# Classification Technique for Improving User Access on Web Log Data

**Bina Kotiyal, Ankit Kumar, Bhaskar Pant and R. H. Goudar**

**Abstract** In the present era, Internet is playing a significant role in our everyday life; therefore, it is very thorny to survive without it. Web log file that keeps track of the users' access on net, if mined, can provide us precious information about the surfers. Similarly, the rapid growth of data mining applications has shown the necessity for machine learning algorithms to be applied to large-scale data. In this paper, we are using the naïve Bayesian (NB) classification technique using Weka for identifying the frequent access pattern. The main objective of this paper is to categorize browsing behavior of the user based on their position. This paper performs experiment and classifies the user access behavior from the large databases, which could result in increasing the efficiency and effectiveness of the system by reducing the browsing time of the user or results in fast retrieval of information from the system.

**Keywords** Data mining · Weka · Classification · Web data · Web usage mining · Preprocessing · Pattern discovery · Naïve Bayesian

B. Kotiyal (✉) · R. H. Goudar
Computer Science and Engineering Department, Graphic Era University, Dehradun, India
e-mail: kotiyalbina@gmail.com

R. H. Goudar
e-mail: rhgoudar@gmail.com

A. Kumar · B. Pant
Information Technology Department, Era University, Dehradun, India
e-mail: mrankitgoyal@gmail.com

B. Pant
e-mail: pantbhaskar2@gmail.com

D. P. Mohapatra and S. Patnaik (eds.), *Intelligent Computing, Networking, and Informatics*, Advances in Intelligent Systems and Computing 243,
DOI: 10.1007/978-81-322-1665-0_111, 

## 1 Introduction

Data mining focuses on the techniques of non-trivial extraction of implicit, previously unknown, and potentially useful information from very large amounts of data [1]. In relation to this, the rapid growth of Internet technology specifically promoted Web mining by applying data mining techniques to Internet data. Among Web mining categories, Web usage mining (WUM) addresses mining of Web log record, which has the following attributes such as the host name or IP address, remote user name, login name, date stamp, retrieval method, HTTP completion code, and number of bytes in the file retrieved [2]. Data mining has several techniques: association rule, classification, cluster, sequential pattern, and time series. However, the vital ultimate purpose of WUM is to discover useful knowledge from Web users' interactive data in order to fulfill the need of the users more efficiently and effectively. Classification method in Weka for Web data classifies the data set which can be further used to enhance the working of the system. However, a major problem in Web data analysis is the loss of focus on important information because lots of ambiguity is there in Web information [3]. This paper describes an existing algorithm Naïve Bayesian (NB) implementation through Weka for classification of Web pages based on the navigation behavior of the users based on their position. The NB approach is one of the most effective and simple method for classification and has exhibited fine results in previous studies conducted for data mining.

The rest of the paper is organized as follows. Section 2 presents a literature survey on previous work done in Web usage mining. In Sects. 3, 4, 5, we will discuss Weka tool, classification techniques, and Naïve Bayesian algorithm. Section 6 presents our method of classifying user access behavior based on certain parameters using NB technique through Weka tool. And at last, Sect. 7 discusses the results of our experiment. The last section concludes the paper.

## 2 Literature Survey

In order to construct a common ground for discussing mining sources, an in-depth review of WUM is presented in Sect. 2.1 for aptly inferring WUM data sources in Sect. 2.2.

### *2.1 Web Usage Mining*

According to Han and Kamber, applying data mining techniques to Internet data covers a new area of Web mining. However, the Web also poses novel challenges to effective data sourcing and knowledge discovery due to its size, the complexity of Web pages, its dynamic nature, the broad diversity of user communities, and the

low significance of useful information [2]. Accordingly, Web mining has been developed into three categories, including Web structure mining that identifies authoritative Web pages, Web content mining that classifies Web documents automatically or constructs a multilayered Web information base, and WUM that discovers users' access patterns of Web pages [4]. From the data source standpoint, both Web content and Web structure mining target the Web content, whereas WUM targets the Web access logs. WUM includes three major processes: the pre-processing, pattern discovery, and pattern analyzing [2].

One of the very important steps in WUM is preprocessing. Pre-processing performs processing on Web log files, covering data cleaning, user identification, session, session identification, path completion, and transaction identification. After this phase, data mining algorithms are adopted such as generating classification, clustering, and association rules.

WUM could be useful in practice and proposes a user browsing behavior model which assumes that a given user's interaction of each page is either for the purpose of 'navigation' or for the purpose of 'actual content,' and this is determined by the page references and associated time obtained in Web server logs [2]. Another example is given in the work of who attempted to decide a Web user's next navigation by defining two types of users based on the navigation strategy: the 'net surfer' who is interested in exploring the cyberspace and the 'conservative user' who is concerned with the contents of a certain site [5]. These two examples show that raw Web log records are used by WUM algorithms to gather the browsing behavior of the users in an application domain.

According to a recent study, Web log records have been conducted to analyze system performance [6], improve system design by Web caching [7], identify best places for Web advertisement [8], predict best browsing paths [9] forming new approaches for efficient search engines [10], and build adaptive Web sites [11].

WUM embeds a connection among browsing data, browsing behavior, and WUM application, which in turn shows similarity to the relationship of data, information, and knowledge. In other words, the Web log data are processed by some mining algorithms, which lead to useful information such as browsing behaviors or patterns, and these are then applied in practice by knowledgeable workers.

## *2.2 Data Source*

The major advantage of Web server logs is their availability and handiness; however, they cannot accurately lock individual or recognize interactive Web pages with dynamic contents and thus may direct to biases in analyses.

WUM can be more effective data sources other than Web log records for personalized applications. A general architecture for WUM has been proposed [2], in which the data come from Web server logs, referral logs, registration files, index server logs, and document and usage attributes. A Web browsing strategies given in a paper describe all possible crossing-point events. [12].

## 3 Weka

Weka is an open source application that is freely available under the GNU *General Public License* agreement. Initially written in C, now Weka application has been completely rewritten in Java and is compatible with almost every computing platform [3]. It provides graphical interface that makes it user friendly. The idea of Weka project is to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and practitioners alike [13].

In our paper, we will provide the filtered data file, which is a log file in the format of Excel spreadsheet, and is converted into CSV format, as CSV format is the format, which is compatible with Weka. The Weka tool is used to categorize the data sets into different classes, which in turn will help to the fast retrieval of information.

## 4 Classification Technique

Classification techniques build (automatically) a model that can classify a class of objects so as to predict the classification or missing attribute value of future objects (whose class may not be known). This is a two-step process [14]. First-step process is based on the collection of training data sets, in which a model is constructed to describe the characteristics of a set of data classes or concepts. This step is also known as supervised learning since data classes or concepts are predefined (i.e., which class the training sample belongs to is provided). In second-step process, the model is used to predict the classes of future data. Various algorithms are used to classify the filtered data set of the log file, such as NB, decision tree.

In this paper, we are using Naïve Bayesian algorithm for the classification or categorization of data sets.

## 5 Naïve Bayesian Algorithm

The Naïve Bayesian classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. It is based on Bayes' theorem with independence assumptions between predictors [15]. Naive Bayesian model is easy to build and does less number of iterations, which makes it good for large data sets. In many practical applications, Naïve Bayesian model utilizes the method of maximum likelihood for parameter inference [16].

Naive Bayesian classifiers have worked quite well in many intricate real-world situations even of their naive design and in fact over-simplified assumptions. The analysis of Bayesian classification problem in 2004 is shown, and there are some theoretical reasons for the apparently unreasonable efficiency of naive Bayesian

classifiers [17]. However, still, an inclusive comparison with other classification methods in 2006 showed that Bayesian classification outperformed other approaches, such as boosted trees or random forests [18].

Advantage of the naive Bayesian classifier is that it only needs a small amount of training data to guess the parameters (means and variances of the variables) essential for classification because autonomous variables are assumed. Only the differences in the variables for each class must be determined and not the whole covariance matrix.

## *5.1 Accuracy and Error Measures*

The performance of the classifiers can be evaluated based on accuracy measures. Factor that measures the accuracy are as follows: true-positive rate, false-positive rate, precision, recall, F-measure, ROC, and confusion matrix. These are explained as follows:

**TP Rate**: It is the proportion of positive tuples that are correctly classified. Sensitivity is also referred to as TP rate.

$$\text{TP Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

**FP Rate**: It is the rate of negatives tuples that are incorrectly labeled/placed.

$$\text{FP Rate} = \frac{\text{FN}}{\text{FN+TP}} \qquad \text{FP Rate} = \frac{\text{FP}}{\text{TP+FP}}$$
$$\text{FP Rate of Class ``Yes''} \qquad \text{FP Rate of Class ``No''}$$

**Precision**: It is the proportion of predicted positives, which are actual positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

**Recall**: Recall is the proportion of actual positives, which are predicted positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

**F-Measure**: A combination of both precision and recall that is why it aggregates the two measures.

$$\text{F} - \text{Measure} = \frac{2 \times \text{Pr} \times \text{Re}}{\text{Pr} \times \text{Re}}$$
$$\text{Pr} = \text{Precision}$$
$$\text{Re} = \text{Recall}$$

**Error Measures**: Along with accuracy, error measures are used to tell how well the instances are classified from the actual known value.

## 6 Experiment

The experiment is conducted on the college data set by using classification algorithm (NB) in Weka tool. We took into account some portion of log files during the 8-h time period in a day. The file used is a preprocessed one in which all irrelevant information are removed, and the filtered file is in the CSV format. Hundred instances were taken for classification that classifies the three classes and results in reducing the browsing time of the users.

### *6.1 Steps for Performing the Experiment*

**Algorithm Applied**: Naïve Bayesian
**Input File**: Data_Set.csv
Step 1: Open Weka *Explorer*
Step 2: Select the file from the *open file* and load the file.
Step 3: Select *Classify* features and then select the *Naive Bayes Algorithm* for classification.
Step 4: Select the *Cross-Validation Folds*. The value taken is five in our experiment for the cross-validation.
Step 5: Now click on *Start* and it will display the correctly classified and incorrectly classified classes. This is shown in Fig. 1.

## 7 Results

In the result section, we can see from the Fig. 2 that Weka using NB has classified 100 instances. And as per the result, the correctly and incorrectly classified instances are

| | | | |
|---|---|---|---|
| Time taken to build model: | 0 s | | |
| Correctly classified instances | 80 | 80 | % |
| Incorrectly classified instances | 20 | 20 | % |

And the confusion matrix shows

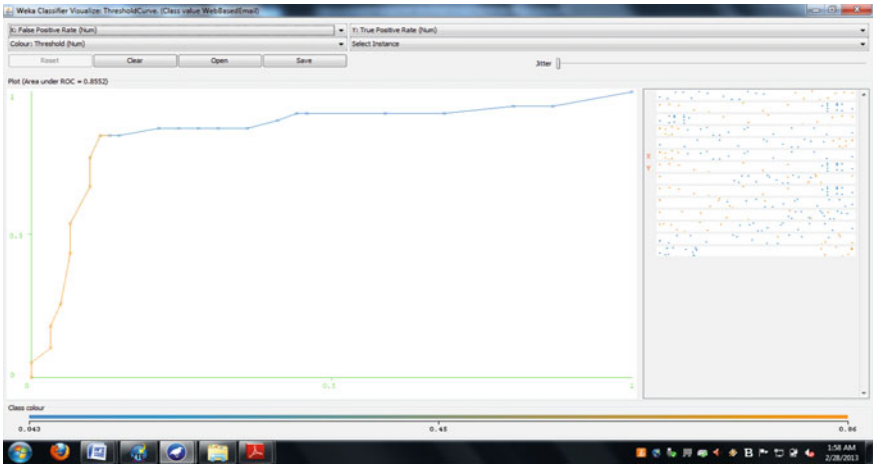

**Fig. 1** ROC showing the true-positive (*TP*) and false-positive (*FP*) rating for the WebBasedEmail

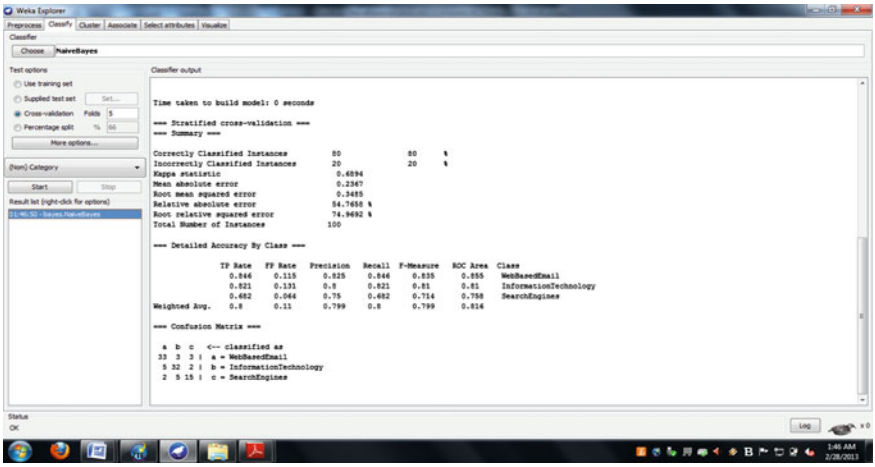


**Fig. 2** Showing the incorrectly and correctly classified classes

Confusion Matrix

| A | B | C | Classified as |
|---|---|---|---|
| 33 | 3 | 3 | $a$ = WebBasedEmail |
| 5 | 32 | 2 | $b$ = Information technology |
| 2 | 5 | 15 | $c$ = Search engines |

The agreement for the confusion matrix is checked diagonally, which again shows that 80 instances out of 100 are correctly classified. As a class has 40 instances, out of that 33 are correctly classified, 3 are incorrectly classified under *b*, and 3 are again incorrectly classified under *c*.

Receiver Operating Characteristic (ROC) Curve

The *ROC Curve* is shown for the WebBasedEmail. The true-positive rate is along the *y*-axis, and the false-positive rate is along the *x*-axis. The color shows the value of the threshold.

## 8 Conclusion

In this paper, we performed an experiment on a filtered college data using the Naïve Bayesian algorithm using Weka tool, which is helpful in finding the user access pattern and the order of visits of the college staff. Naïve Bayesian algorithm can perform multi classification. Once the classification of the user is done, then it also results in increasing the efficiency and effectiveness of the system by reducing the search browsing time of the user by adding the classified links to the Web pages for the types of users. The drawback of NB classifier is that it cannot classify the undefined classes.

## References

1. Agrawal, R., Mehta, M.: SPRINT: a scalable parallel classifier for data mining. The International Conference on Very Large Database, pp. 544–555. Bombay, India (1996)
2. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Academic press (2001)
3. Nasa, C., Suman, S.: Evaluation of different classification techniques for web data. Int. J. Comput. Appl. (0975–8887) **52**(9) (2012)
4. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining World Wide Web browsing patterns. J. Knowl. Inf. Syst. **1**(1), 5–32 (1999)
5. Cunha, C.R., Jaccoud, C.F.B.: . Determining WWW user's next access and its application to pre-fetching. In: The second IEEE Symposium on Computers and Communications, Alexandria, Egypt (1997)
6. Iyengar, A., MacNair, E., Nguyen, T.: An analysis of Web server performance. In: The IEEE Global Telecommunications Conference, vol. 3, Phoenix, AZ, pp. 1943–1947 (1997)
7. Bonchi, F., Giannotti, F., Gozzi, C., Manco, G., Nanni, M., Pedreschi, D.: Web log data warehousing and mining for intelligent web caching. Data Knowl. Eng. **39**(2), 165–189 (2001)
8. Chen, Z., Shen, H.: A study of a new method of browsing path data mining. In: The sixth International Conference of Information Management Research and Practice. TsingHua University, HsingChu (2000)
9. Chen, M.S., Park, J.S., Yu, P.S.: Efficient data mining for path traversal patterns. IEEE Trans. Knowl. Data Eng. **10**(2), 209–221 (1998)
10. Zhang, D., Dong, Y.: A novel Web usage mining approach for search engines. Comput. Netw. **39**(3), 303–310 (2002)

11. Perkowitz, M., Etzioni, O.: Towards adaptive Web sites: conceptual framework and case study. Artif. Intell. **118**(1–2), 245–275 (2000)
12. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World Wide Web. Comput. Netw. ISDN Syst. **27**(6), 1065–1073 (1995)
13. Mark Hall: The WEKA Data Mining Software: An Update, SIGKDD Explorations, vol. 11(1) (2009)
14. Pani, S.K., Panigrahy, L.: Web usage mining: a survey on pattern extraction from web logs. Int. J. Instrum. Control Autom. (IJICA) **1**(1) (2011)
15. Santra1, A.K., Jayasudha, S.: Classification of web log data to identify interested users using Naïve Bayesian classification. Int. J. Comput. Sci. Issues (IJCSI) **9**(1), 2 (2012)
16. Patil, A.S., Pawar, B.V.: Automated classification of web sites using Naive Bayesian algorithm. In: Proceedings of International Multi-Conference of Engineers and Computer Scientists, vol. 1 (2012)
17. Zhang, H.: The optimality of Naive Bayes. FLAIRS 2004 Conference. Available online: PDF (http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf)
18. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning (2006). Available online PDF (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.5901&rep=rep1&type=pdf)