) dataset of Carnegie Melon University (CMU)[iv] comprises of 4 Universities Cornell, Texas, Washington, and Wisconsin data. 20 Newsgroups dataset is also from CMU contains 20,000 online messages from different news groups. And, finally a weblog of www.rahablog.com (rahablog) was also used in the study. Fig. 6 shows the screenshot of rahablog.

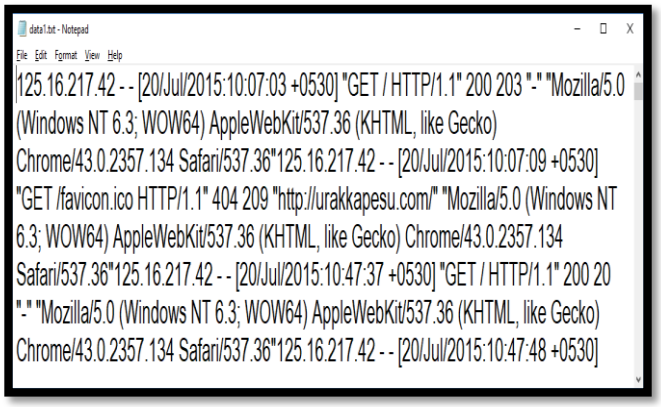

Figure 6: rahablog

Fig. 7 illustrates the flowchart of the experimented, which was conducted in the study. The study has used the raw data of WEBKB and matched with classified dataset of CMU. Rahablog data was pre – processed and classified into three users such as Synthetic, Potential, and Normal users. The raw data was cleaned with specific rules according to the ML technique. The process of tokenization will help to label the tokens and lead to DT. After DT, data will be in a computable format for ML technique. The following section will show the results generated by the study.

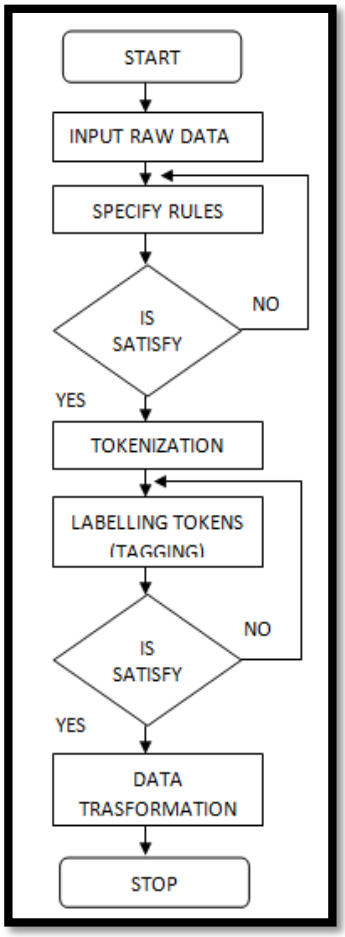

Figure 7: Flowchart – DP

Table 2 shows the details of pre – processing of University dataset. CMU column in the table indicates the data provided by University. They have pre – processed the data and classified into different categories mentioned in the table 2. The fig. 8 shows the relevant graph of table 2. The graph indicates that all tools have reached the optimum level like CMU.

Table 2: WEBKB - University Dataset

Classification / Tools	CMU - Tool	Excel	WEKA	R
Student	1641	1583	1602	1610
Faculty	1124	1134	1201	1194
Staff	137	102	145	109
Department	182	214	165	225
Course	930	1045	908	927
Project	504	490	512	520
Other	3764	3714	3749	3697

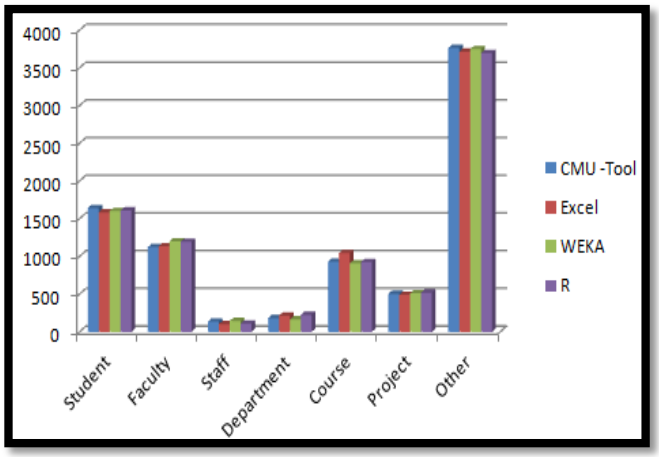

Figure 8: WEBKB - University Dataset Comparison

Table 3 shows the details of pre – processing 20,000 messages of 20 News groups. 20 News groups data are clustered into 6 different categories. C1 is "Politics", C2 is " Entertainment",C3 is " Social Welfare", C4 is "Education" C5 is "Health", and C6 is "Miscellaneous".

The results have proved that Excel is deviated from WEKA and R. Excel has limited functionalities, so accuracy level is not like other tools. Formation of Rules is difficult and complex in Excel. The computation time of Excel is more comparing to WEKA and R. The performance of Excel was better in University dataset because of rules formation.

Table 4 shows the weblog details of rahablog. The weblogs are pre – processed into 3 different categories [18]. The fig. 10 shows the relevant graph of Table 4.

Table 3: 20 Newsgroups Dataset

Classification / Tools	Excel	WEKA	R
C1	1400	2148	2700
C2	5064	5384	4900
C3	4700	5300	4989
C4	1789	1314	1142
C5	4132	3850	3947
C6	2814	1980	2045

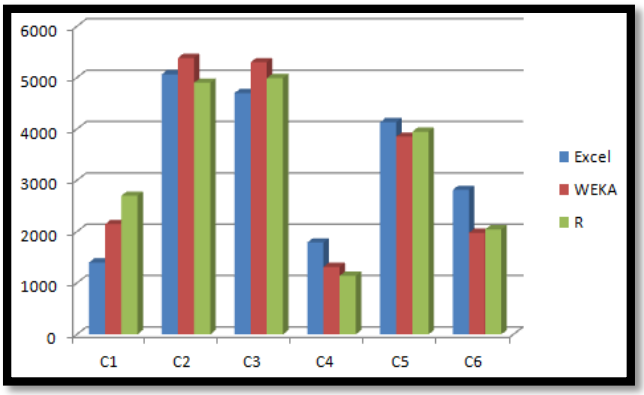

Figure 9: Comparison of DP tools – 20 News groups

Table 4: Weblog - rahablog

Classification / Tools	Excel	WEKA	R
Synthetic	450	752	587
Normal	1487	1854	1754
Potential	1785	1356	1524

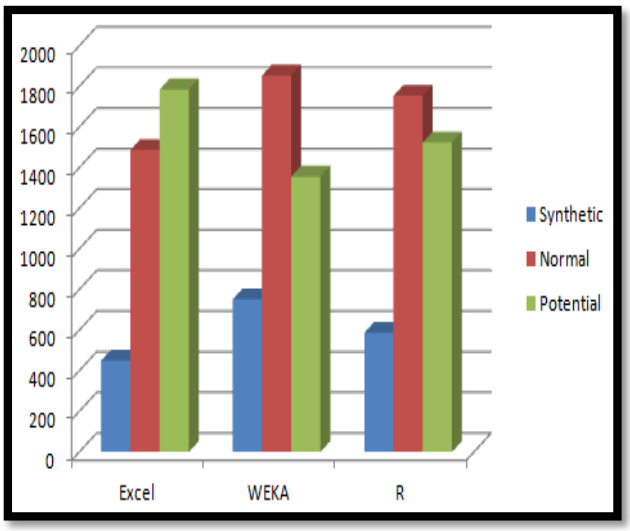

Figure 10: Comparison of DP tools - Rahablog weblog

Finally, Table 5 shows the comparative analysis of pre-processing tools. Different attributes were taken into consideration to derive the performance of DP tools

Table 5: Comparative analysis of DP Tools

Tools / Attributes	Efficiency	User Interface	Analytic Level	Programming Knowledge	DP	DT
Excel	Medium	High	Medium	Low	Medium	Medium
R	High	Medium	High	High	High	Medium
WEKA	High	Medium	High	High	High	High

V. CONCLUSION

Pre-process of data is a part of the process in the extraction of data. The study has followed a systematic way to find familiar pre-process tools

in data mining. Excel, WEKA, and R are familiar tools to perform data analysis. WEBKB datasets of Universities and 20 newsgroups were used in the study for the evaluation of the performance of data pre-process tools. In addition, a weblog of rahablog was also employed in the study. The study has found the following facts. Excel has achieved an approximate of 90 % of accuracy in University and 20 newsgroups dataset. WEKA has reached 96% and R has achieved 97%. Excel is basically a spreadsheet program, not having a good functionality like WEKA and R. The computation time is also more in Excel. A user need not to have better programming skills in Excel. In the case of WEKA, and R, the user should have good programming knowledge to derive a relationship in a large amount of data.

VI. REFERENCES

[1]. Amir E.A.D. et al. (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol ., 31, 545-552.

[2]. Anders S. et al. (2015) HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics , 31, 166-169

[3]. Angerer P. et al. (2015) destiny: diffusion maps for large-scale single-cell data in R.

[4]. Buettner F. et al. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol ., 33, 155-160.

[5]. Campbell K., Yau C. (2016) Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis. bioRxiv , 060442.

[6]. Chikina M. et al. (2015) CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. Bioinformatics , 31, 1584-1591.

[7]. Fan J. et al. (2016) Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods , 13, 241-244.

[8]. Finak G. et al. (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol ., 16, 278.

[9]. Grün D. et al. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature , 525, 251-255.

[10]. Kim D. et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol ., 14, R36.

[11]. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, C. Bouhaddioui, "Big Data Quality: A Quality Dimensions Evaluation", 2016 Intl IEEE Conf on (UIC/ATC/ScalCom/CBDCom/), pp. 759-765, 2016.

[12]. Taleb and M. A. Serhani, "Big Data Pre-Processing: Closing the Data Quality Enforcement Loop," 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI, 2017, pp. 498-501. doi: 10.1109/BigDataCongress.2017.73

[13]. Ikbal Taleb, Mohamed Adel Serhani, Rachida Dssouli, "Big Data Quality: A Survey", Big Data (BigData Congress) 2018 IEEE International Congress on, pp. 166-173, 2018.

[14]. Umar Aftab, Ghazanfar Farooq Siddiqui, "Big Data Augmentation with Data Warehouse: A Survey", Big Data (Big Data) 2018 IEEE International Conference on, pp. 2775-2784, 2018.

[15]. Venkat N Gudivada, Ricardo Baeza-Yates, Vijay V Raghavan, "Big Data: Promises and Problems", IEEE Computer, vol. 48, no. 3, March 2015.

[16]. F.-Z. Benjelloun, Morocco Ibn Tofail Univ., Kenitra, A.A. Lahcen, S. Belfkih, "An Overview of Big Data Opportunities Applications and Tools", IEEE Conference on Intelligent Systems and Computer Vision, March 2015.

[17]. Rul Mao, Honglong Xu, Wenbo Wu, Jiaqiang Li, Yan Li, Minhua Lu, "Overcoming the challenge of Variety: Big Data Abstraction the Next Evolution

of Data Management for AAL Communication Systems", IEEE Communications, January 2015.

[18]. Abdul Rahaman Wahab Sait, Dr.T.Meyappan "An automated web page classifier and an algorithm for the extraction of navigational pattern from the web data", Journal of web engineering, Rinton Press, ISSN: 1540-9589, Vol.16(2017), pp. 126 – 144.

[19]. Ying Zhang, Bin Song, Yue Zhang, Sijia Chen, Algorithms and Architectures for Parallel Processing, vol. 10393, pp. 642, 2017.

[20]. S. Gayathri Devi, M. Sabrigiriraj, "A hybrid multi-objective firefly and simulated annealing based algorithm for big data classification", Concurrency and Computation: Practice and Experience, pp. e4985, 2018.

[21]. Abdul Rahaman Wahab Sait, Dr.T.Meyappan," Improving web contents and detecting malicious activities in websites through web usage mining technique (WUM)", National Conference on MINDS 2014, March 27 - 28th, 2014.

[22]. Abdul Rahaman Wahab Sait, Dr.T.Meyappan "Data preprocessing and transformation technique to generate pattern from the Web log", International conference on Computer Science and Information Systems (ICSIS'2014) Oct 17-18, 2014 Dubai (UAE).

[23]. Ebrahimi, H.; Rajaee, T. Simulation of groundwater level variations using wavelet combined with neural network, linear regression and support vector machine. Glob. Planet. Chang. 2017, 148, 181-191.

[24]. Patle, G.T.; Singh, D.K.; Sarangi, A.; Rai, A.; Khanna, M.; Sahoo, R.N. Time series analysis of groundwater levels and projection of future trend. J. Geol. Soc. India 2015, 85, 232-242

[25]. Sonmez, M.; Akgüngör, A.P.; Bektaş, S. Estimating transportation energy demand in Turkey using the artificial bee colony algorithm. Energy 2017, 122, 301-310.

[26]. Amid, S.; Gundoshmian, T.M. Prediction of output energies for broiler production using linear regression, ANN (MLP, RBF), and ANFIS models. Environ. Prog. Sustain. Energy 2017, 36, 577-585.

[27]. Suryanarayana, C.; Sudheer, C.; Mahammood, V.; Panigrahi, B.K. An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India. Neurocomputing 2014, 145, 324-335