2019 First International Conference of Intelligent Computing and Engineering (ICOICE)

Clustering Web Users Based on K-means Algorithm for Reducing Time Access Cost

Maged Nasser
School of Computing
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
maged.m.nasser@gmail.com

Hentabli Hamza
School of Computing
Universiti Teknologi Malaysia
Johor Bahru,Malaysia
hentabli_hamza@yahoo.fr

Naomie Salim
School of computing
Universiti Teknologi Malaysia
Johor Bahru,Malaysia
naomie@utm.my

Faisal Saeed
College of Computer Science and Engineering
Taibah University, Medina, Saudi Arabia
Alsamet.faisal@gmail.com

***Abstract*— Numerous organizations are providing web-based services due to the consistent increase in web development and number of available web searching tools. However, the advancements in web-based services are associated with increasing difficulties in information retrieval. Efforts are now toward reducing the Internet traffic load and the cost of user access to important information. Web clustering as an important web usage mining (WUM) task groups web users based on their browsing patterns to ensure the provision of a useful knowledge of personalized web services. Based on the web structure, each Uniform Resource Locator (URL) in the web log data is parsed into tokens which are uniquely identified for URLs classification. The collective sequence of URLs a user navigated over a period of 30 minutes is considered as a session and the session is a representation of the users' navigation pattern. In this paper, K-Means algorithm was used to cluster web users based on their similarity in a vector matrix and K-means algorithm implemented several times when k=2,3,4 till k=8 and the results showed the best similarity was when k=8 and the Residual Sum of Squares (RSS) evaluation measure achieved a high intra-cluster similarity value (3.049) when k=8 .**

Keywords*— *Web User Clustering, web usage mining (WUM), Uniform Resource Locator (URL), K-Means, Vector Matrix, Similarity, Residual Sum of Squares (RSS).

I. INTRODUCTION

The Internet has become a major means of life, work, study, and information dissemination. Numerous organizations are providing web-based services due to the consistent increase in web development and the number of available web searching tools. However, information management is becoming troublesome due to the continuous growth in the use and size of the Internet. Hence, there is a need to develop new techniques to improve web performance [1].

Web mining refers to the intelligent analysis of web data; it helps organizations to have a better knowledge of the choices of the web users and help them to run their requirements more efficiently [2]. The clustering of web users based on their similarities is one of the web mining techniques. The web designer can get a better knowledge of the user preferences by analyzing the characteristics of each cluster. This analysis will help in the provision of more suitable and customized services [3].

Virtually all the web clustering methods are based on the similarity of users interests and access patterns; they cluster based on the outcome of these measures. Mining the history of a users' access patterns do not only provide information on the web usage, it also provides some behavioral traits of web users [4]. The need to understand web-users has gained more interest in recent times due to the recent web advancements and the proliferation in the number of web-based applications. Web users can be clustered based on different criteria and useful knowledge can be derived from their access pattern [5]. The knowledge gained can also help in the management of many applications [6]. One of such applications is the prefetching of web pages to assist in personalizing the needs of the user and minimize their waiting time [7]. The other applications may include proxy cache organization [8, 9] and mapping of user access patterns [3]. Few web clustering methods exist [10]; however, their direct application on the primitive user access data is not efficient and fail to establish exciting clusters since web server may often contain several pages which web users may access with different interests [8]. This clustering in this study is focused on the users' navigation pattern. Specifically, a user may visit a website often and spends much time on each visit. The concept of session was introduced to deal with the unpredictable nature of web browsing; it was introduced to serve as the unit of interaction between a web server and a user [11]. The clustering of the user's browsing sessions can help a web developer to understand the browsing pattern of the users and help in the provision of more user-specified services. This knowledge can also contribute to the construction and maintenance of intelligent real-time web servers with dynamic designs to suit future users' needs [12].

Clustering algorithms are used to divide objects into clusters and subsets, with the aim of creating clusters that are internally coherent, but clearly distinct from each other [13]. This implies that objects within a cluster must be as similar as

978-1-7281-4487-0/19/$31.00 ©2019 IEEE

possible and should differ considerably from those in the other clusters. There are several available clustering methods and each of them groups datasets differently. The clustering methods are selected based on the type of output intended, as well as on the known performance of method with the available type of data [13].

The hard clustering algorithms are exhaustive (can assign each object to some cluster) or non-exhaustive (some objects may not be clustered). The hard clustering algorithms can either be flat and hierarchical; the goal of the flat clustering algorithms is to divide the object space into several clusters in that each cluster consists of similar objects and different from the content of the other clusters[14]. Then, the K-mean algorithm is used to compute the similarity between the objects before clustering them. The K-means algorithm was used in this study to compute the similarity between all the web users before clustering them based on their similarity[14].

II.RELATED WORKS

The field of web usage mining (WUM) has recently become an active commercialization and research area. The WUM mainly aims to average the data sourced from users' interactions with the web in order to model patterns that are important for web personalization[15]. Some of the current methods for web usage data mining are statistical analysis, sequential patterns, association rules, classification, and clustering [16-20]. An important WUM topic is web users clustering which involves the discovery of the user clusters with similar information needs, such as users that visits similar web pages. The analysis of the clusters' characteristics can help web developers to understand the user's patterns and be in a better position to provide more customized and suitable services [21].

A comprehensive method in which users' sessions are clustered, evaluated, and interpreted has been presented by Pallis et al. [22]. Xiao et al. also proposed a method of measuring the similarity among the interests of web users based on their past access patterns [23]. Thylashri et al. [24] used K-Mean clustering algorithm for image segmentation for brain tumour detection. Techniques for the improvement of the k-means algorithm by finding fixed centroids and applying a clustering framework to produce similar clusters for each run have been proposed by Chaitraa et al.[25]. Furthermore, Poornalatha et al. suggested the improvements of the K-means algorithm and its application in web sessions clustering [26]. This method addressed the differences in the length of sessions. Duraiswamy et al. proposed the use of matrix for the calculation of sessions' similarity before using agglomerative hierarchical clustering algorithm for the clustering [27].

Previously, we defined the levels of web users' similarities to establish the interests of different web users. However, this definition depends on the application and its function could be based on the number of visits to a page or the number of times a page was visite[28]; it may also depend on visiting the orders of links. Later, two users that visited a page could be clustered into different groups with different interests if they visit the pages in a particular order. A matrix-based framework has been developed for clustering web users in a way that closely related users are clustered together based on their similarity measure [29]. However, an increase in the number of users deteriorated the performance of the clustering method, especially when a threshold number has been reached. Moreover, a page may be visited by the same user several times using either the same or different routes [21]; this makes it difficult to establish the visiting pattern based on similarity. Deep learning has been used for similarity searching in ligand based virtual screening by implemented one of deep learning technique which called deep belief networks with stack of Restricted Boltzmann Machines for reweighting features by giving height weight to the important features that have less error rete to enhance the similarity searching between active and non-active molecules [30].

III.EXPERIMENT STEPS AND RESULTS.

An important step prior to experimentation in web mining is data preprocessing. Many steps have been done for this work during data preprocessing and data clustering as it shown in the Methodology process in Fig. 1, The aim of this step is to transform the log data into a suitable format to be analyzed depending on the needs of the analysis. During data preprocessing, the aim is to meet the demands of the current mining task and as such, the data preprocessing steps in our analysis consists data cleaning, users' identification, session identification, Pages Time Calculation and Vector matrix representation. The data cleaning step involves the removal of the redundant data from the dataset while the user identification step involves knowing the users that visited a website. The session identification step aims to divide the page accesses of each user into individual sessions and this can be easily achieved through a timeout, while vector matrix implies getting the hits information of each users' access to different pages.

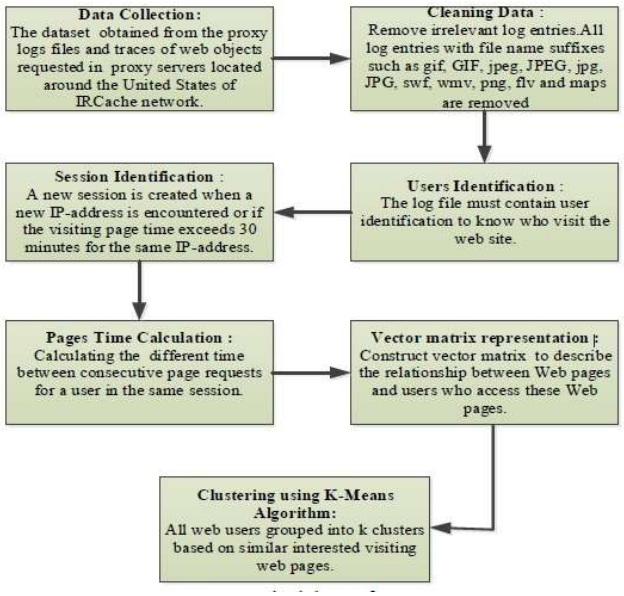


Fig. 1. Methodology of Process.

A. Dataset

The web proxy logs file records information of users' accesses to many web objects in the past. The proxy logs file reflects the behaviors of users and can be considered as a complete and prior knowledge of users' interests. Thus, web proxy logs file can be exploited for predicting future requests. Fig. 2, shows a sample of proxy logs file, which is used for clustering the users based on visited pages.

An access proxy log entry of the proxy logs file is usually composed of 10 fields, including timestamp, client address, elapsed time, request method, URL, size in bytes, user identification, hostname, log tag and HTTP code, hierarchy data, and content type. Figure 1 The data used in this study was sourced from the proxy logs files and web object traces which are requested in BO2 proxy servers hosted around the United States of IRCache network for one day[31] . The meanings of the ten fields for each log entry of the proxy logs file are given in Table 1.

1282348821.049 138 132.55.200.134 TCP_MISS/301 471 GET http://dishnetwork.com/ - DIRECT/205.172.147.51 text/html
1282348917.977 286 249.78.126.183 TCP_MISS/302 1753 GET http://www.hotmail.com/ - DIRECT/64.4.20.184 text/html
1282348954.666 393 132.55.200.134 TCP_MISS/200 22321 GET http://www.msn.com/ - DIRECT/65.55.17.26 text/html
1282348982.398 200 249.78.126.183 TCP_MISS/200 6550 GET http://www.btt.com.ar/foto/t/12/75/1275530489_mark-webb2.jpg - DIRECT/72.232.178.138 image/jpeg
1282348982.437 286 249.78.126.183 TCP_MISS/200 22520 GET http://www.btt.com.ar/foto/t/12/81/1281356302_DSC03379.JPG - DIRECT/72.232.178.138 image/jpeg
1282349600.500 78 50.83.47.141 TCP_MISS/200 344 POST http://app.ninjasaga.com/amf/ - DIRECT/75.126.166.176 application/x-amf
1282361594.519 604 171.11.238.157 TCP_MISS/200 2209 GET http://nt0.ggpht.com/news/tbn/dNDwpcYNGpFYaM/0.jpg - DIRECT/74.125.153.103 image/jpeg
1282361594.561 288 171.11.238.157 TCP_MISS/200 1772 GET http://nt3.ggpht.com/news/tbn/N-iKYy_kXVXPuM/0.jpg - DIRECT/74.125.153.104 image/jpeg
1282411485.94410 60.247.185.54 TCP_MEM_HIT/200 834 GET

Fig. 2. shows the proxy logs file.

TABLE I. EXPLANATION OF THE FIELDS OF LOG ENTRY IN THE PROXY LOGS FILE

Field	Meaning
Timestamp	The time when the client socket is closed. The format is "Unix time" (seconds since Jan 1, 1970) with millisecond resolution.
Elapsed time	The elapsed time of the request, in milliseconds.
Client address	A random IP address identifying the client.
Log tag and HTTP code	The log tag describes how the request was treated locally (hit, miss, etc).But the HTTP status code is the reply code taken from the first line of the HTTP reply header.
Size	The number of bytes written to the client
Request method	The HTTP request method.
URL	The requested URL.
User identification	Always '-' for the IRCache logs.
Hierarchy data and hostname	A description of how and where the requested and Hostname object were fetched.
Content type	The content-type field from the HTTP reply.

B. Data Cleaning

This step involves the application of several filtering techniques to ensure the removal of redundant log entries. Being that both scripts and graphics are downloaded together with Hypertext Markup Language (HTML) file, a user's request to access a page may often result in the generation of several log entries [32]. All log entries with file name suffixes such as gif, GIF, jpeg, JPEG, jpg, JPG, swf, wmv, png, flv, and maps are removed during data cleaning. Data cleaning demands a trace preparation step during which the irrelevant or redundant requests (such as uncatchable and dynamic requests) are cleaned from the log files.

Unnecessary fields should be removed like size, hierarchy data and hostname and user identification because the value of this field always is '-'that does not give any information. The result of reading data set observed the number of log files was 37661 and after cleaned the dataset the number of log files decreased into 3446. Fig. 3, show the pats of these files.

DS | Cleaning DS | Users Session | users matrix | Similarity between session

Timestamp	Elapsed_time	ip	page_name	Content_type	Logtag&HTTPcode
1282360194.091	0.056	249.78.126.183	3963vzpa.css	text/css	TCP_REFRESH_...
1282360230.903	0.26	249.78.126.183	heartbeat.php	text/html	TCP_MISS/200
1282360373.778	0.124	249.78.126.183	ae6gh8wc.js	application/x-javascript	TCP_REFRESH_...
1282360561.26	0.331	249.78.126.183	contenidos1-caja3.css	text/css	TCP_MISS/200
1282360561.813	0.078	249.78.126.183	contenidos4.css	text/css	TCP_MISS/200
1282360561.9	2.559	249.78.126.183	SF_net-%7C-Juegos-...	text/html	TCP_MISS/200
1282360668.092	0.212	249.78.126.183	heartbeat.php	text/html	TCP_MISS/200
1282360831.262	0.183	249.78.126.183	heartbeat.php	text/html	TCP_MISS/200

Fig. 3. Dataset after cleaning

C. User Identification

To reveal the identity of a person that visited a web page, the ID of the person such as the login details must be contained in the log file. However, users are not required to log into some websites, while most web servers do not request for user's login ID when using personal computers. Thus, there is not enough information as per the HTTP standard to distinguish among web users from the same proxy or host. Often, such information is the IP address provided by the Internet Service Provider (ISP) or a corporate proxy server to a user's TCP/IP connection to the site, hence, preventing unique identification [21].

D. Session Identification

User sessions have been used with several studies on recommendation systems for log files to classify and identify the users interesting, but these studies ignored one of the important issue of the works which is the sequential information for each user session [3, 21]. A session refers to the time an activity was initiated to the time it ended. As per W3C, a session refers to the sum of all the activities performed by a user from the time of logging into a site to the time of exit [33]. There is no official login and logout to access and utilize the greater part of the Web destinations. It isn't clear when a session starts and finishes.

Since page demand from different servers are not regularly accessible and a client may visit a site more than once.

Session identification is aimed at dividing each users' page accesses into individual sessions and this can be easily achieved through a timeout. With a timeout, a user is assumed to have started a new session if the time between page requests exceeds a certain limit. A new session is automatically launched if new IP-address is found or if the visiting page time allowed for a particular IP-address has been exceeded. A new session is created when a new IP-address is encountered or if the visiting page time exceeds 30 minutes for the same IP-address. address as shown in Fig.3. Web pages for each user must be reorder based on time visiting before session identification process.

DS | Cleaning DS | Users Session | users matrix | Similarity between session

No_session	ip	timestamp	Name_page	Elapsed_time
3	50.83.47.141	1282350982.984	connect.php	0.264
3	50.83.47.141	1282350989.807	pixel;r=1440723746;fpan=0;fpa=P0-44823...	0.127
3	50.83.47.141	1282350991.862	pixel;r=1123117432;fpan=0;fpa=P0-79448...	0.126
4	50.83.47.141	1282353710.609	5eo1yqin.css	0.039
4	50.83.47.141	1282353710.715	eiim0l2e.css	0.008
4	50.83.47.141	1282353710.833	7mylnyt.css	0.007
4	50.83.47.141	1282353710.948	bw7hgwyd.css	0.007
4	50.83.47.141	1282353711.067	8p5o0tpx.css	0.006
4	50.83.47.141	1282353711.205	7otigqbu.css	0.008
4	50.83.47.141	1282353711.323	6b4rkllkz.css	0.008
4	50.83.47.141	1282353711.443	3963vzpa.css	0.008
4	50.83.47.141	1282353711.573	4wj244ne.js	0.007

Fig. 4. session identification of Web pages

E. Vector matrix for users

Prior to web user clustering based on web logs, a vector matrix for the URL and the user was first constructed and the relationship between the web pages and users that visits these pages was described using a URL-User associated matrix R [34]. Let n and m represent the number of web pages and the number of users respectively, then, the matrix can be represented as:

$$R_{mxn}=\begin{bmatrix} hits(1,1) & hits(1,2) & \cdots & hits(1,j) & \cdots & hits(1,n) \\ hits(2,1) & hits(2,2) & \cdots & hits(2,j) & \cdots & hits(2,n) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ hits(i,1) & hits(i,2) & \cdots & hits(i,j) & \cdots & hits(i,n) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ hits(m,1) & hits(m,2) & \cdots & hits(m,j) & \cdots & hits(m,n) \end{bmatrix}$$

where 'hits' represents a type of user browsing information. The hits of all the users that accessed the web pages over a given time can be directly extracted. From the matrix, users are viewed as the rows, web pages as columns, and the hits count as the values of the elements of this matrix, that is $\text{hits}(i,j)$ as the time spent by user i to access the web page j. The i^{th} row vector $R[\,i,\,]$ records the counts of the i^{th} user access to all the web pages over a specified period, while the j^{th} column vector $R[\,,\,j]$ records the counts of all users who, over the same period, accessed the j^{th} web page. The Fig.5 Show the result of the relationship between the users and the web pages and how much time the user still in the web page.

In Fig. 5, show part of this vector matrix, 3446 web pages was used; number of users = 27, and in the 3rd row and 3rd column indicates that number of users that visited the page 45 over a toke time of 0.007 sec. After getting the hits information vector matrix, we calculated the similarity between users based on visited pages after that K-means algorithm applied to cluster the web users into groups clusters with different k values.

P43	P44	P45	P46	P47	P48
0.006	0.011	0.007	0.006	0.008	0.227
0.006	0.008	0.007	0.008	0.008	0.226
0	0	0	0	0.048	0
0	0	0.023	0	0	0
0	0	0	0	0	0
0.034	0	0.02	0.044	0	0.21
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0.025	0

Fig. 5. Part of Vector matrix

F. Users Similarity

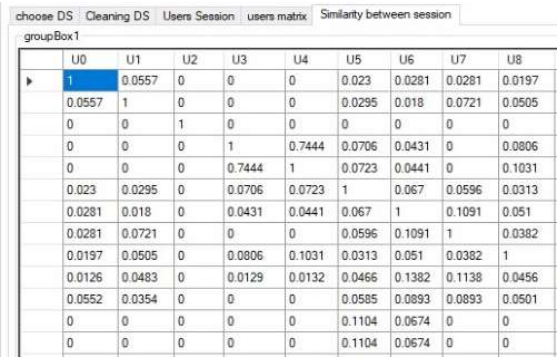
A measure similarity is essential in clustering based on similarity, clustering is based upon grouping samples. The measure should reflect how close or similar two objects. Suppose that, for a given web site, there are n sessions $S = \{s_1, s_2, \dots, s_n\}$ accessing n different web pages $P = \{p_1, p_2, \dots, p_n\}$ in some time interval. For each page p, and each session s, we associate a usage value, denoted as $use(p,s)$ and defined as:

$$use(P_i, S_j) = \begin{cases} 1 \; if \; P_i is \; accessed \; by \; S_j \\ 0 \qquad\qquad otherwise \end{cases} \qquad (1)$$

The $use(*,*)$ vector can be obtained by retrieving the access logs of the site. The users might have similarity and interesting between each other's if they are belong to the same session and accessed the same pages. The similarity can be measured by the number of common pages that are accessed. A precise measure of user's similarity is to consider the actual period each user spent on each page visited. Let $t(P_k, s_j)$ represent the time the user of session s_j spent on page P_k(assume that $t(P_k, s_j) = 0$) if $s_j$does not include page P_k. Here, the similarity between users can be represented as:

$$Cosine\_Sim(s_i, s_j) = \frac{\sum_k (t(P_k,s_i) * t(P_k,s_j))}{\sqrt{\sum_k (t(P_k,s_i))^2 * \sum_k (t(P_k,s_i))^2}} \qquad (2)$$

where $\sum_k (t(P_k, s_i))^2$ represents the square sum of the time the user of session s_i spent while accessing pages at the site, and $\sum_k (t(P_k, s_i) * t(P_k, s_j))$ represents the inner-product over time spent by users of s_i and s_j on visiting the common pages. Even if the same exact pages are visited by two users, their similarity value may be <1 since they spent different times on the page as shown in the Fig.6.



	U0	U1	U2	U3	U4	U5	U6	U7	U8
▶	1	0.0557	0	0	0	0.023	0.0281	0.0281	0.0197
	0.0557	1	0	0	0	0.0295	0.018	0.0721	0.0505
	0	0	1	0	0	0	0	0	0
	0	0	0	1	0.7444	0.0706	0.0431	0	0.0806
	0	0	0	0.7444	1	0.0723	0.0441	0	0.1031
	0.023	0.0295	0	0.0706	0.0723	1	0.067	0.0596	0.0313
	0.0281	0.018	0	0.0431	0.0441	0.067	1	0.1091	0.051
	0.0281	0.0721	0	0	0	0.0596	0.1091	1	0.0382
	0.0197	0.0505	0	0.0806	0.1031	0.0313	0.051	0.0382	1
	0.0126	0.0483	0	0.0129	0.0132	0.0466	0.1382	0.1138	0.0456
	0.0552	0.0354	0	0	0	0.0585	0.0893	0.0893	0.0501
	0	0	0	0	0	0.1104	0.0674	0	0
	0	0	0	0	0	0.1104	0.0674	0	0

Fig. 6. Users Similarity

IV. Results and Evaluation

Recently, The K-means is a popular cluster analysis algorithm [35] which was first introduced in 1967 [36]. With this algorithm, the input dataset is partitioned into k different clusters and each sample is assigned to the cluster that has the closest mean. Each cluster is represented by its sample mean and this mean does not necessarily have to be a sample in the dataset. Similarity measure is used based on cosine similarity when comparing users to the clusters. The K-means is an unsupervised learning framework which can solve most of the established clustering problems[37]. It classifies a given data following a simple and easy procedure to a certain number of clusters (assume k clusters) that has already been fixed. The major idea is to define the k centroids, one per cluster, which should be positioned in a cunning way as different locations can give different outcomes. So, it is better to position them far from each other [37].

The next step involves taking each point that belongs to a given data set and associating it to the nearest centroid. Having associated all the points, the first step is done, and an early group age is equally done. Then, the next thing is to recalculate the k new centroids as the barycenters of the clusters from the previous step. Having determined the k new centroids, a new binding will be done between the same data set points and the nearest new centroid, resulting in the generation of a loop [21].

The algorithmic flow is as follows:

- Place K points into the space represented by the objects to be clustered (these points are the initial group centroids).
- Group the objects based on the closest centroids.
- Having assigned all the objects, recalculate K centroids' positions with based on the objects in the same cluster.
- Repeat Steps 2 and 3 until the centroids are no longer moving.

This procedure explained a easy and simple way to group a given data in to many group clustering based on the similarity and the user interesting by using one of the famous method that had been used for clustering which called K-means algorithm, The main idea is to define k centroids for each cluster. The algorithm is implemented when k=2,3,4,5,6,7,8. All the results are shown in Table 2 and It is observed that all of user divided in to groups based on similarity and k clusters.

In most clustering frameworks, the objective functions try to achieve a high intra-cluster similarity and a low inter-cluster similarity, and this is often considered as an internal clustering quality criterion. One of the measures of internal criterion is the residual sum of squares (RSS), defined as the cosine similarity of each vector from its centroid summed over all vectors [38]. The RSS value is calculated thus:

$$\mathrm{RSS} = \sum_{k=1}^{k} RSS_k \ , RSS_k \sum_{x \in w_k} Cosine\_Sim(\vec{x}, u(w_k)) \quad (3)$$

The aim is to maximize the RSS value as it relates to the maximization of the similarity in the same cluster[38].The RSS clustering results are shown in Table II.

TABLE II. Clustering Results and Evaluation

K-CLUSTER		USERS GROUPING	RSS
2	1	U0,U2,U5,U6,U8,U10,U15,U16,U17,U19,U21,U22,U23,U24,U25,U26	0.872
	2	U1,U3,U4,U7,U9,U11,U12,U13,U14,U18,U20	
3	1	U2,U5,U9,U18,U19,U22	1.114
	2	U0,U1,U3,U4,U10,U15,U16,U17,U21,U26	
	3	U6,U7,U8,U11,U12,U13,U14,U20,U23,U24,U25	
4	1	U2,U5,U19,U22,U24	1.256
	2	U9,U18	
	3	U1,U3,U4,U8,U11,U12,U13,U15,U21,U23,U25,U26	
	4	U0,U6,U7,U10,U14,U16,U17,U20	
5	1	U2,U19,U22,U24	2.25
	2	U0,U6,U7,U9,U10,U14,U18,U20,U26	
	3	U5	
	4	U11,U12,U13,U16,U17	
	5	U1,U3,U4,U8,U15,U21,U23,U25	
6	1	U5,U10,U21,U22,U24	2.564
	2	U19	
	3	U0,U1,U3,U4,U7,U8,U14,U20,U25,U26	
	4	U16	
	5	U2	
	6	U6,U9,U11,U12,U13,U15,U17,U18,U23	
7	1	U2,U7,U10,U15,U18,U20,U21,U25	2.980
	2	U24	
	3	U5	
	4	U22	

	5	U0,U1,U6,U11,U12,U13,U16,U17,U23,U26	
	6	U3,U4,U8,U9,U14	
	7	U19	
8	1	U2,U19,U22	3.049
	2	U25	
	3	U11,U12,U13,U16,U17	
	4	U0,U7,U10,U20,U21	
	5	U5	
	6	U1,U8,U15,U23,U26	
	7	U1,U8,U15,U23,U26	
	8	U3,U4	

As it showed in Table III, we can see that when the k value is 8 the similarity preference of the algorithm was the best, so, the clustering result is shown in Fig.7, based on k = 8.

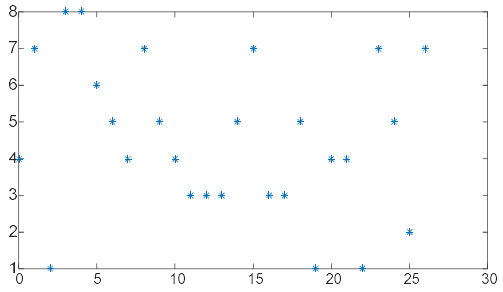


Fig. 7. Users after clustering

The relationship between k and RSS is shown in Fig.8, The RSS was observed to be highest when the value of k = 8, meaning that the users' similarity in the same cluster is high.

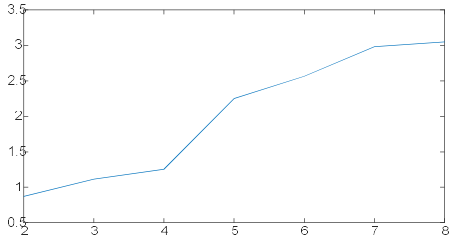


Fig. 8. Relationship between RSS and number of clustering

In most clustering frameworks, the objective functions try to achieve a high intra-cluster similarity and a low inter-cluster similarity, and this is often considered as an internal clustering quality criterion. One of the measures of internal criterion is the residual sum of squares (RSS), defined as the cosine similarity of each vector from its centroid summed over all vectors [21]. The RSS value is calculated thus:

RSS=$\sum_{k=1}^{k} RSS_k$ where$RSS_k \sum_{x\in w_k} Cosine_Sim(\vec{x}, u(w_k))$ The aim is to maximize the RSS value as it relates to the maximization of the similarity in the same cluster. The clustering results are shown in Table III.

TABLE III. CLUSTERING RESULT

K	RSS
2	0.8722
3	1.1147
4	1.256
5	2.25
6	2.564
7	2.9808
8	3.049

V. CONCLUSION

In this paper, we presented the use of the k-means algorithm for the clustering of web users using session-based similarities. The clustering was intended to model the similarities between web users as characterized based on cosine similarity measures. A web user may frequently visit a web page and spend an arbitrary amount of time per visit; users may also access a web page for different reasons. Hence, our web user clustering was based on user sessions rather than the user's entire history. For the data set, the web servers contain 3446 pages and 27 sessions after cleaning and pre-processing. We implemented RSS as a measure of the internal criterion of clustering quality to maximize the similarity in the same clusters. The experiments were conducted, and the results showed that the proposed method can cluster web users with similar interests.

ACKNOWLEDGMENT

This work is supported by the Ministry of Higher Education (MOHE) and the Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under the Research University Grant Category (R.J130000.7828.4F985).

REFERENCES

1 Silverstone, R.: 'Introduction': 'Media, technology and everyday life in Europe' (Routledge, 2017), pp. 19-36

2 Satish Babu, J., Ravi Kumar, T., and Shahana Bano, D.: 'Optimizing webpage relevancy using page ranking and content based ranking', 2018, 2018, 7, (2.7), pp. 5

3 Narayan Jadhav, J., and Arunkumar, B.: 'Web Page Recommendation System Using Laplace Correction Dependent Probability and Chronological Dragonfly-Based Clustering', 2018, 2018, 7, (3.27), pp. 13

4 Catledge, L.D., and Pitkow, J.E.: 'Characterizing browsing strategies in the World-Wide Web', Computer Networks and ISDN systems, 1995, 27, (6), pp. 1065-1073

5 Shahabi, C., Zarkesh, A.M., Adibi, J., and Shah, V.: 'Knowledge discovery from users web-page navigation', in Editor (Ed.)^(Eds.): 'Book Knowledge discovery from users web-page navigation' (IEEE, 1997, edn.), pp. 20-29

6 Yan, T.W., Jacobsen, M., Garcia-Molina, H., and Dayal, U.: 'From user access patterns to dynamic hypertext linking', Computer Networks and ISDN Systems, 1996, 28, (7), pp. 1007-1014

7 Cunha, C.R., and Jaccoud, C.E.: 'Determining www user's next access and its application to pre-fetching', in Editor (Ed.)^(Eds.): 'Book Determining

www user's next access and its application to pre-fetching' (IEEE, 1997, edn.), pp. 6-11

8 Cao, P., and Irani, S.: 'Cost-Aware WWW Proxy Caching Algorithms', in Editor (Ed.)^(Eds.): 'Book Cost-Aware WWW Proxy Caching Algorithms' (1997, edn.), pp. 193-206

9 Cao, P., Zhang, J., and Beach, K.: 'Active cache: Caching dynamic contents on the web', Distributed Systems Engineering, 1999, 6, (1), pp. 43

10 Cooley, R., B. Mobasher, and J. Srivastava, Data preparation for mining world wide web browsing patterns. Knowledge and information systems, 1999. 1(1): p. 5-32.

11 Fu, Y., Sandhu, K., and Shih, M.-Y.: 'Clustering of web users based on access patterns', in Editor (Ed.)^(Eds.): 'Book Clustering of web users based on access patterns' (San Diego, CA. Springer-Verlag, 1999, edn.), pp.

12 Su, Q., and Chen, L.: 'A method for discovering clusters of e-commerce interest patterns using click-stream data', electronic commerce research and applications, 2015, 14, (1), pp. 1-13

13 Yuvaraj, K., and Manjula, D.: 'A performance analysis of clustering based algorithms for the microarray gene expression data', International Journal of Engineering and Technology(UAE), 2018, 7, (2), pp. 201-205

14 Aparajita, A., Swagatika, S., and Singh, D.: 'Comparative analysis of clustering techniques in cloud for effective load balancing', International Journal of Engineering and Technology(UAE), 2018, 7, (3), pp. 47-51

15 Patil, H., and Singh Thakur, R.: 'A semantic approach for text document clustering using frequent itemsets and WordNet', 2018, 2018, 7, (2.9), pp. 4

16 Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-N.: 'Web usage mining: Discovery and applications of usage patterns from web data', ACM SIGKDD Explorations Newsletter, 2000, 1, (2), pp. 12-23

17 Mobasher, B., Dai, H., Luo, T., and Nakagawa, M.: 'Effective personalization based on association rule discovery from web usage data', in Editor (Ed.)^(Eds.): 'Book Effective personalization based on association rule discovery from web usage data' (ACM, 2001, edn.), pp. 9-15

18 Yang, Q., Zhang, H.H., and Li, T.: 'Mining web logs for prediction models in WWW caching and prefetching', in Editor (Ed.)^(Eds.): 'Book Mining web logs for prediction models in WWW caching and prefetching' (ACM, 2001, edn.), pp. 473-478

19 Li, I.T.Y., Yang, Q., and Wang, K.: 'Classification Pruning for Web-request Prediction', in Editor (Ed.)^(Eds.): 'Book Classification Pruning for Web-request Prediction' (2001, edn.), pp.

20 Mobasher, B., Cooley, R., and Srivastava, J.: 'Creating adaptive web sites through usage-based clustering of URLs', in Editor (Ed.)^(Eds.): 'Book Creating adaptive web sites through usage-based clustering of URLs' (IEEE, 1999, edn.), pp. 19-25

21 Nasser, M., Salim, N., and Hentabli Hamza, F.S.: 'Clustering web users for reductions the internet traffic load and users access cost based on K-means algorithm', International Journal of Engineering & Technology, 2018, 7, (4), pp. 3162-3169

22 Pallis, G., Angelis, L., and Vakali, A.: 'Model-based cluster analysis for web users sessions': 'Foundations of Intelligent Systems' (Springer, 2005), pp. 219-227

23 Xiao, J., and Zhang, Y.: 'Clustering of web users using session-based similarity measures', in Editor (Ed.)^(Eds.): 'Book Clustering of web users using session-based similarity measures' (IEEE, 2001, edn.), pp. 223-228

24 Thylashri, S., Mahesh Yadav, U., and Danush Chowdary, T.: 'Image Segmentation Using K- Means Clustering Method for Brain Tumour Detection', 2018, 2018, 7, (2.19), pp. 4

25 Chitraa, V., and Thanamani, A.S.: 'An Enhanced Clustering Technique for Web Usage Mining', International Journal of Engineering Research & Technology (IJERT) Vol, 2012, 1

26 Poornalatha, G., and Raghavendra, P.S.: 'Web user session clustering using modified K-means algorithm': 'Advances in Computing and Communications' (Springer, 2011), pp. 243-252

27 Duraiswamy, K., and Mayil, V.V.: 'Similarity matrix based session clustering by sequence alignment using dynamic programming', Computer and Information Science, 2008, 1, (3), pp. 66

28 Xiao, J., Zhang, Y., Jia, X., and Li, T.: 'Measuring similarity of interests for clustering web-users', in Editor (Ed.)^(Eds.): 'Book Measuring similarity of interests for clustering web-users' (IEEE Computer Society, 2001, edn.), pp. 107-114

29 Sabitha, V., and S.K. Srivatsa, D.: 'An Efficient Modified K-Means and Artificial Bee Colony Algorithm for Mining Search Result from Web Database', International Journal of Engineering & Technology, 2018, 7, (2.20), pp. 5

30 Nasser, M., Salim, N., Hamza, H., and Saeed, F.: 'Deep Belief Network for Molecular Feature Selection in Ligand-Based Virtual Screening', in Editor (Ed.)^(Eds.): 'Book Deep Belief Network for Molecular Feature Selection in Ligand-Based Virtual Screening' (Springer, 2018, edn.), pp. 3-14

31 Romano, S., and ElAarag, H.: 'A neural network proxy cache replacement strategy and its implementation in the Squid proxy server', Neural computing and Applications, 2011, 20, (1), pp. 59-78

32 Cooley, R., Mobasher, B., and Srivastava, J.: 'Data preparation for mining world wide web browsing patterns', Knowledge and information systems, 1999, 1, (1), pp. 5-32

33 Consortium, W.W.W.: 'RDF 1.1 concepts and abstract syntax', 2014

34 Xu, J., and Liu, H.: 'Web user clustering analysis based on KMeans algorithm', in Editor (Ed.)^(Eds.): 'Book Web user clustering analysis based on KMeans algorithm' (IEEE, 2010, edn.), pp. V2-6-V2-9

35 NLANR, M.B.: 'National Laboratory for Applied Network Research', in Editor (Ed.)^(Eds.): 'Book National Laboratory for Applied Network Research' (2006, edn.), pp.

36 Abhari, A., Dandamudi, S.P., and Majumdar, S.: 'Web object-based storage management in proxy caches', Future Generation Computer Systems, 2006, 22, (1-2), pp. 16-31

37 Jain, A.K.: 'Data clustering: 50 years beyond K-means', Pattern recognition letters, 2010, 31, (8), pp. 651-666

38 Singh, V.K., Tiwari, N., and Garg, S.: 'Document clustering using k-means, heuristic k-means and fuzzy c-means', in Editor (Ed.)^(Eds.): 'Book Document clustering using k-means, heuristic k-means and fuzzy c-means' (IEEE, 2011, edn.), pp. 297-301