Analysis of Web Log Clustering based on Customer's Behavior

1st Tin Nilar Win
Faculty of Information Technology Supporting and Maintenance,
University of Computer Studies (Thaton)
Myanmar
tinnilarwin8887@gmail.com

2nd Nang Khine Zar Lwin
Department of Computer Science
University of Technology, Yadanarbon Cyber City (UTYCC)
Myanmar
nangkhinezar@gmail.com

***Abstract*— The vast volumes of widely dispersed, connected, rich, and dynamic information on the World Wide Web is overwhelming. Data mining techniques are used to find and extract important usage patterns from web data in order to understand and satisfy the needs of web-based applications. Web clustering is a research method being used to better understand customer's behavior in the area of web personalized service. One of the most crucial challenges in the web is the proper services provided to consumers in accordance with respect to their requirements. For the website's creators or administrator, providing services tailored to the needs of specific users takes time and is a responsibility. In this research paper, the clustering of K-Means algorithms is utilized to analyze the webpages size for the customer's access behavior after data from web log files have been preprocessed. We can also get other statistics for each cluster (min, max, std, and so on), then visualize those centers, and see how many points we have in that cluster. In this case, we are still exploring, so this is basically allowing us to identify webpages by size with a small, medium, large, etc., without us giving those values explicitly.**

Keywords*— *Web usage mining, Web log files, K-means Clustering, Web Mining, Pre-Processing

I. Introduction

The web mining system emphasizes how extensively data mining techniques are used to gather extensive data from the internet. Web mining is the study of methods for extracting informational data from online records. Web usage mining is a method that employs web mining techniques to discover and examine clickstream usage patterns and associated data that are generated as a result of user interactions with one or more websites in real-time. The objective is to record, model, and evaluate the behavioral patterns and user profiles that are displayed when people interact with a website [1]. The patterns are often shown as collections of objects, websites, or resources that are commonly used or accessed by user groups with comparable requirements or interests. Web content, web structure, and web usage mining are the three subcategories of web mining [3].

The four primary steps of web mining technologies, are the same as they are for every information mining challenge. First, data collection; second, preprocessing phase; third, pattern discovery phase; and fourth, pattern analysis. The initial phases of pre-processing include data cleaning, user identification, session definition, pageviews identification, and data integration. The initial contribution of our research paper focuses on the quantity and quality of the data used in WUM, two outstanding problems in WUM preprocessing. In order to arrange the raw data and remove unnecessary information, specific preprocessing procedures must be developed because the amount of Web usage data that can be analyzed can be 4GB of web log files. This system provides a detailed WUM preparation process that enables the analyst to transform any collection of Web server log files into an organized collection of Web requests. The primary goal of this system is to preprocess web usage mining, clustering of webpage's size and web user's or customers server log files in order to recognize and comprehend the user access pattern. In the stage of pattern discovery, operations are performed on statistical, database, and machine learning data to find hidden patterns that reflect typical user behavior as well as summary statistics on web sites, sessions, and users. In the third stage, the patterns and data that have been found are further processed and filtered, possibly leading to aggregate user models that may be fed into tools for report generation, visualization, and recommendation.

II. Background Theory

In web usage mining, click stream patterns are automatically found and analyzed. Pre-processing is done on the web access Log data, which prepares the data for clustering by cleaning up the data (removing extraneous data), identifying the user, identifying the session and pageviews and integration of data. In this article, K-Means clustering is used to assess users' access behavior. A data mining technique called clustering brings together a group of items with similar characteristics. There are two different categories of intriguing clusters that can be found in the usage domain: user clusters and page clusters [5]. Users who display similar browsing tendencies are often grouped together through user clustering. In this research paper, we analyzed the webpage's size cluster by using k-means. Further analysis of user groups based on their demographic features (such as age, gender, economic level, etc.) may lead to the identification of essential business intelligence.

A. K-Means Clustering

The clustering method known as K-means can be used as an unsupervised learning strategy when you have unlabeled data [1]. Algorithms that are unsupervised usually rely on the assumption that the output has been labeled without considering

the input vectors. The main intention of the K-means approach, where K is the number of groups, is to discover groups in data. It's attempts to find the centroid by figuring out the k number of centroids and assigning each data point to the closest cluster while keeping the center points as small as desirable.

III. System Architecture Design

In this system architecture as shown in figure 1, firstly we collect the log file from Kaggle. Data cleansing, user and session identification, pageviews identification, and data integration are all included in the pre-processing stage. After this, the preprocessed log data are saved in the database. After preprocessing, the K-Means clustering technique is used to assess the size of the webpage and display the analysis's results.

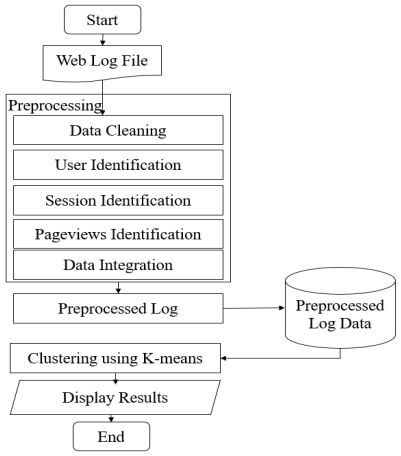

Fig.1. Purposed Model Architecture

A. Data Collection and Pre-processing

Data Collection and Pre-processing are necessary for the effective extraction of useful patterns from the data. The clickstream data is cleaned and separated into a collection of user transactions, which represent the actions that each user took over numerous site visits, during the pre-processing stage. The data set was downloaded from dataverse.harvard.edu/Online Shopping Store - Web Server Logs. Due to the fact that it operates row by row to work, the data size is approximately 4G. In the data preprocessing phase, some of the initial procedures that can be performed to achieve this include cleaning, user and session identification, pageview identification, and data integration.

B. Data Cleaning

The removal of unnecessary data is then done during the data cleaning process so that only valid data can be extracted from the log file to show in figure 2. To improve the quality of analysis, all unnecessary items are removed from the request types and visiting paths, including file name suffixes for the clean files of media extension such as .css,.png, .jpg, .jpeg, .mp3,. js, .cdn, .gif, image,.doc, .html, logo, .ico, .txt, bots, mobile apps, date strings and drop time zone information for easier handling. The values that are less than or larger than 200 are removed from log data by the status codes. Another issue with data cleaning is the HTTP status code.

	client	userid	datetime	method	request	status	size	referer	user_agent
7	37.27.128.139	-	22/Jan /2019:12:38:27 +0330	GET	/browse/Tablet-Arm-Chair/ %D8%B5%D9%80%D8%AF %D9...	200	30604	https://www.zanbil.ir/browse /Classroom-Furnitu...	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.3...
27	85.9.73.119	-	22/Jan /2019:12:38:27 +0330	GET	/site/alexaGooleAnalitic	200	323	https://www.zanbil.ir/browse /compression-stock...	Mozilla/5.0 (Windows NT 6.1; Win64; x64) Apple...
56	37.152.163.59	-	22/Jan /2019:12:38:27 +0330	GET	/site/alexaGooleAnalitic	200	323	https://www.zanbil.ir/product /29314/%DA%A9%D8%...	Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7....
57	5.125.250.100	-	22/Jan /2019:12:38:27 +0330	GET	/m/product/10665/15297 /%D8%B3%D8%B4%D9%88 %D8%A...	200	20552	https://www.zanbil-ir.cdn.ampproject.org/c/s/w...	Mozilla/5.0 (iPhone; CPU iPhone OS 10_3_3 like...
58	178.128.144.232	-	22/Jan /2019:12:38:27 +0330	GET	/order/remainingPayment /187345	200	17614	https://www.zanbil.ir/order /completionAndRemai...	Mozilla/5.0 (iPhone; CPU iPhone OS 10_3_2 like...
...	...	...	...	...	...	...	...	...	...
10364820	87.107.16.82	-	26/Jan /2019:16:01:30 +0330	GET	/browse/home-appliances /%D9%84%D9%88%D8%A7 %D8%...	200	37572	https://www.google.com/	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.3...
10364824	91.99.72.15	-	26/Jan /2019:16:01:30 +0330	GET	/product/32683/63124/%D9%8E %D9%86%DA%A9%D9%87-...	200	41925	-	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_7_3)...

Fig. 2. After Cleaning of Web Log File

C. User Identification

User identifier uses IP addresses to give users with the service they have requested, retrieving each user's access credentials. On the basis of user access behavior patterns, we note that various users can be recognized. A user ID will be assigned to every unique IP address. A user may make multiple visits to a website as part of the user identification procedure. Each user is recorded by the server many times.

D. Session Identification

Every time someone links to the user's page, a session is recorded. It also indicates how frequently a visitor accesses a specific web page. A user's visitation of a certain web page is used to identify their session. In order to identify a distinct user, just one session is formed based on the new IP address. A session is characterized by requests that originate from the same IP address and last for less than 30 minutes. The duration of the longest session is fixed at thirty minutes.

E. Pageviews Identification

From a conceptual standpoint, each pageview can be seen as a group of web resources or objects that each reflect a particular "user event," such as reading an article, visiting a product page, or adding an item to the shopping cart. A user's pageviews during a single visit are collected into a session. Pageviews are things with semantic meaning that mining activities can be done to (such as pages or products). It should be noted that, when it comes to time durations, the amount of time a user spent on the final page they viewed during a session is typically not available. The average time spent on a page during all sessions in which it is not the final one is one often used setting for the last pageview's weight. In some applications, the log of pageview time might be utilized as a weight to reduce data noise. In the figure 3 to show the number of visitors for each pageview.

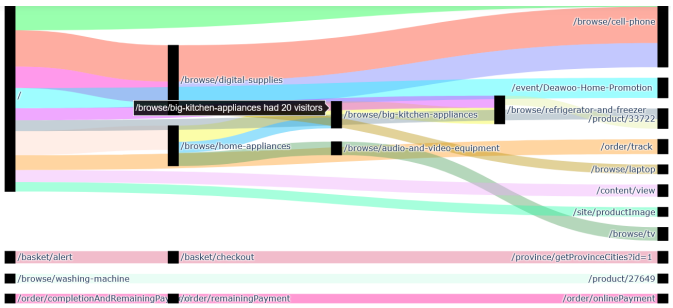

Fig. 3. Number of Visitors by Pageviews

F. Data Integration

The final stage of pre-processing is data integration. The most effective framework for pattern recognition requires the combination of data from a variety of other sources with the pre-processed clickstream data. For various types of data to be successfully integrated, a site-specific "event model," based on which user clickstream segments are gathered and linked to specific events like the adding of a product to the shopping cart, is required. The final transaction database frequently contains the integrated e-commerce data.

IV. WEBPAGE'S SIZE CLUSTERING

After data from web log files have been preprocessed, K-Means clustering algorithms are used to analyze the page size for the customer's access behavior. Additionally, we may obtain additional data for each cluster (such as the minimum and maximum values and standard deviation), show the centers of those clusters, and count the number of points that belong to each cluster. Depending on the application, we can change the cluster count interactively to determine the best number. In this instance, we are still researching, so this essentially allows us to identify pages with a small, medium, large, etc. size, without us explicitly providing those numbers. In the first (low) cluster, we appear to have eight million responses. The average response size for this group is 3,576 bytes, and we can also view the minimum and maximum response sizes as well as other information for this and other clusters. In the figure 4 to show the table of other statistics for five clusters. In the figure 5 to show the result of page distribution by response size for number of five clusters. This diagram shows the x-axis for average page size (bytes) and y-axis display the number of pages in cluster. And then, points are represented the average page size for a cluster of pages. When the user point on each cluster and then to show the result of average page size(bytes), number of pages in cluster, min, max and std as shown in figure 6.

	count	mean	std	min	25%	50%	75%	max
0	8,021,276	3,576	3150.063740	0.000000	1255.000000	2957.000000	4859.000000	15646.000000
1	1,834,990	27,716	9352.256555	15647.000000	18857.000000	27896.000000	36413.000000	50775.000000
2	442,383	73,831	17197.623701	50780.000000	62733.000000	68387.000000	88235.000000	132474.000000
3	55,041	191,115	37639.782577	132489.000000	185207.000000	185207.000000	185207.000000	369714.000000
4	11,175	549,767	61034.302149	374927.000000	538707.500000	549145.000000	567604.000000	1249490.000000

Fig. 4. Table of other Statistics for Five Clusters

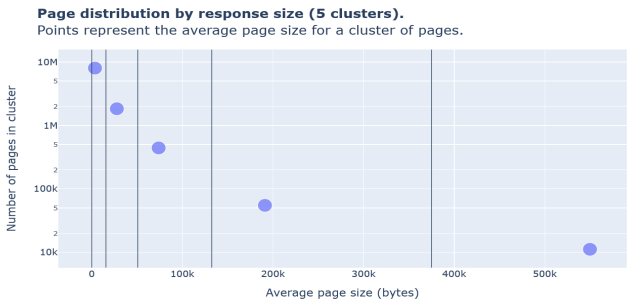

Fig. 5. Page Distribution by Response Size for 5 Clusters

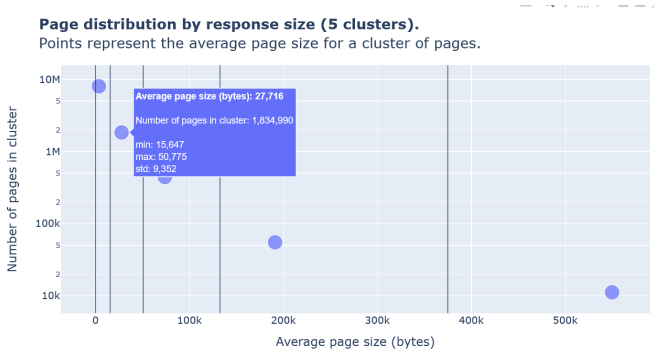

Fig. 6. Display the Result of each Cluster

V. CONCLUSION

There may be a rise in the popularity of online shopping websites where users may buy everything, they desire with a single mouse click. Since it makes it simple to predict browsing and data surfing, web use mining is one of the most significant areas of study. In order to facilitate data availability and accessibility, at the same time, to account for user preferences, clustering of users with similar browsing behaviors is necessary. The advantages of clustering allow us to categorize the size of the webpages based on minimum, maximum, and standard deviation. The findings of our investigation offer insightful information about consumer behavior and content efficiency, assisting online retailers in creating an efficient and successful e-commerce system.

VI. ACKNOWLEDGMENT

First and foremost, the author wishes to express her deepest appreciation to Dr. Aung San Linn, Rector of University of Technology (Yatanarpon Cyber City), for his important guidance and administration. The author wishes to express her gratitude to her respectable supervisor Dr. Nang Khine Zar Lwin of the Faculty of Information and Communication Technology at the University of Technology (Yatanarpon Cyber City), who has worked closely with her at every stage of her research work and provided her with helpful suggestions, in-line supervision, and editing of this research.

REFERENCES

[1] TT.Shwe, Analysis of Web User Clustering based on Users' Access Behavior, University of Computer Studies, Mandalay.

[2] H. Rokham and H.Falakshahi, Web User Clustering Analysis, (IJCSIS) International Journal of Computer Science and Information Security,Vol. 12, No. 9, September 2014.

[3] Dr. RK.Shukla , N.Samaiya , M.Kherajani and P.Sharma, WEB USAGE MINING-A Study of Web data pattern detecting methodologies and its applications in Data Mining, 2nd International Conference on Data, Engineering and Applications (IDEA) ,September 27,2020.

[4] H.Xiang, Research on Clustering Algorithm Based on Web Log Mining, Journal of Physics: Conference Series, ISEITCE 2020.

[5] S.Padmaja, Dr.A Sheshasaayee ,Clustering of user behaviour based on web log data using improved K-means clustering algorithm International Journal of Engineering and Technology (IJET), February 2016.

[6] ZA. Ansari, Web User Session Cluster Discovery Based on k-Means and k-Medoids Techniques, International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 5 No. 12 Dec 2014.