Open Access and Peer Reviewed Journal ISSN 2395-1303 https://www.ijetjournal.org/

RESEARCH ARTICLE OPEN ACCESS

WEB SERVER CLIENT COMMUNICATION FOR WEB USAGE DATA ANALYSIS USING K-MEANS CLUSTERING

V. PRASANNA[1], PROF. G .SREENIVASULU[2]

[1]PG Student, Department of ECE, Sri Venkateswara University College of Engineering, Tirupati

[2]Professor, Department of ECE, Sri Venkateswara University College of Engineering, Tirupati

Abstract

Information from web server logs can be useful to know about user interaction such as page requests, how often users visit a website, and how long users stay at a website. The actual usage of Web pages, though, can include redundant, incomplete or otherwise irrelevant records that do not lend themselves to direct analysis. This paper introduces a web usage data analysis model which uses the web server-client interaction, data preprocessing, feature extraction and K-Means clustering to find different user behaviour patterns. The suggested workflow involves web usage log collection and preprocessing, followed by the extraction of the behavioral features of web usage (e.g., number of visited pages and web usage time). The experimental analysis was done using a dataset of 15 users, clustered with the K-Means algorithm using 3 clusters. The data was processed, clustered, and visualized using Python, Pandas, NumPy, Matplotlib, and Scikit-learn. These clusters correspond to relatively low, moderate and high activity with the centres of the clusters being 8.4 page visits and 3.8 min, 17.8 page visits and 8.8 min, and 28.4 page visits and 14.0 min, respectively. The outcomes show the feasibility of using K-Means to cluster web users based on their simple usage properties and serve as a starting point for the next step of mining the web usage and analysis of the user behavior.

Keywords: *Web Usage Mining, Web Server Logs, Client-Server Communication, K-Means Clustering, User Behaviour, Data Preprocessing, Web Analytics.*

1 Introduction

With the explosion of Web-based applications, there has been a lot of data generated that tracks user interactions with web servers. Web server logs are usually filled with information related to the requests like access time, resources requested, response information, and other attributes that are associated with a request. Information gained from the analysis of these records can be useful for understanding how users interact with a website, and can help improve the design, usability, and service delivery of a website (Cooley et al., 1999).

Web usage mining is a form of data mining that uses data related to the use of Web resources to find meaningful patterns in the use. The conventional web usage mining process has three phases: data preprocessing, pattern discovery, and pattern analysis (Srivastava et al., 2000). One of these stages is especially crucial as

the raw data collected from the server logs can consist of redundant or irrelevant requests and can be used to identify individual users and their browsing sessions, from which meaningful patterns can be extracted (Cooley et al., 1999).

Clustering is an unsupervised technique for grouping users or sessions based on the similarities in their observed behaviour. As an example, K-Means can be used to analyse behavioural features like the number of visits to a page on a website, or the amount of time spent on a website. The algorithm does not need to be set up by the user into predefined user categories but rather clusters observations based on the similarity of their features. This characteristic is ideal for exploratory analysis of web usage data, and this is why clustering is used for this purpose.

In the present study a simplified framework for analyzing web usage behavior is proposed in the following stages: Preprocessing, Feature representation, K-means Clustering and Visualization. The experimental implementation is based on the main behavioural features obtained from the page visit and time spent and a dataset with 15 users. It aims to group users that have relatively similar web usage patterns and show that clustering can be employed as a first stage analysis for web usage data.

2 Literature Review

A lot of research has been conducted on web usage mining to identify useful patterns from data that is collected from user actions on websites. Cooley et al. (1999) indicated the significance of data preparation in Web usage mining and pointed out that Web browsing log data needs to be transformed to a proper form for the purpose of pattern discovery. Because raw web logs might contain information that is not immediately amenable to behavioural analysis, their work made Web usage analysis an important aspect of pre-processing.

In 2000 Srivastava et al. proposed a detailed model for Web usage mining and outlined the various phases of the mining process such as data preprocessing, pattern discovery and pattern analysis. Their research resulted in the realization that web usage mining is able to uncover user navigation and browsing patterns based on web-related information. In this framework, the pre-processing and clustering workflows that are considered in the present study are based on the conceptual foundations.

Facca and Lanzi (2005) have studied the various techniques that can be used to extract useful knowledge from log files. They were interested in how data mining

methods can be used in web usage data, and how well it is important to represent user behaviour. The study suggests that clustering is one of the methods that can be applied for finding the groups of users or sessions with similar characteristics.

Another important topic in web usage analysis is the identification of sessions. Spiliopoulou et al. (2003) analyzed various user session reconstruction heuristics and suggested a framework to assess the effectiveness of these heuristics. Their work shows that the interpretation of web-log records is to a large extent reliant on the way in which individual user activities are clustered into meaningful sessions.

Mobasher et al. (2000) explored the concept of automatic personalisation via web usage mining and showed that user activity patterns can be employed to define user groups and provide personalised service. This is an example of the use of discovered usage patterns for purposes other than description.

K-Means is one of the popular and widely used unsupervised learning algorithms for the clustering stage. Ahmed et al. (2020) gave a detailed study on K-Means algorithm, its simple process, applications, benefits and disadvantages. The algorithm works by assigning observations to clusters based on their distance from the centroids of the clusters, and then computing the new centroids, and repeating the process until a convergence point is reached.

Among the data mining algorithms mentioned by Wu et al. (2008) are K-Means, which they reported on, and they talked about the wide applicability of the algorithm to clustering problems. This algorithm is easy to implement and it can be used to cluster observations based on their feature similarity, allowing for exploratory analysis of user behavior.

From literature reviewed, it can be seen that, for web usage analysis, the proper representation of features and the proper preprocessing of the data is a necessary step before clustering or other web pattern discovery methods are applied. The present study shares this general approach but makes a simplifying experimental analysis by employing Page Visits and Time Spent as behavioural features. The current implementation is smaller scale, experimental dataset with 15 users, and is intended to showcase how the K-Means algorithm can be used to determine website activity levels within groups.

3 Methodology

The proposed methodology includes data preparation, feature selection, K-Means

clustering and visualization. Usually, before the application of any techniques for discovering patterns, the data on the use of a Web site is preprocessed in a way that is called Web usage mining (Srivastava et al., 2000).

Overall workflow starts from web client-server interaction, and user requests generate usage information, which is recorded in web server logs. These records can then be analyzed to provide valuable information on the use of a website and user behaviour (Cooley et al., 1999).

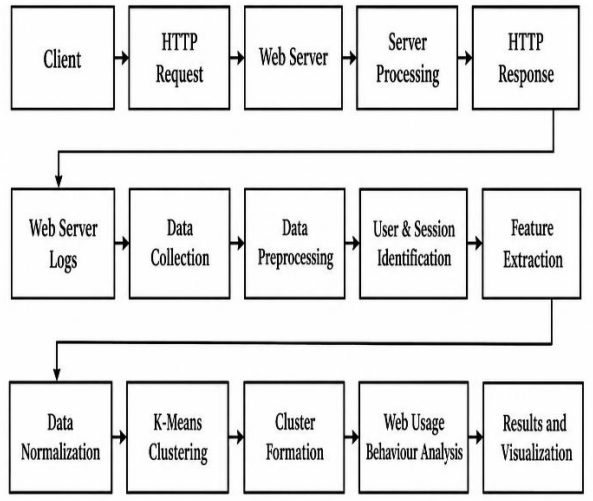

Figure 1 Block diagram of the Proposed methodology

The collected data is preprocessed to get a proper dataset that can be used for clustering. Common web usage preprocessing is to eliminate irrelevant or redundant records and to prepare web usage data for further analysis (Cooley et al., 1999). In the present experimental implementation the prepared set of data is composed of 15 users having, as the main behavioural features, Page Visits and Time Spent. The data for implementation is listed in Table 1.

Table 1 Experimental web usage dataset

User ID	Page Visits	Time Spent (min)
1	10	5
2	15	7
3	8	3
4	20	10
5	25	12
6	5	2
7	30	15
8	18	9
9	12	6
10	28	14
11	7	3
12	22	11
13	27	13
14	14	7
15	32	16

Source: Experimental dataset used in the present implementation.

The two chosen variables are simple representations of users' activity. Page Visits are the number of pages visited by a user, while Time Spent is the time spent on the website. These variables were chosen as the input features used for clustering.

Then the feature matrix is fed to the K Means algorithm. The K-Means method segments the observations into K clusters, where the number of clusters K is a fixed value, by using the following steps: iteratively assign observations to the cluster centroids; iteratively move the

cluster centroids to the mean of observations assigned to them. It is widely applied in unsupervised grouping and has been tested with data mining applications quite thoroughly (Ahmed et al., 2020).

In this study, K = 3 is set. Implementation: The algorithm used is K-Means from the Scikit-learn library, with random_state = 42, and n_init = 10. The resulting cluster assignments are then used to get the mean Page Visits and Time Spent for each cluster.

The obtained groups are then summarized based on the centroid values of these groups. The three cluster centres obtained from the implementation are 8.4 page visits and 3.8 min, 28.4 page visits and 14.0 min, and 17.8 page visits and 8.8 min, respectively.

Table 2 Obtained cluster characteristics

Cluster	Mean Page Visits	Mean Time Spent (min)	No. of Users
0	8.4	3.8	5
1	28.4	14.0	5
2	17.8	8.8	5
Total	—	—	**15**

Source: K-Means clustering output from the implemented model.

The last step is the visualization and interpretation of the clustering results. The cluster assignments, the locations of the centroids, and where the users are located give a direct view of the differences in activity on websites. The entire experimental process is thus described here as an example of transforming a prepared web usage dataset into behavioural groups using the K-Means clustering algorithm.

4 Results and Discussion

Using the K-Means algorithm, the 15 users were divided into three clusters according to the two features of Page Visits and Time Spent. The resulting assignments take a distinct division that can be correlated by the general level of activity on the website. The model implemented gave five users for each cluster and the cluster centres indicated low, moderate and high activity level.

Table 3 User-wise K-Means cluster assignments

User ID	Page Visits	Time Spent (min)	Cluster
1	10	5	0
2	15	7	2
3	8	3	0
4	20	10	2
5	25	12	1
6	5	2	0
7	30	15	1
8	18	9	2
9	12	6	0
10	28	14	1
11	7	3	0
12	22	11	2
13	27	13	1

14	14	7	2
15	32	16	1

Source: K-Means output generated using the experimental dataset.

Cluster 0 contains Users 1, 3, 6, 9, and 11. The centroid of it is 8.4 Page Visits and 3.8 min, suggesting comparatively lower activity of the website. Users 5, 7, 10, 13 and 15 are grouped together and form Cluster 1, whose centroid is 28.4 Page Visits and 14.0 min, which falls in the middle activity level. Within cluster 2, the Users 2, 4, 8, 12 and 14 show an intermediate activity level, with a centroid of 17.8 Page Visits and 8.8 min.

.

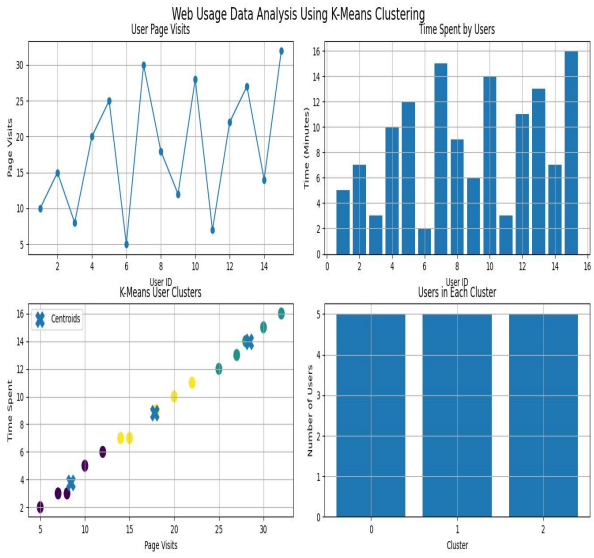

Figure 2 Web Usage Data Analysis Using K-Means Clustering

The correlation between the two features (users who visited more pages and the amount of time spent) is consistent in the clustering result. This relationship helps the K-Means algorithm to create different groups that have different activity characteristics. The second diagram, a graphical representation, further shows the spatial separation of the

Table 4 Summary of identified user behaviour groups

Cluster	Behaviour represented by centroid	Mean Page Visits	Mean Time Spent (min)
0	Lower activity	8.4	3.8
1	Higher activity	28.4	14.0
2	Intermediate activity	17.8	8.8

Source: Calculated from the K-Means cluster centres.

The following characteristics are associated with this specific experimental data set, and are not generalizable to the distribution of web users: 5 users per cluster. It is evident from the results that K-Means is able to cluster the supplied observations based on similarity in the features in behaviour selected. The implementation thus offers a simple framework of analysis that can be expanded to larger and more representative web usage data sets.

Clustering could aid the identification of user clusters, which have comparable browsing behavior, from a web usage mining point of view. Clustering and other data-mining methods have been shown to

be useful in finding patterns in web usage data in previous studies (Srivastava et al., 2000; Facca and Lanzi, 2005). In the present implementation, this concept is demonstrated with two features that are directly interpretable, and not a large number of behavioural variables.

The present experimental analysis has some limitations as only two behavioural features have been used and a smaller number of data points. For this reason the clusters identified need to be understood as an experimental glimpse rather than a representative characterization of real users of the websites. Improved representation of user behaviour would be obtained from larger datasets with extra characteristics like session length, frequency of requests, resources accessed and temporal properties. Research on Web usage mining has pointed out the need for proper pre-processing and feature construction for making meaningful patterns from data gathered from server logs (Cooley et al., 1999; Spiliopoulou et al., 2003).

The findings revealed that overall, the adopted K-Means approach can classify the experimental users into various groups based on the Page Visits and Time Spent. The cluster centres offer a brief quantitative description of these clusters, and show that the technique of unsupervised clustering can be used as a preliminary method in web usage behaviour analysis.

5 Conclusion

In this study, a web usage data analysis framework was proposed, which combined the steps of preprocessing, behavioural features extraction, K-Means clustering and visualization. Web usage mining offers a systematic way to analyze web-server traces and extract useful information from them, and two key steps in the analysis process are pre-processing and the discovery of patterns (Srivastava et al., 2000; Facca and Lanzi, 2005).

We analysed 15 users in the experimental implementation with the behavioural features of Page Visits and Time Spent. The results of the K-Means algorithm (3 clusters) were three groups of points with centre coordinates of (8.4 page visits, 3.8 min), (17.8 page visits, 8.8 min) and (28.4 page visits, 14.0 min). The outcomes of these results show that users in the experimental data set can be distinguished by their actual level of use of the website.

The principle results of the study are:

- K-Means can be used to classify web users into clusters based on web usage attributes.

- The Page Visits and Time Spent metrics are simple representations of the user's activity and are used for exploratory clustering.
- The resulting centroids of the clusters give a quantitative description of the user groups identified.
- The visualization of clusters helps in understanding the difference between the two levels of user activities.
- Small scale demonstration of the present experiment and the results should not be extended to other web user groups.
- A more complete web usage analysis could be obtained by using more real-world datasets from server logs in the future and adding more behavioural characteristics.

The study proposes the foundation of a computational method for mapping web usage data into meaningful behavioral segments, which is then developed into a practical one.The study thus sets up a basic computational method for mapping web usage data into meaningful behavioral groups which is then developed into a practical one. Additional features in the framework may include more complex log processing, user/session attributes, larger data sets, and comparing and contrasting various methods of log clustering.

References

[1] Cooley, R., Mobasher, B. and Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1), 5–32.

[2] Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.-N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2), 12–23.

[3] Facca, F.M. and Lanzi, P.L. (2005). Mining interesting knowledge from weblogs: A survey. *Data & Knowledge Engineering*, 53(3), 225–241.

[4] Spiliopoulou, M., Mobasher, B., Berendt, B. and Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, 15(2), 171–190.

[5] Mobasher, B., Cooley, R. and Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8), 142–151.

[6] Ahmed, M., Seraj, R. and Islam, S.M.S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8), 1295.

[7] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach,

M., Hand, D.J. and Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1–37.

[8] Tan, P.-N., Steinbach, M., Karpatne, A. and Kumar, V. (2018). *Introduction to Data Mining*. 2nd ed. Pearson.