ENRICHED BIG DATA PRE-PROCESSING MODEL WITH MACHINE LEARNING APPROACH TO INVESTIGATE WEB USER USAGE BEHAVIOUR

N. Silpa

Research Scholar,
Department of Computer Science and Engineering,
Centurion University of Technology and Management, Odissa, India
nrusimhadri.silpa@gmail.com

Dr. V V R Maheswara Rao

Professor,
Department of Computer Science & Engineering,
Shri Vishnu Engineering College for Women(Autonomous), Bhimavaram, AP, India
mahesh_vvr@yahoo.com

Abstract

In the present, the web has become the environment to live, learn, entertain, and socialize individually or as a group through digital platforms where users with high aspirations. As a result, investigating the web user behaviour is most active research even in the present and demands re-innovation in potential analytics to provide reliable and quality customized solutions. To perform this, the weblog is the primary source and poses tremendous challenges for the web researchers with complex sequence of processing steps and abundant information of weblog. Further, limited distributed storage models, partial parallel computing techniques, typical identification of appropriate attributes in the weblog analysis demands the high competitive performance models for effective characterization of web users. The importance of pre-processing in the entire process of weblog analysis is so critical while it is popular among researchers, nonetheless, the studies are limited. In addition, existing pre-processing studies focus on elicitation, reduction and transformation of web user usage data individually not comprehensively.

Towards this, the present paper proposes Enriched Pre-processing Model (EPPM) that comprehensively concentrating on all the stages of pre-processing of weblog data in the framework of apache spark. The EPPM enables the capability of processing real time streaming data along with batch data as to sustain the validity of web user behaviour extracted from historical data also requires the strategy of processing real time streaming data. In addition to all pre-processing steps, EPPM integrates a machine learning approach to discard the search engine accessed logs from weblog as they are excessive in noticing the web user behaviour. The performance of EPPM is validated by conducting a series of experiments on a server side weblog data in a standard execution environment. The experimental results are also included.

***Keywords*: Web Analytics, Weblog Pre-processing, Machine Learning, Search Engine Access, Apache Spark, Big Data.**

1. Introduction

Learn to understand and forecast the human behaviour on World Wide Web is a high demand of many industries including Finance Sector, Healthcare, Retail Industry, Law Enforcement, Entertainment, Public Security and safety, to meet their ever changing aspirations. In response to this trend, researchers increasingly turn to advanced big web data analytics in tackling potentially predicting web user usage behaviour. To perform this, weblog is a great distributed resource consists of all the details of web users' activities. However, there is a combination of five V's problems involved in Weblog data as below:

- Weblog - High Volume, consists of petabytes and surprisingly Exabyte of huge data and too big to be processed by the existing state of techniques and technologies

- Weblog - High Velocity, due to the proliferation usage of WWW has led to an unprecedented rate of data which is continually being generated at a pace that is tough handle by conventional systems.
- Weblog - High Variety, is a measure of heterogeneity and challenge for new management technologies
- Weblog - High Veracity, need confidence and trust to process the uncertain data, is major quality concern
- Weblog - High Value, opportunity to extract actionable intelligence from above mentioned typical characteristics of data by an innovative big data technologies and machine learning techniques.

Due to these complex characteristics as shown in Figure 1, yet fundamental limitations remain in data collection, modeling and retrieving knowledge from weblog data. In addition, the other challenge involved in this direction

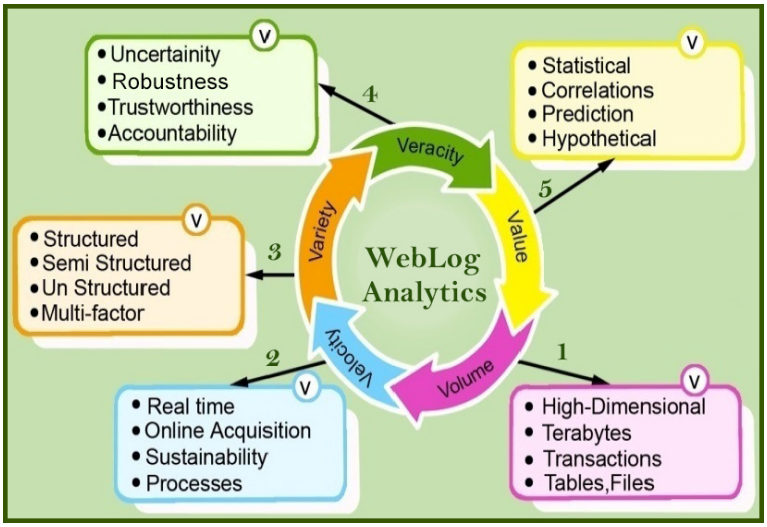

Fig. 1. Characteristics of Weblog Data

is detection and discarding of transactions made by web crawlers from human user access, as web crawlers are software programs that navigates the websites autonomously. These challenges will have been solved efficiently by machine learning integrated big data analytics.

The most challenging issue in this model is consideration of processing typical characteristics of weblog data in a reasonable time. Towards this, Apache Spark is a most suitable big data platform and can execute the machine learning algorithms in an efficient manner.

Spark is designed with a unified processing engine to handle the big data by following the parallel programming model on batch, streaming and interactive data and capable of handling computational challenges by multiple nodes. The spark is employed in many real time applications from web based services to health care systems to security & safety issues.

Among all the stages of big data analytics, preparing weblog data is a critical stage as the data is not in an appropriate state for doing the analysis and it is notoriously impact on the success of complex knowledge retrieving process. Due to the importance of big data preprocessing, in the present paper the authors exclusively concentrates on the stages of big data preprocessing in the framework of machine learning and Apache Spark.

The remaining sections of the paper are presented as follows. Section 2 presented the related work that consists of valuable research contributions made in this area. The detailed novel Enriched Pre-Processing Model (EPPM) is proposed in section 3. In the next section, the achieved experimental results of proposed EPPM are explained. At last, the last section summarizes the concluding remarks and offers future research paths.

2. Related Work

From the past two decades, there has been prominent research has done in the field of web mining to review the usage characteristics of the web user. However, the recent digital shift and advanced technologies demands more innovative accomplishments to structure and analyze the weblog data for analyzing web user usage behavior and it became as a promising present research area. With this motto, the authors have carried out the literature survey over a period of two decades and mention the noticeable points as follows.

The research works [1, 7, 8, 9, 19, 28] have motivated the present researchers to take-up the problem of finding the usage characteristics of web user in the era of big data. To accomplish it, big data Analytics has undergone rapid development in recent decades by several significant research studies [1, 3, 4, 20, 24, 27]. The promising studies [6, 11, 12, 15, 18] reported the characteristics of big data with associated challenges, big data Analytics pipeline, machine learning paradigm with big data in various domains. Particularly the authors [2, 5, 10, 14, 16, 17, 23] have paid attention on deriving functional web usage patterns from big web log data, still, the researchers expressed the need of more attention on data preparation stage in the overall process of big data analytics.

In the same line, some of the research works [21, 22, 25, 29, 34] paid attention on data cleaning, one of the important stage of big data preparation. They find-out the issues and approaches that are involved in cleaning the complex and noise web log data to tackle efficiency and scalability of analytics. The other approaches [29, 30, 32, 33] are devised to address the remaining individual stages of data preprocessing, procedures of parsing of weblog entries [19, 25, 29], feature identification [13, 19 ,25], feature selection [13, 19, 25, 29, 34], user identification [32, 35, 36], sessionization [26, 32, 33] and path completion [32, 33, 36] etc. In addition to that, very few studies [23, 31, 32, 33] identified the necessity of discarding the transactions that are performed by web robots or crawlers, although the investigators strongly recommend the efficient learning algorithms to differentiate humans from web crawlers.

By conducting a deep review on the above mentioned literature, the present study concludes that preparation of weblog data play a vital role in the overall process of data analysis that triggered the proposed work. The novelty of present research study concentrates on all the stages of weblog data preparation and it is multi-objective to filter-out noncontributory data as well as structuring weblog data in an efficient manner to investigate the web user behavior with high accuracy.

3. Proposed Enriched Pre-Processing Model (EPPM)

The correctness and efficiency of any web data processing techniques on complex big weblog data is fully dependent on the pre-processing stage for any application. The investigation and reinvestigation of web user behaviour is mandatory as the web user usage behaviour may change time-to-time. In pursuit of this, traditional pre-processing techniques consume more time, so, intelligent pre-processing techniques are required to prepare rightly useful data.

To achieve the useful data, it is a right recommendation for the web data processing engineers to integrate machine learning approach at the pre-processing stage. It is also advisable to concentrate on historical data along with current streaming data to discover potential insights from rapidly growing weblog data. To derive quality decisions that are close to reality, the authors proposed an innovative Enriched Pre-Processing Model (EPPM) is presented in figure 2. The EPPM is torched on both data collection and pre-processing to generate enriched weblog data to accomplish high performance of web data analytics.

3.1. *Collection of Web Log Data*

The proposed EPPM provides comprehensive solution by inclusion of techniques of data collection and the pre-processing for the formation of enriched weblog. The work also focuses on current challenges involved in

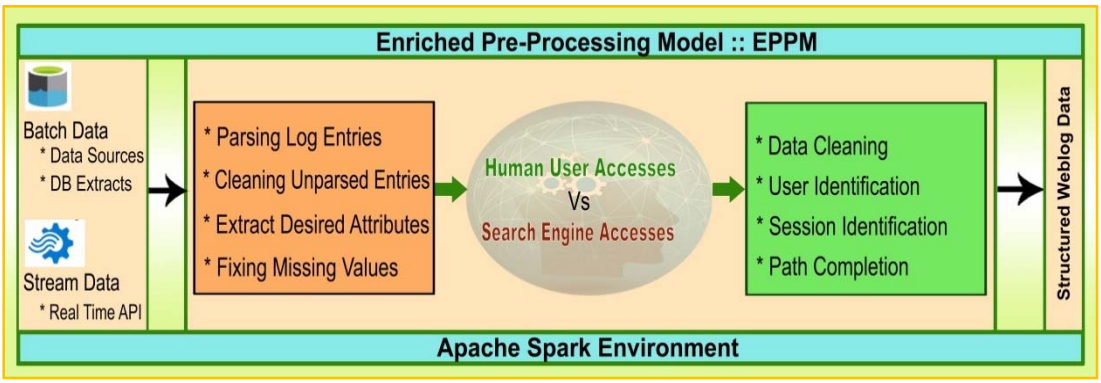

Fig. 2. Proposed Enriched Pre-Processing Model (EPPM)

importing weblog data into the Spark environment incubated with batch and streaming modes as web user usage behaviour changes at any time with a rapid speed.

For this model, the researchers taken the raw weblog of an educational society comprises the usage data of students, academicians, technical and administrative people of engineering, pharmaceutical, medical, pure sciences. In this network, as the weblog grows incrementally the EPPM considers the batch and streaming weblog data and leverage the efficiency by paying attention on the programmed parallelization, speedy recovery of fault, and cooperating with right distributed storage model.

3.2. *Pre-Processing of Web Log Data*

The learning algorithm of proposed model considers each web user click stream data collected at firewall database server of an Educational Institution as an input. To identify rightly and time-to-time web user usage behaviour the

model works on the input in both batch and streaming mode. The algorithm implements with a series of stages in the spark environment to derive structured enriched weblog data in a machine learning nutshell.

Algorithm - Enriched Pre-Processing Model

Input : Raw Weblog Data (RW_D)
Output : Enriched Weblog Data (EW_D)

Stage 1. Start
Stage 2. Load RW_D into HDFS Hybrid Framework
Stage 3. For each RW_D entry in Spark
Do parsing using Regular Expression
Stage 4. End for
Stage 5. For each unparsed RW_D entry
Fix Missing Values
Go to Stage3
Stage 6. End for
Stage 7. Identify task relevant attributes
Stage 8. Build LearningUsageAccessesTree( )
Stage 9. For each Human User Usage Accesses Data
Remove Unwanted Data
Accomplish User Identification
Execute Session Identification
Perform Path Completion
Stage 10. End for
Stage 11. Return EWD
Stage 12. Stop

To achieve real time processing capabilities the stage2 of learning algorithm archives the raw weblog big data available in batch and streaming mode into the HDFS - Spark hybrid framework. The hybrid ecosystem of spark is able to with structured, semi-structured and unstructured nature of real time weblog. The in-memory computation capability of Spark enables the EPPM to achieve high processing speed.

In stage3, the EPPM employs a defined multi-threading and concurrent execution capability of spark job driver to parse the unstructured weblog data into separate features as extracting individual attributes play a great role in categorizing the user usage behaviour. To do this, the authors taken the methodology of proven regular expression based parsing.

In the subsequent stage, the EPPM concentrates on cleaning the remaining unparsed noisy log entries by fixing the missing values and invoke the Spark Job driver to complete the parsing the cleaned data.

Later, to explore desired interesting insights, the EPPM extracts the most relevant attributes from weblog data by applying standard in-built chi-square heuristic function of Spark Machine Learning Library. This also minimizes the induction of accidental correlations by discarding irrelevant attributes. This attribute identification process further helps in visualizing the results of user behaviour.

The researchers built a learning algorithm in the EPPM to implement the LearningUsageAccessesTree( ) by taking the support of MLLibrary of Spark to derive the real insights from the weblog using notable attributes. This learning algorithm is designed on the intrinsic separable identified features in usage log data by web users and crawlers.

The output of the learning algorithm is clearly distinguishable user accessed data and crawler induced data, which is fed to the next phase of the EPPM.

The integration of spark machine learning program of this learning algorithm produces high statistically correlated data with its in-memory computing framework.

The Terminate_Rule( ) of the algorithm have the capability of preparing optimal learning model from the spark imported log from HDFS by scaling with enough training data and fallen within the threshold value.

The Test_Rule( ) uses Attribute_Best_Split_Rule( ) to safeguard goodness and builds the learning tree with the properties of human and crawler accesses as listed in the Figure 3 and Figure 4 by creating meticulous conditional nodes along with proper labels assigned by Assign_Label_Rule.

Tree_Spread _Rule( ) of the algorithm is continuously expand the learning tree by using the Test_Rule( ) upto the scope of Terminate_Rule( ) with assigning rights conditions and labels.

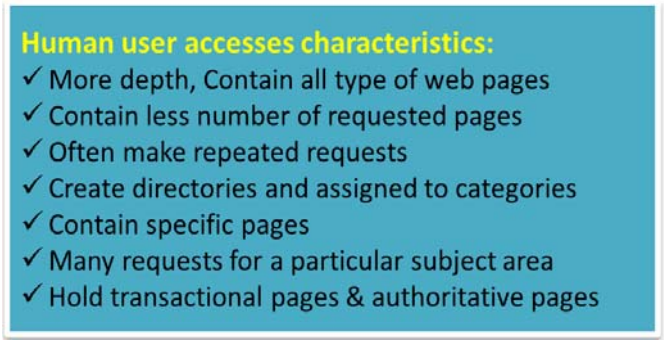

Fig. 3. Characteristics of Human User Accesses

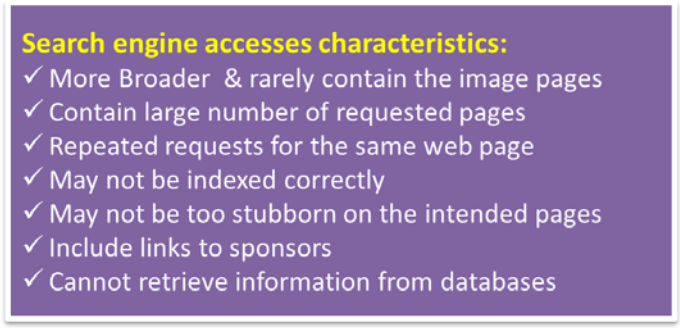

Fig. 4. Characteristics of Web Search Engine Accesses

The Spark enabled proposed Machine learning algorithm in a parallel fashion identifies the attributes and extract knowledge in recursive way of build the decision tree along with the correct assignment of decisions at internal nodes and labels at leaf nodes. A sample learning tree of Stage 8 is presented in the Figure 5.

At the end of this stage, the Spark learning algorithm discarded the search engine accesses to minimize noise and human user usage data is kept ready for further stages of EPPM.

The repeated Stage 9 of EPPM works on identified user usage accesses and effectively performs the remaining

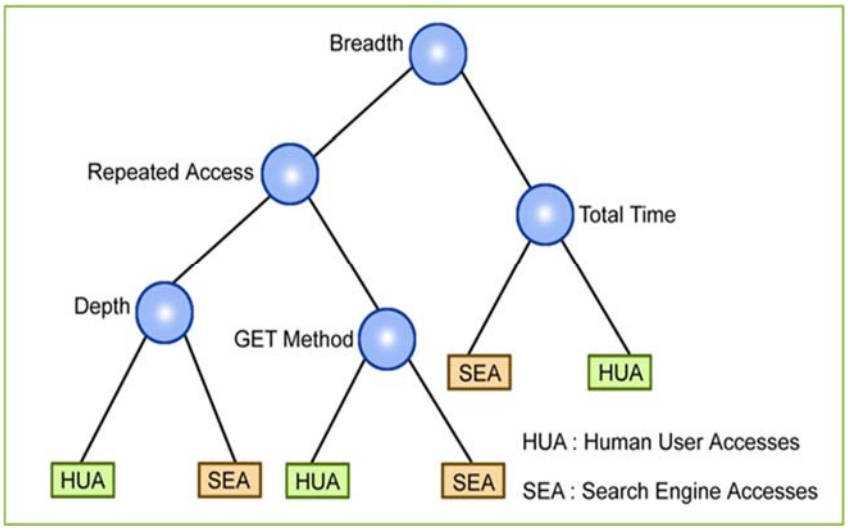

Fig. 5. Sample proposed Learning Tree

organic co-stages using SparkR API. Initially it filters-out unwanted data such as scripts, images, audio, video, etc. files using the filtering functionality of SparkR to proceed with valuable data. And then for identifying the unique users, selection and aggregate functions of the API returns all IP addresses along with accessed log records by following proposed rules,

- Rule1 : Every different IP addresses is considered as different user

- Rule2 : If IP Address is same and Operating System or Browser is different, also be considered as different user.

To accomplish the sessionization process, the EPPM drills the data such as time oriented heuristics with a defined threshold for session duration, duration of stay, website topology etc. by integrating the Spark and its AWS Services. It also pays attention on imperfect datasets, fault-tolerance issues, repeated data etc.

Path Completion is a successive iterative process of EPPM after the sessionization as it is potentially intensive in the whole pre-process. Path completion is a method of adding access to the page that is not in the weblog, but should eventually occur. The authors at this stage took the standard approach for the path completion that is comparison of referral URL with the present URL along with the reference length of the URL so as to include the insights of missed pages also. Figure 6 represents an example of the method.

The EPPM takes the crucial stage of pre-processing of raw weblog Data (RW_D) and prepares the Enriched Weblog Data (EW_D) and which offers notorious influence on the progress of further stages of data processing.

4. Experimental Results

The researchers create execution environment by setting up of apache Spark framework for effective implementation of proposed EPPM with well taken care of necessary computing resources to process growing

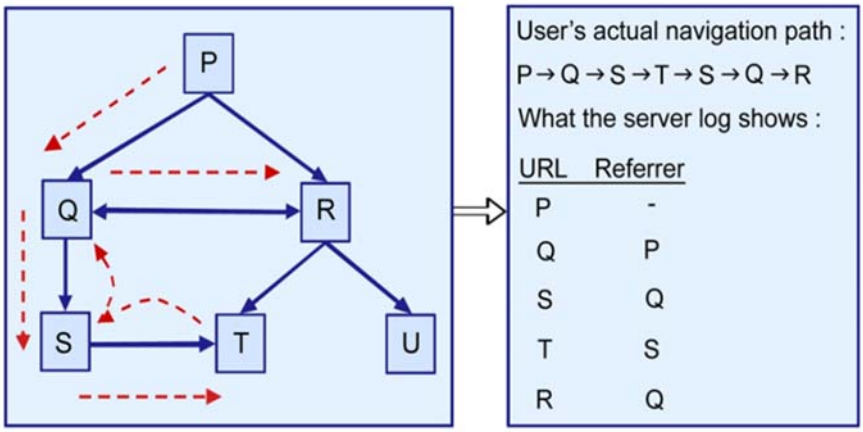

Fig. 6. Example for Webpage Path Completion

weblog.

A raw weblog data on the educational institution server side is collected and registered each month and it is considered for experimental analysis of the proposed research work. For 12 weblogs, a graph is plotted in figure 7, between pre-processed weblog size and raw weblog size, collected per month. The graph shows that for all weblogs, the EPPM consistently decreases the raw weblog size. In relation to computational time, this performance increases the efficiency of further stages of data processing.

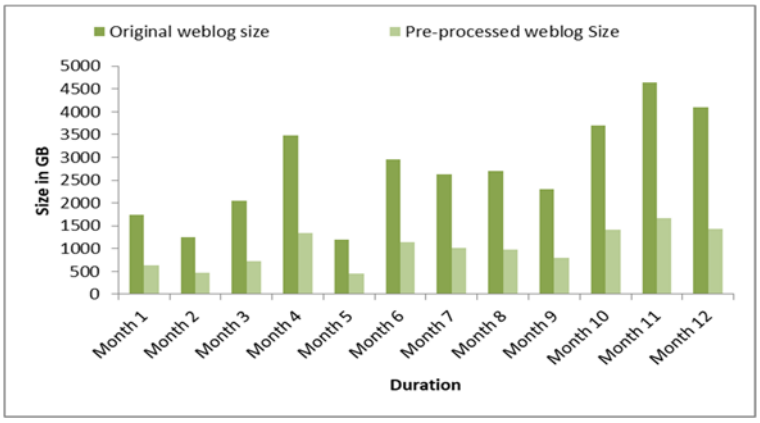

Fig. 7. EPPM Outcome with Decreased Size of Pre-Processed Weblogs

The organic pre-processing stages of EPPM decreases the promising percentage as presented in the figure 8 of the scale of the raw weblog. Thus, as the weblog size increases, pre-processing plays a crucial role in the performance of data processing stage.

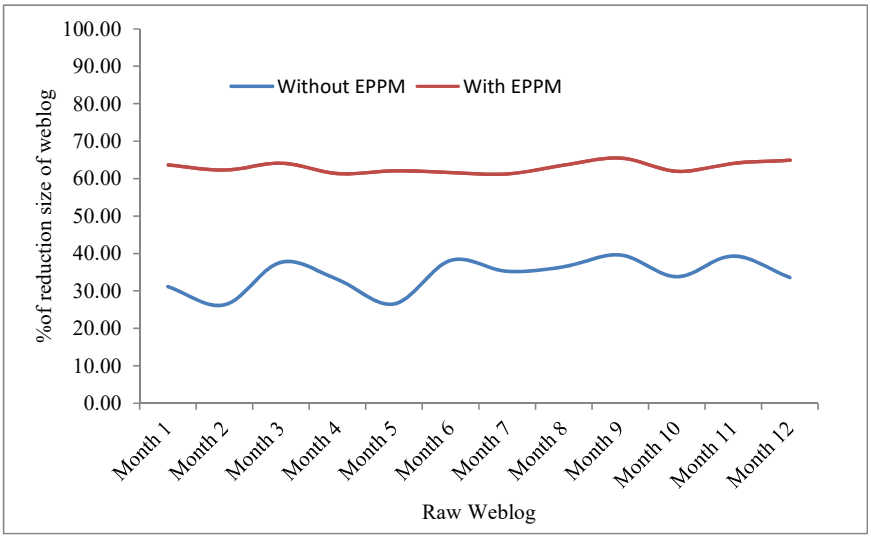

Fig. 8. Reduction size of weblog with and without EPPM

The educational institution weblog is recorded and tabulated for 12 months in column 2 of table 1. The EPPM classified and discarded the access data of the search engine using the machine learning algorithm in each experiment and is tabulated in column 3 of table 1. Column 4 shows the data accessed by the specific human user. Next, the unwanted data is identified and tabulated in column 5. By subtracting the search engine access and unnecessary data from the raw weblog, tabulated in column 6, the pre-processed weblog size is determined. Column 7 of Table 8.1 reveals that with the size of the raw weblog, the size of the pre-processed weblog is reduced to 62%-66%. For the purposes of the data processing, this reduced data is not relevant.

1	2	3	4	5	6	7
Month	**Raw weblog size (in GB)**	**Search engine accesses size (in GB)**	**Human user accesses size (in GB)**	**Unwanted data size (in GB)**	**Pre-processed weblog size (in GB)**	**% of reduction size by EPPM (in GB)**
1	1733.10	563.70	1169.40	539.90	629.50	63.68
2	1245.00	448.50	796.50	327.15	469.35	62.30
3	2040.30	540.90	1499.40	768.40	731.00	64.17
4	3489.00	986.70	2502.30	1153.50	1348.80	61.34
5	1190.70	423.70	767.00	315.70	451.30	62.10
6	2958.00	694.50	2263.50	1128.60	1134.90	61.63
7	2630.70	684.60	1946.10	926.70	1019.40	61.25
8	2706.90	736.20	1970.70	986.60	984.10	63.64
9	2300.40	596.70	1703.70	911.20	792.50	65.55
10	3692.70	1040.10	2652.60	1247.50	1405.10	61.95
11	4646.70	1151.60	3495.10	1827.20	1667.90	64.11
12	4097.40	1285.70	2811.70	1375.90	1435.80	64.96

Table 1. Summary of the performance of ML algorithms

In summary, as shown in table 2, the number of IP addresses and the number of unique users are reported day by day. Sessionization is carried out on the basis of time-oriented heuristics, the approach is implemented on the same weblog, and the number of day-wise sessions is recognized and tabulated.

Day	No. of IP Addresses	No. of Unique Users	No. of Sessions
1	5460	7089	21294
2	5940	7723	23166
3	4650	6054	18135
4	4980	6470	19422
5	5220	6876	20358
6	5040	6525	19656
7	4980	6447	19422
8	5790	7658	22581
9	5610	7392	21879
10	5670	7254	22113
11	5700	7458	22230
12	5040	6641	19656
13	6030	7842	23517
14	6540	8546	25506
15	6150	7958	23985
16	5580	7354	21762
17	4440	5784	17316
18	4650	6145	18135

Table 2 Results of user identification and sessionization

5. Conclusion and Future Work

To create effective cross marketing strategies and to meet the specific requirements of the user timely, now-a-days every organization should pay the attention on the era of web engineering techniques. Towards this, the scalability, efficiency, usability and even feasibility issues are the challenges with the rapid growing nature of weblog and pre-processing is an essential phase as the accuracy of web mining techniques is highly dependent on the quality of data. With this aim, the authors in the present paper mainly concentrated on the stage of pre-processing and proposed Enriched Pre-Processing Model with a right integration of machine learning approach. The experimental results are evident that as a whole, the EPPM effectively generates enriched structured weblog which helps the data analytics in making quality decisions that are closure to the reality.

The present research effort geared the direction of future research towards integration of machine learning based pre-processing models with the further stages of big data processing to retrieve robust solutions in the real time analysis.

Acknowledgements

The authors expressed their gratitude to the authorities of Centurion University of Technology and Management (CUTM), Odissa for their continuous encouragement and constant support. The authors also reported their acknowledgments to the authorities of Shri Vishnu Engineering College for Women (Automonous), Bhimavaram, A.P., India for their cooperation.

References

[1] Ali Mostafaeipour, Amir Jahangard Rafsanjani, et.al., “Investigating the performance of Hadoop and Spark platforms on machine learning algorithms”, The Journal of Supercomputing, 2020, pp.01-28.

[2] Jaber Alwidian, Sana Abdel Rahman,et.al., “Big Data Ingestion and Preparation Tools”, Modern Applied Science; Vol. 14, No. 9, 2020, pp.12-27.

[3] Archana A. Chaudhari and Preeti Mulay, “SCSI: Real-Time Data Analysis with Cassandra and Spark”, Springer Nature Singapore Pvt Ltd, 2019, pp.237-264.

[4] Eman Shaikh, Iman Mohiuddin, et.al., “Apache Spark: A Big Data Processing Engine”, 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), 2019, pp.01-06.

[5] Mitali Srivastava, Atul Kumar Srivastavab, Rakhi Gargc, “Data Preprocessing Techniques in Web Usage Mining: A Literature Review”, Int. Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019), 2019, pp.466-476.

[6] N. Silpa, Dr. Maheswara Rao V V R, “A Complete Research on Techniques & Technologies of Big Web Data Preparation to Web User Usage Behavior”, published in International Journal of Recent Technology and Engineering(IJRTE), ISSN:2277-3878, Vol. 8, Issue 2S11, September 2019.

[7] Taiwo Kolajo, Olawande Daramola and Ayodele Adebiyi, “Big data stream analysis: a systematic - literature review”, Journal of Big Data, Vol.6:47, 2019, pp.1-30.

[8] Yun Li, Yongyao Jiang, et.al., “A Cloud-Based Framework for Large-Scale Log Mining through Apache Spark and Elasticsearch”, applied sciences, MDPI, Vol.9, 2019, pp. 01-13.

[9] Amine Ganibardi, Chérif Arab Ali, “Weblog Data Structuration - A Stream-centric approach for improving session reconstruction quality”, iiWAS 2018, ACM, 2018, pp.01-09.

[10] Dilip Singh Sisodia, Vijay Khandal and Riya Singhal, "Fast prediction of web user browsing behaviours using most interesting patterns", Journal of Information Science, SAGE, Vol.44(1), 2018, pp.74-90.
[11] Marlina Abdul Latib, Saiful Adli Ismail, et.al., "Analysing Log Files For Web Intrusion Investigation Using Hadoop", ICSIE, Association for Computing Machinery, 2018, pp.12-21.
[12] Alexandra L'heureux, Katarina Grolinger, "Machine Learning With Big Data: Challenges and Approaches", Vol.5, 2017, 7776-7797.
[13] Dimitrios Sisiaridis and Olivier Markowitch, "Feature Extraction and Feature Selection: Reducing Data Complexity with Apache Spark", International Journal of Network Security & Its Applications (IJNSA) Vol.9, No.6, 2017, pp.39-51.
[14] Pinjia He, Jieming Zhu, et.al., "Towards Automated Log Parsing for Large-Scale Log Data Analysis", Journal of Latex Class Files, 2017, pp.01-14.
[15] Alex Liu, "Apache Spark Machine Learning Blue Prints", PACKT Publishing,2016.
[16] David Stodder, "Improving Data Preparation for Business Analytics", tdwi publications, 2016.
[17] Ilias Mavridis, Eleni Karatza, "Performance evaluation of cloud-based log Þle analysis with Apache Hadoop and Apache Spark", The Journal of Systems & Software, 2016.
[18] Salman Salloum, Ruslan Dautov, "Big data analytics on Apache Spark", Int J Data Sci Anal, Springer, 2016, pp.145-164.
[19] Salvador García, Sergio Ramírez-Gallego, et.al., "Big data preprocessing: methods and Prospects", Big Data Analytics, Vol. 1:9, 2016.
[20] Abdul Ghaffar Shoro & Tariq Rahim Soomro, "Big Data Analysis: Ap Spark Perspective", Global Journal of Computer Science and Technology:Software & Data Engineering, Vol. 15, Issue.1, 2015, pp.01-09.
[21] Andoena Balla, Athena Stassopoulou and Marios D. Dikaiakos, "Real-time Web Crawler Detection", 18th International Conference on Telecommunications, IEEE xplore, 2015, pp.01-05.
[22] Dr. Maheswara Rao V.V.R., N. Silpa, "A Comprehensive Study on Potential Research Opportunities of Big Data Analytics to Leverage The Transformation In Various Key Domains", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 5, No.5, 2015, pp.01-18.
[23] Federico Castanedo, "Data Preparation in the Big Data Era – Best Practices for Data Integration", O'Reilly Publication, 2015.
[24] Matei Zaharia, Reynold S. Xin, "Apache Spark: A Unified Engine for Big Data Processing", Communications of the ACM, Vol. 59, No. 11, 2015, pp.56-65.
[25] Mitali Srivastava, Rakhi Garg, P. K. Mishra, "Analysis of Data Extraction and Data Cleaning in Web Usage Mining", ICARCSET 2015, ACM, 2015, pp.01-06.
[26] Peng Huang, Dehua Chen, Jiajin Le, "An Improved Referrer-Based Session Identification Algorithm Using MapReduce", Ninth International Conference on Natural Computation (ICNC), IEEE, 2015, pp.1072-1076.
[27] Sara Landset, Taghi M. Khoshgoftaar, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem", Journal Big Data, Journal of Big Data, Vol2.24, 2015, pp.01-36.
[28] Sonali Agarwal, Bakshi Rohit Prasad, "High Speed Streaming Data Analysis of Web Generated Log Streams", IEEE 10th International Conference on Industrial and Information Systems, ICIIS 2015, 2015, pp.01-06.
[29] Zuhair Khayyat, Ihab F. Ilyas, et.al., "BigDansing: A System for Big Data Cleansing", SIGMOD'15, ACM, 2015, pp.01-17.
[30] You Joung Ham, Hyung-Woo Lee, "Big Data Preprocessing Mechanism for Analytics of Mobile Web Log", Int. J. Advance Soft Compu. Appl, SCRG Publication, Vol. 6, No. 1, 2014, pp.01-18.
[31] Dusan Stevanovic, Aijun AN, and Natalija Vlajic, "Detecting Web Crawlers from Web Server Access Logs with Data Mining Classifiers", Springer-Verlag Berlin Heidelberg 2011, 2011, pp.483-489.
[32] Maheswara Rao V.V.R., Dr. V. Valli Kumari, Dr. KVSVN Raju "An Intelligent System for Web Usage Data Preprocessing" presented in The First International Conference on Computer Science and Information Technology - CCSIT-2011, January 2 - 4, 2011. Bangalore, India, Springer LNCS-CCIS, ISSN: 1865-0929, ISBN: 978-3-642-17856-6, Vol. 131, Part 1, pp. 481-490, 2011.
[33] Maheswara Rao V.V.R., Dr. V. Valli Kumari, Dr. KVSVN Raju "Study of Visitor Behavior by Web Usage Mining" presented in International Conference on Recent Trends in Business Administration and Information Processing - BAIP 2010, India, and proceedings published by Springer LNCS-CCIS, ISSN: 1865-0929, ISBN: 978-3-642-12213-2, Vol. 70, pp. 181-187, 2010.
[34] Pablo E. Roman, Robert F. Dell, and Juan D. Vel´asquez, "Advanced Techniques inWeb Data Pre-processing and Cleaning", Advanced Techniques in Web Intelligence, Springer, 2010, pp.19-48.
[35] Jie Zhang and Ali. A. Ghorbani, "The Reconstruction of User Sessions from a Server Log Using Improved Time-oriented Heuristics", Proceedings of the 2nd Annual Conference on Communication Networks and Services Research (CNSR'04), IEEE, 2004, pp.01-08.
[36] Fang Yuan, Li-Juan Wang, Ge Yu, "Study On Data Preprocessing Algorithm In Web Log Mining", Proceedings of the Second International Conference on Machine Learning and Cybernetics, IEEE, 2003, pp.28-32

Authors Profile

N Silpa, is currently pursuing her Ph.D. in Computer Science and Engineering at Centurion University of Technology and Management (CUTM), Odissa, India. She has 10 years of teaching experience and 5 years of Research Experience. Her Research interests include Data Mining, Web Mining, Big Data Analytics, Text Mining, Data Science, Artificial Intelligence and Machine Learning.

Dr. V.V.R. Maheswara Rao, is a leading Researcher & Academician in Computer Science & Engineering and holds Ph.D. degree. He is currently working as a Professor in the Dept of Computer Science & Engineering at Shri Vishnu Engineering College for Women (A), Andhra Pradesh, India. He is actively involved and successfully implemented three projects funded by DST. He has 31 research papers, six of which are Scopus-indexed and four of which are Web of Science-indexed. He has 23 years of experience that include 6 years of Industry experience, 17 years of Teaching experience and 13 years of Research experience. His Research interests include Data Mining, Web Mining, Big Data Analytics, Data Science, Artificial Intelligence and Machine Learning.