International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Performance of Web Traffic Activities using Web Mining and Machine Learning Techniques

Abha Narwal, R. K. Chauhan

Abstract: ***In recent days Web mining gathers all tools, approaches, and algorithms that had to retrieve information and knowledge through data-based data. The portion of this methodology is aimed at analyzing users ' behaviors, to continue improving the framework and content of websites visited consistently. A relevant question then arises: how much more the attempt to enhance the services provided via a website breaches the privacy of visitors? The use of important retrieval resources including web mining can threaten the privacy of users. This paper would concentrate on developing approaches to speed up the weblog mining process and also to show data visualization as a consequence of the log mining process and evaluate algorithms for data mining. The right metrics to equate algorithms will be used for the analysis of the classification methods, accuracy RMSE and MAE. The fundamental goal of the case study is to evaluate the usefulness of the expert-driven system and data-driven method for the classification of authenticated network traffic, in particular, SSH traffic from traffic log files.***

Keywords: Web Usage Mining, J48 algorithm, SSH, weblog, classification

I. INTRODUCTION

Mainly due to its applicability in e-business, recent research has focused on web usage analysis. We expect the research community to demonstrate similar, if not more, interest in the privacy issues, distributed cloud mining, and semantic web mining. Nevertheless, the expanded use of cloud mining technologies would entail the resolution of privacy issues. Furthermore, the services provide website owners and network managers with valuable information, enabling changes to the structure and functionality of a site to be created, taking into account the principle of optimizing the user's experience and therefore making access to the site simpler for the visitor. To do this, many methods have been created to collect information from the web data produced by each website access. While work in this field is driven by unselfish ideals," excessive help" may impinge on the privacy of the individual, especially due to persistent personal data requirements. Several studies have shown that websites including content tailored to users can establish a loyal association to their visitors [1], but there are questions about the compensation [2]. Although the interaction between privacy and technology has been discussed on several occasions, the dominant strategy of the IT community has been to follow a restricted definition of data privacy, essentially relating to the contradiction of public and private sectors, based on the nature of data, including whether they are personal data or anonymous data.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Abha Narwal*, Research Scholar, Department of Computer Science and Application, Kurukshetra University, India.
R. K. Chauhan, Professor, Department of Computer Science and Application, Kurukshetra University, India.

While this is not subjective, because it is meant to have a valuable definition that allows functional guidelines to be established, significant negative effects can be predicted. We focus mainly on web mining (WUM), described as "an application that applies data mining techniques to discover patterns of use of WWW data" [3]. Web usage mining can be defined as a user access trend discovery and review, through mining log files and related website details. The important feature associated with web usage mining is, reveal some fundamental ideas like "what are the most popular and least popular pages?" "And" what metropolitan region was made up of most users?" Classifying web traffic operations specifically by device categories using data mining techniques is a major task for getting a notification about traffic congestion from the unauthorized systems. Several countermeasures have been examined and used to find the best solutions in network/system administrators to block or control the problem.

Precise network traffic analysis can assist network operators successfully in many network activities such as bandwidth control and protection. Checking the content of every network packet is a traditional way of classifying network traffic. If the payload is not secured, this approach can be surprisingly accurate. There may nevertheless be concerns about privacy concerning the examination of (arbitrary) user data, and solicitations like SSH (Secure Shell) [4], that can encrypt the payloads, indicating an opaque payload. Another solution to traffic grouping is therefore to use major TCP / UDP port numbers. But the method turns to be more and more unreliable.

The main contributions and organization of this paper are summarized as follows: In section 2 we describe background details of web traffic mining treatment. Section 3 discusses the proposed work. Section 4 deliberates results and discussions. Finally, in section 5, we concluded the paper.

II. BACKGROUND WORKS

In [5], the authors spoke about the nuances in semantic web 2.0, as well as about the advantages and disadvantages of both approaches. The author states that the semantic network is special but needs only a certain basic structure of the site to improve usability and stability. In [6] the authors first addressed a model on the semantic website which provides a complete description and design of the functions of each element. The writers have discussed the scalability and the findings of the semantic web [7].

In [8] the authors described and introduced a new approach to the identification of sessions. The world-wide-web growth is unbelievable as it is now visible. We use mining plays a major role in personalized web services, transforming websites and improving the performance of web servers.

Data mining methods are used to find web access trends from site log data. Server log data should be assembled into sessions to identify connection trends.

In [9], the authors examined the mining of websites and concentrated on strategies that can forecast the actions of the user when communicating with the Internet. It tries to understand the data generated by sessions or behaviors of web surfers. There is an attempt to give an overview of the state-of-the-art in web mining research while discussing the most relevant tools available in the field, and the niche requirements lacking in the current tools.

In the light of the current scenario, the authors [10] develop a flexible methodology to analyze the effectiveness of various variables in different dependence variables, all of which are time series, and in particular displays them in terms of the efficacy of different dependent variables Visitors to search engines have several effects on page views that cannot be explained by a single regression.

On the one hand, referral visitors are well equipped with low impact linear regression. Similar users on the other side had an enormous impact on page views. The faster connection speed does not mean a higher effect on on-site visits and web pages content and visitor territories may better explain user behavior at connection speed. Visitors returning have some resemblances with direct visitors.

III. METHODS AND METHODOLOGY

In this paper Figure, 1 demonstrates this work's basic framework. In this post, our details are log files derived from server storage. In CLF format the web access log contains the information about the visitor of a particular web page with his/her IP address and also date and time of page. It also consists of the URL address of the page that is essential for knowing the weblogs. The protocol is the communication device used, e.g. HTTP/1.0. The state is the mark of completion. E.g., 100 is the performance code. The field size indicates the bytes transmitted by page order. In addition to metadata, the expanded log format provides information about the requested page and proxy in the browser.

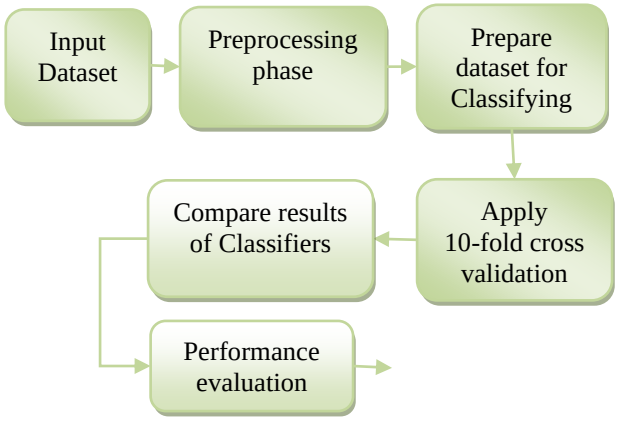

Figure 1. Flow chart for proposed framework for the weblog mining process

Input Dataset: The initiation starts with the cycle of data comprehension. There are many features in Weblog files, therefore at least 10 of the study on data mining are checked. It is clear that practice after recognizing templates for data mining, data mining techniques and relevant principles. The 4th week of log files from the School of Computer Science repositories will be obtained from 25 December 2017 to 31 December 2017. The details about the file logged in this paper are the NCSA Unified Log file format utilizing cookie exclusion personalization.

The standard NCSA composite log file format looks like [7]: *Host, rfc931, username, date: time, request, status code, bytes, referrer, user agent, cookie*. For example: *125.125.125.125 - dsmith [10/Oct/1999:21:15:05 +0500] "GET /index.html HTTP /1.0" 200 1043 "http: //www.ibm.com/" "Mozilla/4.05 [en] (WinXP; I)" "USERID=Customer A; IMPID=01234"*.

As mentioned the address and request of the host do not screen for business intelligence purposes. The IP address is used to evaluate a transaction. It had been believed that if IP address is the same as when the person is on the transaction, it also proposed that 30 minutes to distinguish between different users but in [11] 24 hours as people of modern ages are usually multi-tasking and they typically need more time for processing that data. In this scheme, the use of WEKA as a process for mining of text that can essentially be used to access the same functionality.

Preprocessing phase:

The primary requirement for the preprocessing phase is to ensure that raw data should undergo a reformatting before applied to the WEKA tool that requires the transformation by WEKA of unstructured raw data into a standardized data set. WEKA can understand the common formats arff, CSV, Matlab and so on. The need for the filtering process a useful feature for process attributes or instances monitored or not. This paper uses Microsoft Excel 2007 to pre-process before loading the file into WEKA. Microsoft Excel can load and convert text files to CSV or arff.

Prepare dataset for Classifying algorithms:

Feature selection is a method in which only desirable attributes and insignificant attributes are selected and reworked into a file format that can comprehend data mining applications. The selection of features can enhance the performance of the system and decreases the noise level in the model. Finding the important features must be differentiated that of a selection of features. In this process, the most significant feature is picked from the irrelevant features that have less significance and resultant is a feature subset. Such attributes can be extended to a classifier. Classification also functions well in supervised learning, implying that we have predefined laws to be grouped for instance. The clustering strategies referenced in this study were tested independently on various data sets but somehow good results on accuracy were reported, to obtain useful results of the comparison.

Apply 10-fold cross-validation:

During most of the comparative analysis, 10 cross-validation folds are drawn. The initial sample is uniformly partitioned in 10 equal size subsamples in 10-fold cross-validation. Of the 10 sub-samples, the validation data for testing the model are retained as a single sub-sample, the rest of the data more suitable for the learning process of the network.

Compare results of Classifiers:

The option of the correct algorithm may be the most difficult however crucial decision to be made, with each algorithm producing a different outcome, some even yielding more than one outcome. Furthermore, for every problem there is no need to limit an algorithm to produce different views on datasets.
Many potential considerations have been established when choosing the appropriate web mining set of rules.
a) All the constraints of data mining tools and procedures for these tools can implement.
b) Key goals of the data sets issue and configuration
c) Adapting to one algorithm for the anticipated consequences
d) It could be very fast and provide complete information of the process
e) Web miner must recognize this algorithm and check other input files if necessary before they can be applied on the actual data sets.

Modified J48 decision tree classifier:

The modified decision tree classifier J48 explores the typical knowledge gained from a selection of a fragmented data attribute as demonstrated by Figure 2.

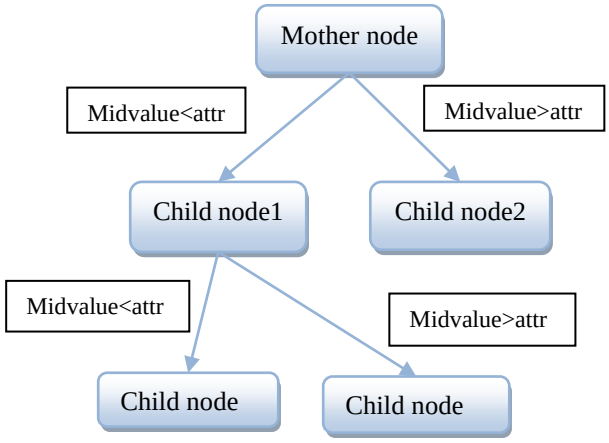

Figure 2. Structure of the modified J48 decision tree

The maximum uniform data gain factor can be used for evaluating. It implies that the system is reappearing in limited subsets. When all cases are large to about the same class in a subclass, the separating process is prevented. The decision tree then generates a leaf node that tells us to select whichever class. It produces a higher decision node in the tree that used the estimated value of the class. When the CADL LSB value and the attribute properties are the same, the condition will be marked as Class' 0,' so Class' 1' condition is recommended. The first step of Tree is a single mother node. It's only the node of a child's reference. The second level of the tree comprises of 2 sub-trees called 1-2.

Naive Bayes: This classifier assumes that there is no connection between the existence of a particular element in a class and any other function.
ZeroR and OneR: ZeroR is a simple but accurate classification algorithm which produces one rule for each data predictor and then selects the rule as the 'one rule' which contains the smallest total error. Only OneR is the simplest way to identify the goal and neglect all the predictors.

Case study for Web traffic privacy preserving:

The information-driven architecture also represents the network as traffic flows and hence the interpretations could be used for the dynamic specified SSH protocols. The information-driven architecture better describes traffic as streams as well as the interpretations have been used as an input vector for both the dynamically specified algorithms for SSH traffic analysis. Throughout this case, the purpose of each method is to map each aspect of the traffic flow for SSH and NonSSH (characterized by a vector attribute). By attaching a client device to 4 SSH servers anywhere in our testbed via the Internet, we simulate an SSH link, Figure 3. Our aim in this report is to compare/assess the utility of an expert controlled system and a data-driven System, in specific SSH traffic from traffic log files, to categorize encrypted network traffic without the use of IP addresses, port numbers or payload information. To meet this goal, we match the two strategies with almost the same data set. The data sets, their labeling methods and the measuring measures used to evaluate the latter strategies are described below.

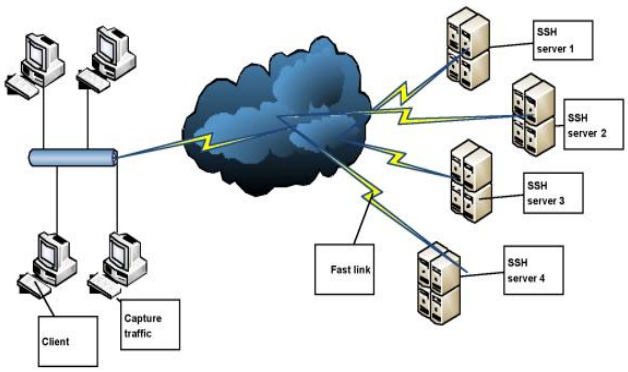

Figure 3. Simulating network traffic on local network and capturing this traffic

Dalhousie traces (UCIS) are labeled as PacketShaper, a proprietary device that would be a fast packet analyzer [12]. The latter thus gives us the foundation reality for the traces of Dalhousie. PacketShaper classified each traffic as SSH or Non-SSH. The attributes sets and rules are created by the optimization algorithms used for the detection/classification of SSH traffic. As already stated, two separate artificial optimization algorithms–RIPPER andC4.5 –are used in an information-driven framework methodology and one learning system is contrasted with the above expert-driven process.
Repeated Incremental Pruning to Produce Error Reduction (RIPPER): It is a rule-based model, which learns the laws explicitly from those in the data [13]. RIPPER discontinues incorporating laws in which the definition duration of the rule base is 64 bits (or several) longer than that of the strongest definition duration.
C4.5 decision tree: This is a process to shape a decision tree by determining the gain size, in which the significant gains will be seen as a preliminary node or root node. Next, it chooses the maximum gain cost factor, and then generates a branch for each feature. For both the divisions, too, the process continues for each branch until the branches has the same class in all situations.

IV. RESULTS AND DISCUSSION

WEKA will be the tool in this paper for applying data mining algorithms for simulation purposes. This tool supports several high-level languages like Java that is compatible with ML techniques for visualizing the patterns. The WEKA classification panel provides formulas to be used and evaluated with the data set. Also as discussed before, the data obtained from the simulation for the data-driven model by applying the C4.5 method generates improved recital with the NIMS and Dalhousie data sets than the expert driven system since Machine Learning algorithms can extract more patterns and can compute large data sets.

Root Mean Square Error (RMSE): It measures how much error there is between two data sets. In other words, it compares a predicted value and an observed or known value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n}} \quad (1)$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values

$y_1, y_2, \ldots\ldots y_n$ are observed values

n is the no.of observations

Mean Absolute Error (MAE): It is the average of all absolute errors. The formula is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - x\right| \quad (2)$$

Where:

n = the number of errors,

Σ = summation symbol (which means "add them all up"),

$|x_i - x|$ = the absolute errors.

Detection Rate (DR), in Eq. (3), and False Positive Rate (FPR), in Eq. (4). In this case, the DR indicates the number of SSH flows correctly categorized while FPR indicates the number of Non-SSH flows incorrectly categorized as SSH. Of course, it should be maintained that high DR rates as well as low FPR to be anticipated consequences.

They are calculated as follows

$$DR = 1 - \frac{\#FN\ classicications}{Total\ no.of\ SSH\ classifications} \quad (3)$$

$$FP = 1 - \frac{\#FP\ classicications}{Total\ no.of\ Non\_SSH\ classifications} \quad (4)$$

Table.1 Classifiers types

Classifiers	Correctly classified instances using 10-fold cross-validation (%)
ZeroR	89.87
Modified J48	98.04
Naïve bayes	41.32
OneR	63.63

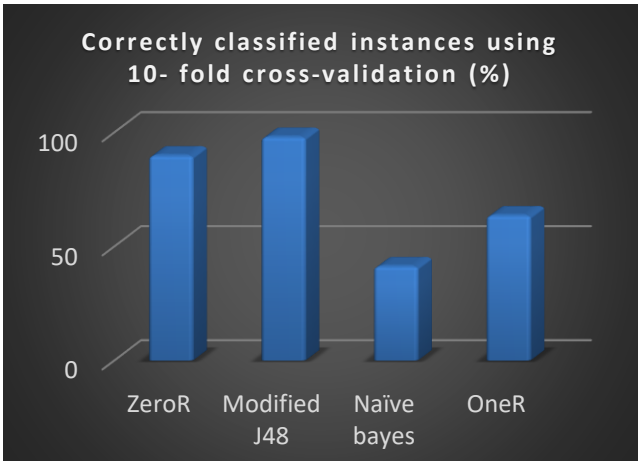

Figure 4. Comparison of classifiers accuracy

Table.2 Classifiers types and its MAE

Classifiers	Mean Absolute error
ZeroR	0.05561
J48	0.0024
Naïve bayes	0.0599
OneR	0.0154

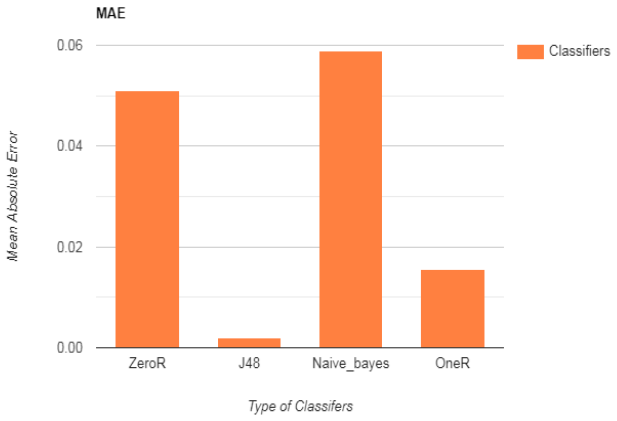

Figure 5. Comparison of classifiers for mean of absolute error

Figure.5 clearly indicates that the MAE of Modified J48 is less as related to the other techniques.

Table.3 Comparison of proposed model RMSE with other classifiers

Classifiers	Root mean squared error
ZeroR	0.1659
J48	0.0386
Naïve bayes	0.1552
OneR	0.1238

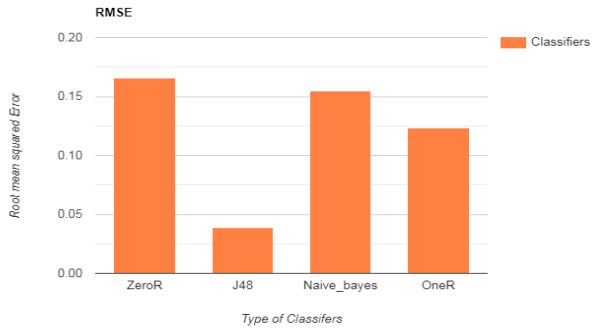

Figure 6. Comparison of proposed model RMSE with other classifiers

Figure.6 clearly indicates that the RMSE of Modified J48 is less as related to the other techniques.

Table.4 Comparison of proposed model error classification with other classifiers

Classifiers	Classification error (%)
ZeroR	7.28
J48	0.42
Naïve bayes	5.64
OneR	3.83

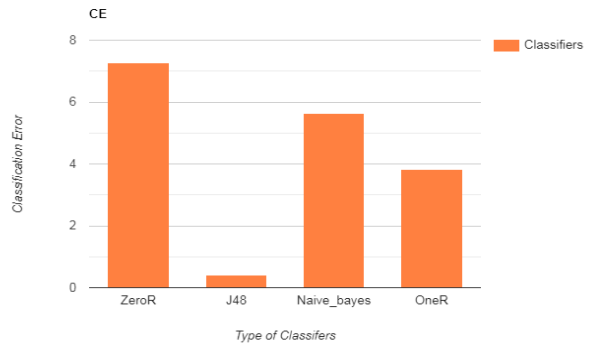

Figure 7. Comparison of proposed model error classification with other classifiers

Figure 7 designates the SSH and Non-SSH attribute classification error for all classifiers. On the other hand, these results are generated from the Weka tool for 10 cross-validations. J48 has fewer errors of classification due to the mother node can distinguish the correct class with the help of child nodes than that of other techniques.

Table.5 Classifiers types and their DR and FPR

Classifiers	Detection Rate (DR) (%)	False Positive Rate (FPR) (%)
RIPPER	97.2%	1.2%
C 4.5 decision tree	99.9%	0.4%

V. CONCLUSION

The current study is based on Modified J48 to evaluate web traffic's influence on the weblog dataset. The research has shown the potential use of this method to derive relevant information from current secondary web data. Based on the analysis, we can determine the correct classification strategies for log mining the Bayes Network and the Naive Bayes algorithm. In our tests, in the worst-case scenario, the C4.5 classifier would identify SSH traffic by obtaining a 99.9 percent DR and 0.4 percent FPR score (in one network, but another test).

REFERENCES

1. Kobsa, A. (2001). Generic user modeling systems. User Modeling and User-Adapted Interaction, 11, 49–63.
2. Cavoukian, A. (2008). Privacy in the clouds. Identity in the Information Society, 1, 89–108
3. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N. (2002): Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations 1, pp. 12-23.
4. SSH, http://www.rfc-archive.org/getrfc.php?rfc=4251
5. Parth R. Agarwal, "Semantic Web In Comparison to Web2.0", 2012 Third International Conference on Intelligent Systems Modelling and Simulation, DOI 10.1109, IEEE, 2012.
6. WANG Yong-gui and JIA Zhen, "Research on Semantic Web Mining", 201O International Conference on Computer Design and Applications (ICCDA 2010), IEEE, 2010.
7. R. Guha, Rob McCool and Eric Miller, "Semantic Search", WWW2003, May 20-24, 2003, Budapest, Hungary. ACM 1.58113-680-3/03/0005.
8. NehaSharma & PawanMakhija (2015), "Web usage Mining: A Novel Approach for Web user Session Construction", Global Journal of Computer Science and Technology: E Network, Web & Security Vol.15, No 3, PP: 15-20.
9. ChhaviRana (2012), "A Study of Web Usage Mining Research Tools", Int. J. Advanced Networking and Applications, Vol.3, No.6, PP: 1422-1429.
10. Mohammad Amin Omidvar, Vahid Reza Mirabi And NarjesShokry (2011), "Analyzing The Impact Of Visitors On Page Views With Google Analytics", International Journal of Web & Semantic Technology, Vol.2, No.1,PP:14-32.
11. Spilipoulou M., Mobasher B, Berendt B. (2003) "A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis," INFORMS Journal on Computing spring.
12. PacketShaper, http: //www.packeteer.com /products /packetshaper/, last accessed Jan. 2008.
13. Alpaydin E., "Introduction to Machine Learning", MIT Press, ISBN: 0-262-01211-1.

AUTHOR PROFILES:

Abha Narwal is associated as Research Scholar with Department of Computer Science and Application, Kurukshetra University, Kurukshetra Haryana. Her current area of interest is in data mining and Data science.

R. K. Chauhan is currently working as a Professor in Department of Computer Science and Application, Kurukshetra University, Kurukshetra Haryana and has also served as chairperson from 2007 to 2010. He has guided 18 PhD students in different domains and has published more than 150 research papers in National & International Journals.