Analysis of Some Popularly Used Techniques of Click Stream Analysis

[1]Umang Goyal, [2]Neeraj Sharma, [3]Kawaljeet Singh
[1]Research Scholar, [2]Professor, [3]Director
[1]Department of Computer Science, Punjabi University, Patiala
[2]Department of Computer Science, Punjabi University, Patiala
[3]University Computer Centre, Punjabi University, Patiala
[1]u.goyal60@gmail.com, [2]sharma_neeraj@hotmail.com, [3]singhkawaljeet@rediffmail.com

ABSTRACT

Websites and other online business marketing are becoming more effective and powerful way to deal and interact with users. The research study focuses on the working and techniques of click-stream analysis. Click-Stream Analysis is a comprehensive body of data that is used for describing the sequences of all the activities that have been happened between a user's browser and the other internet resource like a website and a third party ad server. The website chosen www.indiatourandtrip.com which is used for knowing the interest of the customers so that we can enhance the business and make it better. We have used data mining techniques for the extraction of valuable data, this data have been taken in order to predict the users behaviour and their interest. The research study focus on different classifiers and the performance evaluation of each classifier is done to get better results. The main motive of this evaluation is to upgrade the tour and travel site in order to make it more convenient for booking the tour packages and hotel on reasonable value price and it's the effective way to optimize the website for the improvement of the booking and marketing with the help of software called MATLAB.

I.INTRODUCTION

A Click-Stream is the demonstration of recording the search and clicking history on a web program or some other programming application. The moment a client clicks any place on the website page or the web application, that activity is put away or recorded on the web server as customer signed in [6]. The utilization data about client is recorded in web logs. Analyzing web log records to separate useful patterns is called web utilizing mining. Clickstream is an ordered sequence of website page saw by a client that is a session of visited site pages; pages are displayed one by one of every a column at any given moment. Finish click grouping contain the data of all clients and analyzing what number of URLs the client followed by the search result. Clickstream analysis is utilized for the extraction of the data from the log record. Log file produced by web server contains huge measure of site page information that is significant for understanding the movement of site guest [2]. Time spend on a specific page is a well evaluation of the client consideration in that site page. In this given analysis session of the client all together to compute significance of the URLs.

II.METHOD OF EXPERIMENT

Implementation of the scheme has been done through the MATLAB. MATLAB stands for matrix laboratory. It is distributed by Math Works Inc. It is a multi-paradigm numerical processing environment. It is fourth-era programming language. It is developed by MathWorks. It allows matrix manipulations, plotting of functions and data, execution of algorithms, formation of UIs, and interfacing with programs written in different languages, including C, C++, Java and Python. In spite of the fact the MATLAB is projected fundamentally for numerical computing, an optional toolbox uses the MuPAD symbolic engine. An extra package, Simulink which includes graphical multi-domain simulation and Model-Based Design for embedded frameworks. It gives graphical yield to the client.

III.TECHNIQUE PROPOSED

The research aims on the study of indiatourandtrip.com website where People from different places visit this website to book different tour packages, hotels etc. By the use of Click Stream Analysis we get to know the interests of different users.

- **Overview of Indiatourandtrip.com**

Indiatourandtrip.com is an E-Commerce website to develop for customers, who want to visit at different places in India. This site helps the customers to book car as well as hotel according to their requirements and interests. The website offers links to a large variety of tour packages but does not concern itself with the end transaction.Indiatourandtrip.com links are grouped into categories such as Visiting places, Hotel booking, Tour Packages etc. Our objective is to create a system that allows users to find what they are looking for with greater ease through collaborative user profiling. The system should be able to learn a profile model based on past user interactions with the site and then use the model to modify the website's content so as to better assist future users. The user profile and web page content adaptation should be tightly coupled to the website HTML generation so as to provide a rapid response time. The Indiatourandtrip.com website logic will track the categories that a user visits in sequence – this is often referred to as a "click stream" [9]. A click stream captures the behaviour of a previous user because it describes the path taken by that user through the site in search of a product. In aggregate, the click streams of many previous users provide the raw data for developing a model for predicting the category a new user is likely to next wish to visit. Different classification methods have been l be used to develop the model. Given the current category as input, the classifier model will output the next best category links.

- **Implementation**

The implementation of our solution is broken down into feature extraction, feature training and testing components. The first component collects the click-stream data generated by users and transforms that data so that it can be used as training examples for a classification system. Then training has been carried out by five different classifiers i.e. ANN, KNN, SVM, Naïve-Bayes and Decision trees.

- **Data Collection and Preparation**

The first step in building classifiers training data is collecting a set of training examples. For our problem these training examples are users' click-streams through the site. First each category is given a numeric identifier which depends upon the page links. In order to track where a user goes we created a cookie that keeps track of every category a user visits. All categories are similar in that each has a webpage that consists of all the links that lead to clients to different pages. When accessed, each of these pages writes its particular identifier to the cookie. There are 45 unique pages in the website which are divided into five category types. After a visit to Indiatourandtrip.com, a user's click stream might look like the following:

Table 1: Click streaming example of an IP address

Click Streaming example of an IP address	
Click From	**Clicks To**
6	28
28	3
3	40
40	7
7	41
41	39
39	44
44	24
24	22
22	32
32	28
28	32
32	44
44	23
23	41
41	6
6	21
21	38
38	1

From the stream we must create training examples that can be used by a classifier. Chosen classifiers are excellent at learning sequences such as the user moved from category 6 to 28, 28 to 3, 3 to 40,40 to 7, then from 7 to 41, then from 41 to 39, etc. Breaking this down, the training examples for a classifier must consists of one input and one output as follows:

(6 28),(28 3),(3 40), (40 7), (7 41), (41 39), (39 44), …

Before we can ask the classification system to learn these examples, more data preparation is needed. The page categories are nominal in value - they cannot be placed on a metric scale from 1-5. Therefore, the category identifier must be transformed from a numeric value to individual discrete variables that can be properly learned by the classifiers. In our implementation a category is represented by a series of 5 '0's, with only a single '1' appearing in the nth place, where n-1 is the category number (keep in mind we are starting from 0). For

example, the category 2 would be represented as 0 1 0 0 0. In summary, the click-stream is first gathered into a sequence of pairs of 2 category numbers (input and output) and those number pairs are then transformed into a series of '0/1' representations. The resulting binary representation is used to train the neural network.

- **Procedure 1** (Feature Extraction)

1. Read log file and divide it into three columns that is page, date and time and IP address of user.
2. Convert the page strings into discrete form and give a unique number to each page link.
3. Convert the IP address into unique discrete representatives so that unique IP address ID's can be represented by a unique number.
4. Divide each page link into a particular category as there are five different categories available for the generated website that are booking, hotel booking, Single link, Tour Packages & Visiting places.
5. Generate Sessions according to unique IP address ID and extract click streaming sequences for each unique user.
6. Repeat the step 5 for each unique IP address in order to get click Streaming Sequences for each user.
7. Generate pair based input-output pattern from click streaming sequences generated in previous steps in which three columns will be generated that are click from, click to and category of clicked page.
8. Repeat step 7 for all the clicks and concatenate the features into a single array.
9. The column 1 and column 2 are feature vectors for training and testing purpose.
10. Add Five more columns to add Boolean features vector based on category of output page link.

- **Procedure 2** (Classification)

1. Read Input feature vector generated in Procedure 1.
2. Read Category column of each feature set as a ground truth or target matrices based on which training has been carried out.
3. Train the feature sets using Five different classifiers that are KNN, ANN, decision tree, Naive Bayes, SVM.
4. Test the feature sets using five different classifiers.
5. Evaluate Performance in terms of Sensitivity, Specifying and accuracy parameters for all classifiers.

IV.PERFORMANCE EVALUATION

A. Classification Accuracy

The classification accuracy is the extent to which the classifier is able to correctly classify the exemplars and is summarized in the form of confusion matrix for testing the data. This is defined as the ratio of the number of correctly classified patterns (TP and TN) to the total number of patterns classified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \text{.....(1)}$$

B. Sensitivity

The sensitivity of a classifier is the fraction of the Plant samples correctly classified as that specific species class. It is defined by the equation below:

$$Se = \frac{TP}{TP + FN} \text{.....(2)}$$

C. Specificity

It is the fraction of normal species correctly classified as normal class. It is also known as selectivity.

$$Sp = \frac{TN}{TN + FP} \text{.....(3)}$$

Confusion matrix has been evaluated for all discussed classifiers. But, Decision tree shows most effective results as shown in below table. It shows the total no. of categories is five and total no. of pairs in decision tree is 1336 for all the categories. It also shows exact no. of pairs for each category.

stats_decision_tree.confusionMat

	1	2	3	4	5
1	334	0	0	0	1
2	0	176	0	0	0
3	0	0	536	0	0
4	0	0	0	119	0
5	0	0	0	0	171

Figure 1: confusion matrix after testing using decision tree

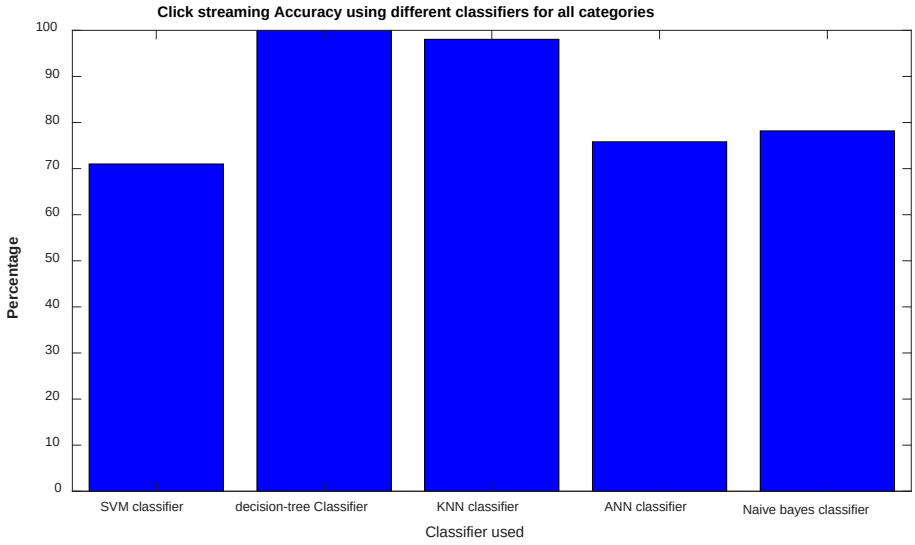

Figure 2: Click streaming Accuracy using different classifiers for all categories

Fig 2 shows that accuracy of the click stream analysis using different classifiers. It is clearly shown decision tree classifiers give the more accuracy for the click stream analysis as compared to others.

Table 2: Average accuracy of category prediction using different classifiers

Average accuracy of category prediction using different classifiers				
SVM Classifier	**Decision tree Classifier**	**KNN Classifier**	**ANN Classifier**	**Naïve-Bayes Classifier**
71.009	99.9700	98.055	75.82	78.18

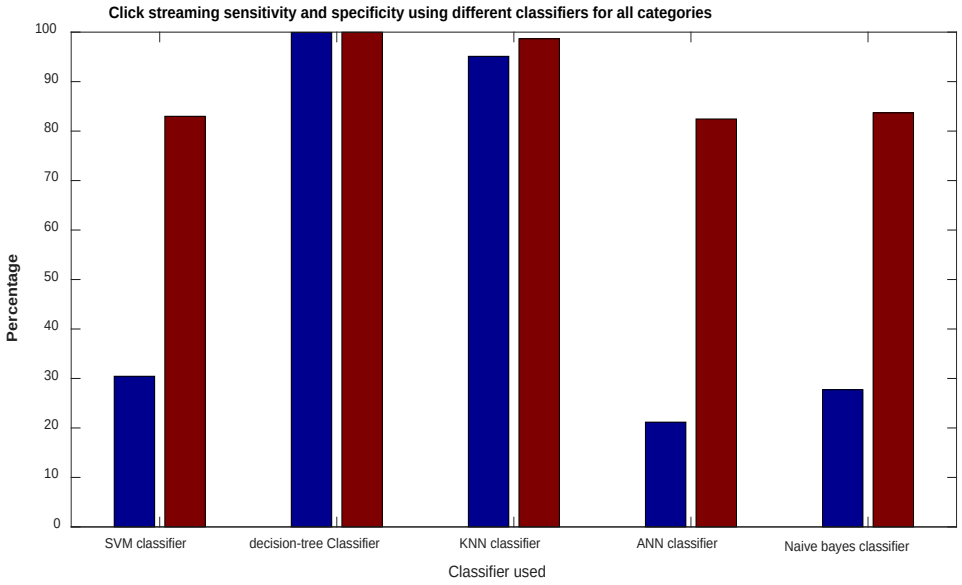

Figure 3: Click streaming sensitivity and specificity using different classifiers for all categories

This figure shows the sensitivity and specificity of the click stream analysis using different classifiers for all categories.

Table 3:Average sensitivity and specificity of category prediction using different classifier

Parameter	**Average sensitivity and specificity of category prediction using different classifiers**				
	SVM Classifier	**Decision tree Classifier**	**KNN Classifier**	**ANN Classifier**	**Naïve-Bayes Classifier**
Sensitivity	30.44	99.94	95.10	21.17	27.73
Specificity	82.98	99.98	98.66	82.44	83.69

V.CONCLUSION

Analysis of click-stream shows how a website is navigated and used by its visitors- by determining the users who are visiting the site and user's interest. We have implemented five different types of classifiers on different categories that is decision trees, SVM, KNN, ANN and Naïve Bayes to evaluate the results based on user's behaviour and interests. After the evaluation of the classifiers, decision trees have proven to be the best classifier by giving results with more accuracy and clarification. The website indiatourtrip.com has been taken as a platform to applied different type of classifiers and 5 categories to determine user's interest and behaviour. The future scope of the click stream analysis is to used it or implementing it for the website that have heavy traffic of the visitors in terms of number of clicks and booking by using more classifiers and more categories. . Furthermore we can also try to improve the other techniques of click stream analysis which have not given better result as the decision tree so in the future there is a possibility to work more on these techniques so that these techniques can also deliver best results for click stream analysis.

REFERENCES

[1] G, R.C.a.K., "An Efficient Preprocessing Methodoloy for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network," Fifth International Conference on Information Processing, 2011. Springer -Verlag.

[2] Maheswara Rao.V.V.R and Valli Kumari.V, "An Enhanced Pre-Processing Research Framework for WebLog Data Using a Learning Algorithm," Computer Science and Information Technology, DOI:, pp. 1-15, 2011.

[3] C. Kaushal and H. Singh, " Comparative study of recent sequential pattern mining algorithms on web clickstream data," *2015 IEEE Power, Communication and Information Technology Conference (PCITC)*, Bhubaneswar, 2015, pp. 652-656.

[4] Losarwar, Vijayashri and Joshi, Dr. Madhuri (2012) "Data Preprocessing in Web Usage Mining".

[5] Kansara, Akshay and Patel, Swati (2013) "Improved Approach to Predict user Future Sessions using Classification and Clustering", ISSN: 2319-7064, Volume 2, Issue 5.

[6] A. Surya and D. K. Sharma, "A comparative analysis of clickstream as web page importance metric," *2013 IEEE Conference on Information & Communication Technologies*, JeJu Island, 2013, pp. 776-781.

[7] Qiang Su, Lu Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data" Published in: Electronic Commerce Research and Applications Volume 14, Issue 1, January–February 2015, Pages 1-13

[8] Juhnyoung Lee, Mark Podlaseck, Edith Schonberg, Robert Hoch, "Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising" Published in: Applications of Data Mining to Electronic Commerce pp 59-84

[9] F. Nottorf, "Modeling the clickstream across multiple online advertising channels using a Bayesian mixture of normals", Issue 1, Volume 13, 2014, pp. 45-55.

[10] Gang Kou, Chunwei Lou, "Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data" Published in: Annals of Operations Research August 2012, Volume 197, Issue 1, pp 123–134.

[11] Navjot Kaur, Dr. Himanshu Aggarwal, "Survey of Various Techniques for WebUsage Minning" Published in: An International Journal of Engineering Sciences ISSN: 2229-6913 Issue Dec. 2011, Vol. 5.

[12] Eesha Goel, "Data Warehousing and Data Mining in Business Applications" Published in: An International Journal of Engineering Sciences, Issue December 2014, Vol. 3, ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online).