International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 781-784 www.ijaem.net ISSN: 2395-5252

An Approach for the Analysis of users' future request by Web Mining

[1]Hinal Rathod, [2]Anita Anand
[1] PG Scholar, Hinal J. Rathod, LDRP ITR, Gandhinagar, Gujarat, India
[2] Professor, Mrs. Anita Anand, LDRP ITR, Gandhinagar, Gujarat, India

Submitted: 05-05-2021 Revised: 17-05-2021 Accepted: 20-05-2021

ABSTRACT - In the web servers, log repositories plays a key role as it keeps record of user pattern for different users and thus it is great source of knowledge. Web Usage Mining is an area, where the navigational access behaviour of users' over the web is tracked and analyzed. So that websites owner can easily identify the access patterns of its users'. By collecting and analyzing this behaviour of user activities, websites owner can enhance the quality and performance of services to catch the attention of existing as well as new customers. An approach is proposed to get more accurate prediction results by using both kmean clustering and logistic regression. The predictions of users' future access requests by this manner can improve accuracy of results and will helps in order to reduce the searchtime.

Keywords: web mining, K-mean clustering, Logiistic regression, python,web usage mining.

I. INTRODUCTION

Information dominates the world more than any time before. As the volume of data increases, it becomes a very tedious and tough task to comprehend it. Information technology has made it possible to analyze as well as manage large amount of data electronically and to be able to search for probably very interesting knowledge inside this deep ocean of data. Data mining seems the only solution to this ever growingproblem.

Data mining is primarily a conception of as the process of extracting implicit, previously unknown and potentially useful information from the large set of databases. Exercising large amount of data for superior decision making by looking for interesting patterns in the data has become prime task in today's environment. Hence, the significance of data mining is arising conspicuously. The techniques such as association rules mining, classification, clustering, genetic algorithms and other statistical patterns etc. are often used to discover useful patterns and knowledge that are previously unknown to the system and users. Also data mining has been used in various applications such as marketing, engineering, customer relationship management, decision making, expert prediction, web mining, crime analysis, medicine, and cloud computing, amongothers.

With the technological advancement as well as by the raised popularity of World Wide Web (WWW), many websites typically experience thousands of visitors and its users' daily. So there has been huge interest towards web mining.

The web mining is intended to discover useful patterns from large sets of web data, where at least one of structure or usage data is used in the mining process. Web based data includes different kinds of information like web structure based data, document based data, log based data and data of user profiles. For small sites, an individual Web designer's intuition along with some straightforward usage statistics may be adequate for predicting and verifying the users' browsing behaviour. However, as the size and complexity of a website increases, the simple statistics provided by existing Web log analysis tools are inadequate for providing meaningful insight into how a website is being used.[12]

The web server automatically generates the usage information of the websites and it is stored in the web server as log file. Information of each page requested and accessed by web users' stored in log file commonly referred to as web access log. So web service providers only need a tool to analyze these logs. Web Usage Mining (WUM), a sub part of web mining is used for thispurpose.

II. RELATED WORK

2.1 Clickstream DataPre-processing

Whenever the users' hit search for any webpage every clickstreams they are made stored in the web server log files. The clickstream data is defined as a sequence of Uniform Resource Locators (URLs) browsed by users within a particular time period. It needed to be pre-

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 781-784 www.ijaem.net ISSN: 2395-5252

processed to remove irrelevant entries from it before it is taken foranalyze.

2.2 User Access Matrix

To recognize 'how many pages accessed by particular user' and 'how many users accessed particular page', there is a need to generate web access matrix of users. The retrieved data pattern from previous stage is converted into web user access matrix U_{AM} in which rows represent users and columns represent pages of website. It is used to describe relations between web users and pages who accessed by users. The element a_{ij} of U_{AM}representsthefrequencyofthe user$u_i$of uvisitthe page$p_j$ofpduringagiven period oftime.

$$a_{ij} = \begin{cases} hits(u_i, p_j), if\ p_j is\ visited\ by\ u_i \\ 0, otherwise\ take\ it\ as\ zero \end{cases}$$

Where hits (u_i, p_j) is the count of user u_i accesses the page p_j during a given period of time

2.3 K mean clustering

The k-means algorithm captures the insight that each point in a cluster should be near to the center of that cluster.

In this approach, the data objects ('n') are classified into 'k' number of clusters in which each observation belongs to the cluster with nearest mean.

It defines 'k' sets (the point may be considered as the center of a one or two dimensional figure), one for each cluster $k \leq n$. The clusters are placed far away from each other.

Then , it organizes the data in appropriate data set and associates to the nearest set. If there is no data pending, first step is complicated to perform, in this case an early grouping is done. It is necessary to re-calculate 'k' new set as barycenters of the clusters from previous step.

After having these 'k' new sets, the same data set points and the nearest new sets are bound together.

Finally, a loop is generated. As a result of this loop, the 'k' sets change their location step by step until no more changes are made.

Finally, this algorithm aims at minimizing an objective function as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where,

$x_i^{(j)}$=data point

c_i=cluster center

n=Number of data points

k=Number of cluster

$||x_i^{(j)} - c_i||^2$ = distance between a data point $x_i^{(j)}$ and cluster centre c_j.

2.4 Logistic regression:

Regression analysis is a statistical process for estimating the relationship between two variables.

One of these variable is called predictor variable whose values is gathered through experiments. The other variables is called target variable whose value is derived from the predictor variable. One target variable multiple predictor variable.

This article discusses the basics of Logistic Regression and its implementation in Python. Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.

Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as "1".

III. PROPOSEDWORK

3.1 Overview of proposed work :

In proposed scheme first-of-all the web log data are collected from the web server which contains hits made by the users to access the WebPages. As the log file is in the unstructured format we have to clean it from unwanted entities and convert it into useful format and for that pre-processing of log file will be done. After that to recognize how many pages accessed by particular user and how many users accessed particular page, there is a need to generate web access matrix of users from the click-stream data, and coherent clusters will be generated by bi –clustering

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 781-784 www.ijaem.net ISSN: 2395-5252

technique. The greedy search technique is opted to refine the clusters and resulting patterns generated from these integrated approach are used to predict future page access of users'.

3.2 Flow of proposed system:

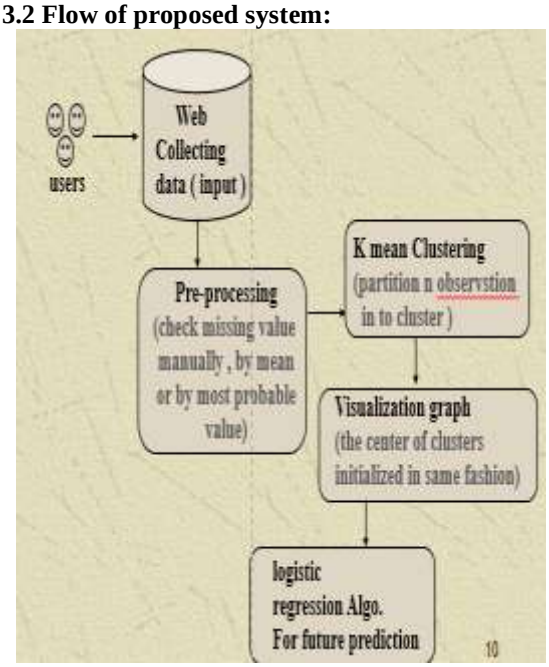

Fig. Flow of proposed system

3.3 Proposed Algorithm :

The following steps are performed in the proposed system:

Input: Web collected data file.

Step 1: Collect data or convert data in required format.(csv file)

Step 2: Apply data pre-processing/cleaning on collected data.

- check missing value.

Step 3: apply k mean clustering algorithm to processed data.

-partition n observation into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers)

Step 4: visualization of cluster graph.

Step 5: Apply Multiple linear regression algorithm for predictive analysis.

- $y = mx + c$

y= dependent or target variable

x= independent or predictor variable

m= slope

c=intercept

Step 6: Make future request predictions for user

IV. CONCLUSIONS AND FUTURE WORK

As the information of page of web site is huge and develops rapidly, increasing the user's browsing speed efficiently as well as possible and reducing the loading of web server become very important issues. As per this thesis, we will try to generate an integrated approach to recognize the frequent access patterns by analysing past users access behavior and based on those retrieve patterns the browsing behaviour of the user will be analyzed which is useful to predict the next page access requests from the user. The proposed approach will be used to improve the accuracy of predictions for users' future requests to better the web performance.

In future work, work can be extended by applying it on different kinds of websites to assess its performance. Work can also be extended by implementing the proposed approach on the parallel as-well as on the cloud technology to evaluate its effectiveness.

REFERENCES

[1]. Sowmya H.K., Dr. R.J. Anandhi, "Web Usage Mining Algorithms: A Survey", AICAAM, April 2019.

[2]. PanjawaniHeena, PoojaJardosh, "WebPage Recommendation in web usage mining using Genetic Algorithm", IJARIIE-ISSN, 2017.

[3]. Pooja Solanki, Jasmin Jha, "Web Page Recommendation System using Biclustering with Greedy Search and Genetic Algorithm", June 2015.

[4]. KaushalKishor Sharma, Prof. Kiran Agrawal, "A Hybrid Approach for Predicting User"s Future Request", IEEE, 2014.

[5]. Dilpreet Kaur, A.P. Sukhpreet Kaur, "User Future Request Prediction Using KFCM in Web Usage Mining", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 2, Issue 8, August 2013.

[6]. A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications (IJCA), Volume 8- No. 11, October 2010.

[7]. PriyankaMakkar, Payal Gulati, Dr. A.K. Sharma, "A Novel Approach for Predicting User Behavior for Improving Web Performance", International Journal on Computer Science and Engineering (IJCSE), Vol. 2, No. 04, 2010.

[8]. V. SUJATHA, PUNITHAVALLI, "Improved User Navigation Pattern Prediction Technique from Web Log Data", Procedia Engineering 30, Elsevier, 92-99, 2012.
[9]. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Volume 1, Issue 2, 1-12, 2000
[10]. Robert -Walker Cooley, "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", May 2000.
[11]. Yuhefizar, Budi Santosa, I Ketut Eddy P., Y. K. Suprapto, "Two level clustering approach for data quality improvement in web usage mining", Journal of Theoretical and Applied Information Technology (JATIT), Vol. 62, No. 2, 404-409, April-2014.
[12]. Raymond kosala, Hendrik Blockeel, "Web mining Research: A Survey", ACMSIGKDD, Volume 2. Issue 1. 1 – 15. July 2000.