A Comprehensive Research Study on Web Usage Data Analytics to Investigate the User Behavior with Machine Learning

N. Silpa[1], Prof. Anurag[2] and Dr. V V R Maheswara Rao[3]

[1]*Research Scholar, CUTM, Odisha & Asst. Prof., Shri Vishnu Engineering College for Women, Bhimavaram*

[2]*Pro Vice Chancellor, CUTM, Odisha*

[3]*Professor, Shri Vishnu Engineering College for Women, Bhimavaram*
[1]*nrusimhadri.silpa@gmail.com,* [2]*anurag@cutm.ac.in,* [3]*mahesh_vvr@yahoo.com*

Abstract

In the light of wide spread use of WWW, the organizations shakeout with fled flowing transactional data and accumulating trillions of bytes of information about their customers, suppliers with their operations. Moreover, in this digitized era users are leaving enormous amount of data, and it is widely distributed, high dimensional, interrelated and dynamic in nature. As a result, it become tolerant and golden mount, in turn led to big data paradigm. With this trend, the web usage has profoundly influenced many aspects of organizational needs, nevertheless, the legacy systems and techniques proved to be inefficient and poses great challenges and led forward to actionable analytics. This situation gains the focus of present researchers towards big web data analytics equipped with the Self, Artificial Intelligent and Machine Learning algorithms.

The authors in the present paper conducts a comprehensive research study on modelling web usage data in the light of machine learning approach, shinning a bright spot light on the insights of the big web data to investigate the web usage behavior as only web usage data represents the actual behavior of the web user. In this regard, the authors articulate a systematic framework with a focus on each stage of finding the user usage behavior. Initially, the study focusses on the possible ways of collecting the web usage data in various environments. Later, the study concentrates to explore the methods of on the preparation of the data and filter out the unwanted data. Subsequently, the study showcases the techniques to perform exploratory data analysis. Finally, the study especially torches the light on employment of machine learning techniques and technologies and their value addition to get actionable analytics for investigating the web user usage behavior and to leverage in optimal decision-making process.

Keywords: *Web Usage Data, Big Web Data, Web Data Collection, Web Data Preparation, Web Data Modelling, Web Analytics, Big Data Analytics, Machine Learning, Web User Behavior*

1. Introduction

The World Wide Web (WWW) is rapidly and significantly increasing with immense amount of data along with massive number of users. The usage data in the web is widely distributed, highly dimensional, interrelated and dynamic information [9, 15, 26]. The users on the web have high aspirations on reliability. With this trend and explosion of information technology, web services and web-based applications, the WWW has become a golden mount. As a result, it influenced many aspects of user lives and brought changes in the fields of communication, business, sharing knowledge and so on. However, the abundant information on the web, which is structurally complex, exponentially growing in nature become web big data and poses great challenges in the era of web usage data analysis.

Web usage data analysis, also known as web data analytics, is the process of automatic discovery and investigation of trends and patterns in click streams and associated data collected or generated as a result of user interactions with web resources on websites [3, 15, 17]. The main goal of web usage analysis is to capture, model and analyze the behavioral patterns of users interacting with websites. The discovered patterns are usually represented as collection of pages, objects or resources that are frequently accessed by groups of users with common needs or interests.

In technology perspective, web big data is the possibility of [1, 10, 12, 13, 14, 25] improved storage -Volume, the ability to process the information and make it available in real time - Velocity and the ability to deal with various kinds of data sources, including structured, semi-structured and unstructured ones -Variety. Inclusion of Veracity as the fourth big data attribute emphasizes the importance of addressing and managing for the uncertainty inherent with in some times of data. The technology exists, so the essential issue is how carriers can make sense of the massive volumes of data and deliver as fifth attribute Value to business. Fundamentally, web big data means not only 5Vs of web user usage data but also describes a new generation of technologies and architectures, design to economically extract value from very large volumes of a wide variety of web usage data, by enabling velocity capture, discovery and analysis. This definition congregates the five dimensions: Volume, Velocity, Variety, Veracity and Value as shown in Figure 1.

Figure 1. Dimensions of Web Usage Data

Further, the web has really grown-up as a channel and further got focused deep demand and held as accountable for the organizations. Thus, many organizations spent millions of dollars for web channels and web analytics regarding their business. Yet consistently, the ability of learn accurately and making optimal decisions from that web usage data is become number one challenge to the organizations. To learn from such data, it is necessary to adopt Machine Learning techniques [1, 4, 6, 7, 8] which explores web usage data and develops learning mathematical models and algorithms. They fundamentally understand the key insights in data and define the new world of actionable web analytics.

This paper is organized as follows. Section 2 describes related work. Later, section 3, presents proposed comprehensive research study in each phase, specifically suitability of machine learning approach. Finally, in section 4 conclusions and future work are mentioned.

2. Related Work

The comprehensive research study in the present paper contains literature survey from 2010 to current year. In specific, the authors in the present study concentrate on each phase of web usage data analytics and explored the complete research challenges & opportunities extensively and comprehensively. Towards this the research review and the contributions made by the authors [1, 12, 13] noticed and emphasizes that the web usage data analytics includes interdependent stages namely data collection, data preparation, data exploration and data modelling. Furthermore, the recent literature study [14, 25] has recognized that implementation of this model with machine learning is a promising and practical research area.

Many of the earlier authors as observed in the literature have explained the importance, criticality and efficiency of comprehensive approach in the process of web usage data analytics, which has been considered as the formal basis for the present study.

3. Proposed Comprehensive Research Study on Web Usage Data

In the recent past, many organizations are doing their business or services through legacy systems and internet. They also realized the need of understanding the behavior of stakeholders, in view of vital signs of business in this digital era [12, 13, 27]. In addition, the top leaders in every organization have high expectations on capturing value from insights of such big data. In achieving fair and real knowledge, big data analytics with machine learning are more suitable and its processing stages plays a critical role to overcome the technology challenges posed by the characteristics of big web usage data. It is essential for all the organizations to look at external data and compare internally generated data to make actionable analytics. This opportunity motivates the present authors to conduct a comprehensive research study on focusing each stage of finding web user usage behavior. Towards this, the authors carefully designed a systematic framework with an attention on each stage as shown in Figure 2.

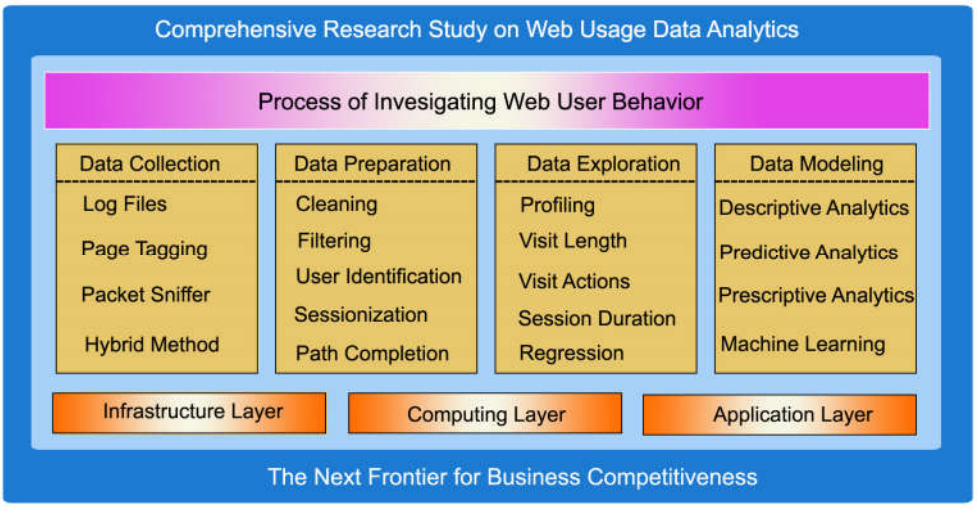

Figure 2. Systematic Framework of the Proposed Study

The review initially focuses on techniques and technologies involved in data collection as it is very typical and important in the web scenario. Subsequently, the authors emphasize on the methodologies in the data preparation includes filtering unwanted data. Later, the framework concentrates on the stage of exploratory data analysis on web user usage data to draw the empirical insights. Finally, the authors focus on employing of techniques and technologies of machine learning [4, 6, 7, 8] in producing actionable analytics which use optimization algorithms to advice on possible outcomes.

3.1. Web Usage Data Collection

The study of web user usage and activities performed by the user, produces valuable marketing intelligence like: Assurance, Insight and Optimization. Data collection is the prime step in web user usage process [12, 13, 15] and it is the process of extracting the task relevant data from diverse logs. It is very important and difficult task to get details of web user usage data from various sources as shown in Figure 3, since the capability, comprehensiveness, and timeliness of web data analytics solution depend on the methods used for capturing Web usage data. The authors in the literature, explored several methods.

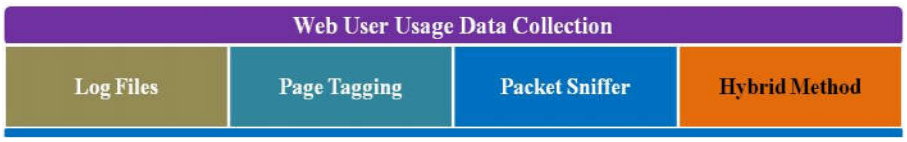

Figure 3. Web Usage Data Collection

3.1.1 Log Files: Many authors explored that [15, 22, 26], Log files are the prime source to gather web user usage data. Basically, the web user usage information is retrieved from three types of raw web log files: access logs, referrer logs and agent logs. This method of data collection is based on server-side and depending on the configuration the data is recorded in various formats. However, the common log format consists a variety fields of web user usage data that includes: remote host field, identification field, authuser field, date/time field, HTTP request, status code field, transfer volume field, referrer field and user agent field. All web servers automatically create these logs and store them on its respective servers with a freedom to change their analytics and strategies. The method need not require any extra bandwidth when loading a page, and since everything is recorded at server-side, it is possible to log both page request successes and failures also.

3.1.2 Page Tagging: Page Tagging is another approach in the literature that records visitor activity in invisible image when a page was successfully loaded, also solves accuracy problems in the logfile analysis. Eventually, this method evolved into script-based data collection which allots a cookie to individual user, investigates their behavior on the web, and then processes the data remotely. Unlike a log files, the data received via page tagging is parsed as it comes in and allows for near real-time reporting. More specially, the popularity of page tagging is easier to record additional information about the visitor that does not involve a request to the Web server. The authors [22, 26] also reported that the page tagging method consumes more bandwidth and also fails to record unloaded or failure pages.

3.1.3 Packet Sniffer: Packet sniffer is another way [22, 26] to capture the data by deploying either a software or a hardware tool over a network. Here the packet is decoded as per the defined configuration of the software or hardware tool. As a result, this method led to different formats for the collected data, and also requires additional process.

3.1.4 Hybrid Method: However, in many of the practical cases hybrid method suits to collect the data which allows combination of multiple data collection methods. Many of the authors in the literature endorsed that this method gives greater accuracy of tracking sessions across multiple domains, eliminating the caching problems, and tracking detailed web usage data metrics. Many of the authors [22, 26] in the literature noticed and expressed the importance of the data collection is the basic activity.

The web user usage data collected any of the method is unstructured since the data contains different types of entries. These entries do not have definite number of attributes, identifiable structure and defined relation. This resulted in ambiguities and difficult to understand. The characteristics of such web user usage data are heterogeneous, incremental, distributed and have more insights. These characteristics coverage the five dimensions: volume, velocity, veracity, variety and value [1, 3, 10] in turn define and describe the big data.

3.2. Web Usage Data Preparation

The accuracy of any web usage data modelling on raw web logs consists of big data is highly dependent on the data preparation stage in the process of investigating the web user usage behavior as presented in Figure 4. To prepare the useful data, it is necessary for the web data analysts to employ data preparation techniques [3, 6, 7, 8] before modelling the web usage data. This helps in making the optimal actionable decisions, closer to the reality.

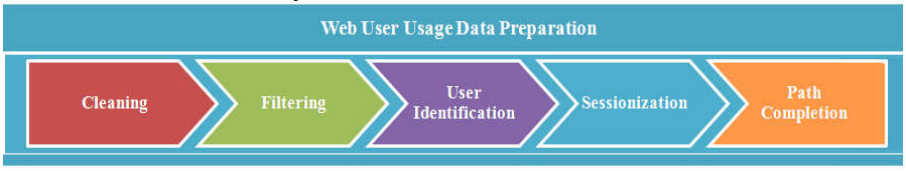

Figure 4. Web Usage Data Preparation

3.2.1 Data Cleaning: Once the data collection is done, the web usage analyst focusses to clean up the data. Data cleansing is usually [1, 3, 9, 13, 15, 20] site-specific, and involves tasks such as, removing extraneous references to embedded objects that are not important for the purpose of analysis, including references to style files, graphics, or sound files. The cleaning process also involve the removal of at least some of the data fields (e.g. number of bytes transferred or version of HTTP protocol used, etc.) that may not provide useful information in data analysis tasks.

3.2.2 Data Filtering: In addition to the typical nature of web usage data, the web log also stores nonhuman access data [15, 23, 24] along with human user usage data. The search engine data is not useful with the aim of identifying the web user usage behavior. In order to reduce the data preparation time, initially it is important to separate and discard the web search engine accesses from web usage data like spiders, crawlers, and other automatic web bots are constantly crawling around the Web. The behavior of the bots differs qualitatively from human behavior and is not considered interesting from a web usage analysis standpoint. The most useful method in the literature for eliminating bots, spiders, and crawlers from the web usage data is to identify the spider's name in the user agent field, when supplied.

3.2.3 User Identification: The task of user identification is to identify who accesses website and which pages are accessed [3, 9, 20, 23, 24]. The analysis of web usage data does not require knowledge about a user's identity. However, it is necessary to distinguish different users, since a user may visit a site more than once, the server logs record multiple sessions for each user. Ideally, this would be accomplished easily if the user provides his or her registration information, such as user name and password, each time the website was accessed. Unfortunately, the free-form structure of the Internet means that most user accesses to most websites are done anonymously, so that registration information is not available. According to the literature, another way of describing this situation is to say that the Internet is essentially stateless, meaning that each request for a web page gets treated as an isolated event, unrelated to all other requests for the site's web pages. User identification is one way of introducing a state into this stateless system. The remote host field, or IP address field, may in principle be used to identify users. However, the widespread use of proxy servers, corporate firewalls, and local caches renders problematic use of the IP address as a substitute for user identification.

3.2.4 Sessionization: The concept of user session is important because it corresponds to what is often considered to be a visit to a website. Sessionization is the process of segmenting the user activity record of each user into sessions, each representing a single visit to the site. Websites without the benefit of additional authentication information from users and without mechanisms such as embedded session ids must rely on heuristics methods for sessionization. The goal of a sessionization heuristic is to re-construct, from the web user usage data, the actual sequence of actions performed by one user during one visit to the site. That is, for each visit, determine which pages were requested, the order of the requests, and the duration of each page view. Also, try to estimate when the user left the website. Perhaps the most straightforward sessionization heuristics fall into two basic categories: time-oriented or structure-oriented. Many authors [3, 9, 20, 23, 24] in the literature have identified various heuristics for sessionization.

3.2.5 Path Completion: Path completion is another potentially important data preparation task after sessionization. Path completion is process of adding the page accesses that are not in the web usage data but those which should actually occur [20, 23]. Client or proxy-side caching can often result in missing access references to those pages or objects that have been cached. Many people use the "Back" button on their browsers to return to a page viewed previously. When this happens, the browser returns to a page that has previously been cached locally rather than accessing the web server again. This leads to "holes" in the web server's record of the user's path through the website. Knowledge of site topology must be applied to complete these paths. Once the missing pages have been

identified, they are inserted into the session file along with an estimate of the duration spent on the missing page. These duration estimates may be classified according to whether the missing page is a navigation page, with a shorter duration estimate, or a content page, with a longer duration estimate.

In turn, in the recent past the web usage data analysis got attention of many researchers [12, 13, 15], yet, the data preparation in investigating the web user usage behavior has received less attention than it deserves. Many researchers [3, 9, 15, 20, 23, 24] are working on data preparation that involves user identification, session identification, path completion and so on. They endorsed that web usage data preparation is very important and crucial task in entire process. This stage can be strengthened by choosing and employing various intelligent techniques. This helps any organization to track the user behavior in order to meet the desires of specific users and cross marketing strategies effectively.

3.3. Web Usage Data Exploration

After the heavy lifting of web usage data preparation and before modeling the web user usage information, it is preferable to accomplish exploratory data analysis. Specially, exploratory data analytics allows the analyst to investigate deeper into the web usage data, inspect the interrelationships among the fields, and reveal interesting patterns of the log records.

Profiling the web users to the website, the important factor is visit length, in terms of number of visit actions are made. Also, the analysts would be interested in obtaining various statistics related the aggregate behavior of web user in terms of the numbers of visit actions and mean number of visit actions per session.

Apart from the number of user actions per session, another important variable is the time duration per session [3,9,23] that the user spent on the website. Unfortunately, the time spent by the last page session is not recorded in the usage data, the exact amount of time per session will prove to be elusive. Thus, when calculate the session duration, need to restrict the sessions to those that contained more than a single action; otherwise, there is no measured duration at all for a single web log entry.

Next, it is turn to the finding the relationship between session duration and the number of visit actions by the web user. One of the important method to find the relationship between session duration and visit actions is regression analysis. The regression analysis estimates equation to make predictions for the session duration, given a particular number of visit actions of web user on site.

Moreover in the literature[20,23], many of the researchers noticed the importance of average time per page and duration of individual page also. Visits to websites tend to be rather short in terms of page requests. Thus, the analyst is interested in investigating the behavior of user for the benefit of making recommendation and actionable decisions as quickly as possible.

3.4. Web Usage Data Modelling

The final and important stage of proposed comprehensive research study on web user usage data is data modeling, here, the web analysts apply the all possible skills to identify the patterns and trends from the hidden insights of the web usage data. The web analysts focus with an objective of collect, measure, track and analyze quantitative web usage data [5, 13]. Thus, these patterns help to evaluate and understand the behavioral aspects of web user with a focus on providing actionable and optimized initiatives. This situation endorsed with great challenges and in turn motivated the researchers towards discovering actionable insights. Towards this, the authors in the literature [14, 15] and the organizations contributed with solutions which deliver knowledge in turn to take smarter

decisions for better business outcomes. These solutions considered in three dominant types: Descriptive Analytics, Predictive Analytics and Prescriptive Analytics as represented in Figure 5. However, the common ability of all the analytics is to leverage the business outcomes.

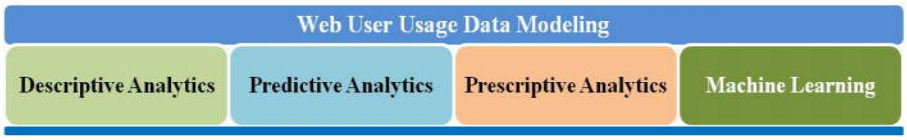

Figure 5. Web Usage Data Modelling

3.4.1 Descriptive Analytics: Descriptive analytics [2, 5, 11] focus on analyzing historical data for the purpose of identifying patterns and trends. These conventional analytics create business intelligence and data analysis with facts and figures in understandable format. These analytics are useful as they allow learning from past behavior and further creating the path to understand the influence on the future. These analytics summarize with certain groups along with simple counts. In addition, some more researchers expressed that, descriptive analytics closes the analysis at surface level of the data and the validity of the results unable to interpret if large number of attributes presented in the data.

3.4.2 Predictive Analytics: Predictive analytics [2, 5, 10, 11] works with an intension to predict what might happen by understand the past. Predictive analytics started by providing actionable insights to the organization. These analytics are able to provide estimations of future outcomes. Predictive analytics are probabilistic in nature and provide better recommendations better than the business intelligence. Predictive analytics able to capture relationships among various attributes beyond the surface area of the data. This modeling is able to mathematically represent underlying relations in the historical data and make predictions. Predictive models operationalized in machine critical transaction system and summarize large quantities of data to amplify its value. The predictive models analyze multiple aspects of individual behavior to forecast the future behavior by providing actionable results.

3.4.3 Prescriptive Analytics: Relatively from the above two analytics, the prescriptive analytics allows to prescribe a number of possible and actionable solutions. These analytics are able to quantify the effect of future decisions in order to advice on possible outcomes before making the decisions. These analytics are comparatively complex in nature and many organizations are not yet employing for their day to day activities. Prescriptive analytics are also useful for the traditional analysts to dig and determine the appropriate action on their own and in turn make more efficient. Another important aspect of prescriptive analytics is their ability to analyze the feedback coming from the using the rules, in order to take effective actions. The authors [2, 11] in the literature explored the techniques deployed for these analytics starting with decision trees, fuzzy based system, neural network and support vector machines.

3.4.4 Machine Learning: In order to create a comprehensive view on the data it is necessary to use the combination of the above analytics appropriately. Specifically, Machine learning [1, 4, 6, 7, 8, 14] has also greatly enhanced the speed of predictive and prescriptive analytics and the breadth of data that can be incorporated. Machine Learning, a subset of AI, is the area of computation science that is divided into three folds: supervised learning, unsupervised learning and reinforced learning. To make actionable decisions from the unprecedent volumes of web usage data it is necessary to understand the relations among the data. To do this, reinforcement algorithms are more adaptable from the entire learning paradigms. Reinforced learning algorithms enables the algorithms learn from feedback received from interactions with external environments. In addition, reinforced learning algorithms focuses on analyzing and interpreting patterns and

structures in web user usage data to enable learning, reasoning and that allow for autonomous decision making.

4. Conclusion

The literature [12, 15] is clearly evident that, the capable and successful handling of web usage data management by using sophisticated web usage analytics is a promised research area with great challenges, is one of the main motivations to the present research study. This extensive research study noticed and emphasized that to create more optimize and actionable analytics techniques to serve the increasing demands of each organization to satisfy their demands [16, 21, 25] and solve the problems encountered on the web has become the further motivation to the present investigation. Many authors [2, 5, 7, 8] have expressed the importance, criticality and efficiency of comprehensive analysis of web user usage data yet, most of the works concentrate on individual stages only. Furthermore, the literature survey has recognized that employing machine learning and self-learning algorithms in web usage data analytics creates potential and practical research paths.

Acknowledgments

The authors recorded their acknowledgements to Centurion University of Technology and Management, Odisha and Shri Vishnu Engineering College for Women, Bhimavaram authorities for their constant support and cooperation.

References

[1] Ahmed Oussous a, Fatima-Zahra Benjelloun, et.al, “Big Data technologies: A survey”, Journal of King Saud University – Computer and Information Sciences, Vol. 30, pp: 431-448, 2018.

[2] Dr. Venkatesh Naganathan, “Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends”, International Research Journal of Engineering and Technology, Vol. 05 Issue: 05, pp: 19488-1964, 2018.

[3] Muhammd Jawad Hamid Mughal, “Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview”, International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, pp:208-2015, 2018.

[4] Suryabhan Pratap Singh, Umesh Chandra Jaiswal,“ Machine Learning for Big Data: A New Perspective”, International Journal of Applied Engineering Research, Vol. 13, No.5, pp. 2753-2762, 2018.

[5] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, et.al. “Critical analysis of Big Data challenges and analytical methods”, Journal of Business Research, pp: 263–286, 2017.

[6] Zoila Ruiz, Jaime Salvador, “A Survey of Machine Learning Methods for Big Data”, Springer International Publishing, IWINAC 2017, Part II, pp. 259–267, 2017.

[7] Junfei Qiu, Qihui Wu, “A survey of machine learning for big data Processing”, Journal on Advances in Signal Processing, pp: 1-16, 2016.

[8] Lidong Wang, Cheryl Ann Alexander, “Machine Learning in Big Data”, International Journal of Mathematical, Engineering and Management Sciences, Vol. 1, No. 2, pp: 52–61, 2016

[9] Sunena, Kamaljit kaur, “Web Usage Mining-Current Trends and Future Challenges”, International Conference on Electrical, Electronics, and Optimization Techniques, pp:1409-1414,2016.

[10] Amir Gandomi, Murtaza Haider, “Beyond the hype: Big data concepts, methods, and analytics”, International Journal of Information Management, pp: 137-144, 2015.

[11] Dennis Oliver Huisman, “To What Extent do Predictive, Descriptive and Prescriptive Supply Chain Analytics Affect Organizational Performance?”, IBA Bachelor Thesis Conference, University of Twente, pp: 1-9, 2015.

[12] Dr. V V R Maheswara Rao, Dr. Valli Kumari, and 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, Vol.5, No.5, pp:1-18, 2015.

[13] Dr. V V R Maheswara Rao, Dr. Valli Kumari, and N. Silpa, "An Extensive Study on Leading Research Paths on Big Data Techniques & Technologies", International Journal of Computer Engineering & Technology, Vol. 6, Issue 12, pp: 20-34, 2015

[14] Lisbeth Rodríguez-Mazahua1, Cristian-Aarón Rodríguez-Enríquez, "A general perspective of Big Data: applications, tools, challenges and trends", Springer Science+Business Media New York 2015, pp: 1-49, 2015.

[15] Han Hu, Yonggang Wen, et.al., "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", IEEE Access, Vol. 2, pp: 652-686, 2014.

[16] Xindong Wu, Xingquan Zhu, et.al., "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, pp: 97-107, 2014.

[17] Center for Media Justice, "Consumers, Big Data, and Online Tracking in the Retail Industry", A Case Study of Walmart, pp: 1-30, 2013.

[18] Serendio, "Big Data Management and Predictive Analytics as-a-service for the Retail Industry", Executive Summary, pp: 1-6, 2013.

[19] Xin Luna Dong, Divesh Srivastava, "Big Data Integration", Proceedings of the VLDB Endowment, Vol.6, No. 11, 2013.

[20] Amog Rajenderan, "Data Preparation for Web Mining-A survey", International Journal of Advanced Computer Research, Vol. 2, No.4, pp: 1-7, 2012.

[21] Danyel Fisher, Rob DeLine, et.al., "Interactions with Big Data Analytics", pp: 1-10, 2012.

[22] Lakhwinder Kumar, Hardeep Singh, et.al., "Web Analytics and Metrics: A Survey", International Conference on Advances in Computing, Communications and Informatics, ACM, pp: 966-971, 2012.

[23] Liu Kewen, "Analysis of Preprocessing Methods for Web Usage Data", International Conference on Measurement, Information and Control, pp: 383-386, 2012.

[24] Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems", pp: 1-5, 2012.

[25] SaS Institute, "Big Data Meets Big Data Analytics", White Paper, pp: 1-13, 2011.

[26] Brijendra Singh, Hemant Kumar Singh, "Web Data Mining Research: A Survey", 978-1-4244-5967-4/10, IEEE, pp: 1-10, 2010.

[27] David Bollier, Rapporteur, "The Promise and Peril of Big Data", Aspen Institute, Washington, pp:1-66, 2010.