International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 16, Issue 8 (August 2020), PP. 18-22

A Website Analytics System Considering User's Category

Ayush Sharma[1], Prashant Kumar Mishra[2], Snehal D Chaudhary[3]
[1,2,3]*Department of Information Technology,*
Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune,India.

ABSTRACT
One of the most important field of e- commerce is, Website analysis and it has been the top main concerns of the website administrator. The most focused or important issue in this field is, the interaction and the exploration of the web content done by the users and where their attention lies in the process of navigation. However, different studies have been made for the analysis of the all kinds of categories of users. It can be said that different categories of person have a different need from a particular website. A particular element may be useful fora particular person but it may be useful for the second one too, as both the persons don't show up the same interest.
Hence the paper proposes a website analysis method which is based on users' categories, categorizes users using two different methods of Data mining, Classification and Clustering. The first approach deals with the Organizational categories and the second approach deals with the users' behaviour on the website. The two approaches have been implemented on the University of Tehran's website. With the use of heat map, we have presented the important items in a page for respective category of users which can be used to restructure the pages appropriate to each category
KEYWORDS: *Web analysis, Heat Maps, Data Mining, Clustering, Data Classification*

Date of Submission: 15-08-2020 Date of Acceptance: 01-09-2020

I. INTRODUCTION

The companies that don‟t get engrossed in e- commerce can also be benefited from an effective website, as customers often find business through internet searching. There are hundreds of reasons a company may perform a website analysis. If we take up an example, the customers might show curiosity how well their website is functioning or what do their competitors do with the web pages. A website analysis can also help to determine the design of the particular site once it has been created. The focal pointof the website analysis to tell how supportive the site becomes for a company‟s goal or target.

As website analysis becomes the most important part in e-commerce and has been the main concern and challenge of the web administrator. To achieve the deep understanding the users‟ behaviour, web administrators need to keep analysing their website to know or examine the extent of their goals achieved, also develop the future goals and strategies, and try to improve the user experience. Hencethese tools help the administrator to identify the strength and the weakness of their website and develop the future strategies.

Web analysis

There are several numbers of tools for analysing the website like Google analytics1 and piwik2. These different tools offer a variety of reports, like website traffic, browser type, visitor‟s geolocation, most viewed pages and the different versions. In the present time, these tools are designed for the purpose like, CrazyEgg3,MouseFlow4and Hotjar5. These different tools analyse the behaviour of the users and presents a report which is based on heat map. It shows the usage of heat maps to display reports of website analysis. The Heat map is an illustrative diagram, where the colour is representative of the amount of the users click on the page or is also representative of the areas of the page which attracts more visitors. In the map, the warm colours are representative of the areas which have been most visited by the users whereas the cold colours are representative of the few visits. To maintain the growth of this graphical tool, one of the most important reason is the ability to make easy transfer of the concepts in different conditions, which synchronizes the data collected over a period of time and reduces the complexity of data and gives a comprehensive portrait of multi-dimensional analysis result. As the development of the mobile devices took place, the web analytic approach

was designed to modified touchbased interactions like, CrazyEgg3, MouseFlow4 and Hotjar which are designed to record.zoom and even drag gestures to display in heat map. Although, the devices which have a small screen , a touch event may not always represent a user‟s intended area. For example, if a person uses a smartphone, if the user touches the screen to scroll the page or to zoom the page and the fingers may be moved intentionally outside the region which contains the information of interest, to not occlude it.

Proposed Approach

A. Classification on the basis of user's type

The initial and the most focused step in data mining and classification of data preparation. The data represented in this paper includes the links that have been visited by the users. Google Analytic is one of the websites which is used widely by the user so, to automate data preparation process, we have used Google Analytic to represent the data.

„All Pages‟ reports are used to provide for training data, this includes the links that are being viewed by the different types of users. The data is required to label. The labels used in this research paper are the different types of users (student / faculty / staff). We have done the analysis with use of custom dimension. A custom dimension is used to analyse and collect the data, which is not auto – tracked by google. With the use of this feature, we have set user type as the custom dimension for the logon user and send to google analytic. To classify the access to different pages, we used Random forest, Decision tree, SVM and Naïve Bayesian classification.

B. Clustering users based on user's behaviours using web Access logs

Web server access logs has been the primary data sources for this approach. Data has not been labelled in this approach and user categorization has been done on the basis of user‟s behaviours. The access log contains the raw data, so, pre- processing of the data was beforehand needed.

1.) Data cleansing

Data cleansing or data cleaning – it is a process of detecting and correcting (or remove) the corrupt or the inaccurate records from record set , table or data base and it refers to identifying incomplete , inaccurate , incorrect or the irrelevant parts of the Data , and afterwards it replace , modify and delete the dirty or coarse data . It may be done interactively with data wrangling tools or as batch processing through scripting.

- Data indicates the access of crawlers and the bots
- Data which were requested by the user during the visit of a page without the explicit request, such as Java script, CSS, photos and more.

2.) Session Identification

User session can be defined as the collection of pages visited by a particular user during a visit to a website. In the approaches discussed in this paper the user‟sidentities were not needed. Though there is a need to differentiate between the users from each other. Important fact to remember is that a user may visit a website once twice or more than that hence, the server records multiple session for respective user.

If there is non- appearance of an authentication mechanism, we need another way to differentiate within the user. As the IP address of a particular computer that is used by user for accessing a website is not always unique. If we take this in simple words, there is a possibility that from a single IP address, we may have multiple access, also the browser version or operating system also can be different.

3.) Data Preparation

After the identification of the user‟s sessions, there is a need to create a data set for clustering process. The main purpose of dataset for this paper is to represent the number of links per visit of the user (page view) and this is obtained by calculating the difference between the first and the last user requests. Once the dataset is prepared the next step is clustering of data.

4.) Pattern Discovery

The main goal of clustering is to assign the data point to k cluster, Or in simple words,

„To impute label that indicate the membership of each data to the cluster‟. In this analysis we have used k-mean algorithm for clustering. The users are being categorized into three clusters (k =3) on the basis of time spent on the website and number of pages visited:

- Active users
- Moderate users
- Low activity users

3.3. EXPERIMENTAL RESULTS

A. Experimental Results of Classification

As explained in II.A, data collected from Google analytics and labelled them with users‟ types.
shows a part of the output obtained. For example, page with URL /fa/page/253:

• Viewed by 19506 staff/faculty members (user Type =
5).
• The average time this type of users spent on this page is a minute and 57 seconds.
• 2132 times this page was the first page in a session.
• 33.16% of entrances on this page where from userswho did not interact with the website any further.
• In 21.64% of pageviews this page was the final page

SVM also has a relatively better performance than the other algorithms because it is not sensitive to the number of dimensions. In this approach, the dataset is averaged, and the percentage and the number of data rows and features are low.

C.Ex***perimental Results of Clustering***

As explained in II.B, data collected, pre-processed, and prepared for clustering from the access log. is sample of data for clustering.

C. Heatmaps

• In this study, heat map representation is used to represent the importance of the pages based on users‟ activities.

II. METHODOLOGY

The main objective of this paper is to propose a method of concepts for the improvement of the website analytic system.

Currently there are number of tools for the analysis of website such as Google Analytics and Piwik2,eachrespective tool offers, different variety of reports, like – website traffic, visitor‟s geolocation, browser type, the different version and the most viewed pages. And these specific tools have been designed for Currently the purpose of Crazy Egg3, MouseFlow4 and Hotjar5. These tools are meant to analyse the behaviour of users‟ and present a report based on heat map. In [2] and [3], usage of heat maps is being represented to display the reports of the web analysis. Heat map can be defined as the illustrative diagram, where colours are used as the symbols to represent the amount of user clicking the website of a page, or for the areas which page has attracted how much visitors. Warm colours represent the areas most visited and the cold colours represent the fewer visit.

1 https://analytics.google.com/
2 https://piwik.com/
3 https://www.crazyegg.com/
4 https://mouseflow.com
5 https://www.hotjar.com

PROPOSED SYSTEM

The main aim of the proposed system is developing a system along with improved facilities of website analytic system. The system can be developed with the help of using Django and Python with a proper database connectivity. This system can break all types of limitation of the existing system. The new system provides a better accuracy than the previous applied algorithms. The main beneficial point of the website analytic reduces the dependence upon the expensive legacy system hardware and enables parallel processing of a very large amount of data across inexpensive,standard,commodity server is used to store and process the data without any volume limitation. When we look forward for more categories of users in the purpose for classification of the operations, like, alumni, regular visitors and breakdown of the staff and faculties of each one and create a separate class for each.

To design a script for collection of data and a set of data, for classification operations, instead of using Google analytics just to improve the accuracy of the classifier.

Google Analytics

It is web Analytics service which is to be offered by google itself that helps in tracking and reporting website traffic, and currently it is as a platform inside the google platform(marketing) brand.
Google has launched it in back November 2005 after acquiring urchin.

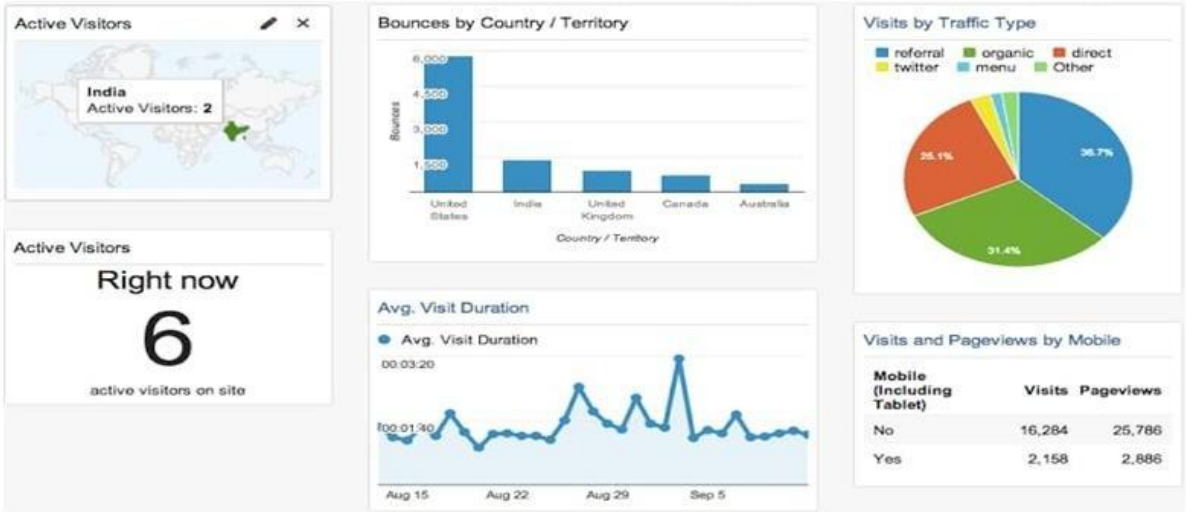
As the survey of 2019, Google analytics is the most used web analytics tool used. It provides a software development kit that allows collecting usages data from android, iOS and other apps. Which is also known as Google Analytics for Mobile Apps. It may be blocked by browsers or browser extensions and various firewalls and others.

Get Started with Google Analytics

To Start gathering Basic Data from Our Website:

1. Create Your account or sign in to your account if you already have
Just go to google.com/analytics.
2. now the next step is we have to set up a property in the analysis account which represents your website and is the gathering point in analytics for the data from your site or app.
3. now the next step is we have to set up a reporting view in our propertywhich lets you create filtered perspectives of your data.

EXPECTED CONCLUSION



Combining the two approaches used in this study. First, users are classified according to the existing labels and then clustered. The results of these operations include categories such as active Users, moderately active User, and low active User.

This Research surely will give more accuracy than the SVM random Forest.

III. CONCLUSION

If we combine the two approaches used in this paper, in first we observe Users are being classified according to the existing labels and then they are clustered. The results of these operations were in categories such as active user, moderately active user, and low active user.
This Work would surely give more accuracy than SVM, Random forests.

REFERENCES

[1]. Jarvinen, J and Karjaluoto, H. "The use of web analytics for digital marketing performance measurement", Ind. Mark Manag, Vol.50, 2015, pp. 117-127.
[2]. Ehmke, C and Wilson, S "Identifying web usability problems for eye tracking data", 21st century British HCI group annual conference on people and computer, Vol.1, pp.119-128.
[3]. Chen, M.C, Anderson, J.R and Sohn, M.H "What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing", CHI"01 Extended abstracts on human factors in computing systems, 2001, pp.281-282.
[4]. Eichinski, P and Roe, P "Heat maps for aggregating bioacoustics annotations", Information visualisation (IV) 2014 18th internationalconference,2014, pp.88-93.
[5]. Lex, P, Partl, C, Streit, M and Schmalstieg, D "Comparative analysis of multidimensional quantitative data", 2010.
[6]. Lamberti, F, Member, S, Paravati, G, and Cannav, A "Supporting web analytics by aggregating user interaction data from heterogenous devices using viewport- DOM based heat maps", IEEE Trans. Ind. Informatics, vol.13, 4, 2017, pp.1989-1999.
[7]. Huang, J and Diriye, A "Web user interaction mining from touch-enabled mobile devices", HCIR workshop, 2012.
[8]. Park, S, Suresh, N.C and Jeong, B.K "Sequence- based clustering for web usage mining: A new experimental framework and ANN-enhanced K-means algorithm", Data Knowl. Eng., vol.65, no.3, 2015, pp. 512-543.
[9]. Pirolli, P, Pitkow, J and Rao, R "Silk from a sow"s ear: Extracting usable structures from the web", Proceedings of the SIGCHI conference on human factors in computing system, 1996, pp.118-125.

Ayush Sharma, et. al. "A Website Analytics System Considering User"s Category." *International Journal of Engineering Research And Development*, vol. 16(8), 2020, pp 18-22.