Elysium Journal of Engineering Research & Management

P-ISSN: 2347-4408
E-ISSN: 2347-4734

Similarity Measurement Of Web Navigation Pattern Using K-Harmonic Mean Algorithm

K.Abirami[1] and Dr. P.Mayilvaganan[2]

[1]Research Scholar,School of Computing Sciences, VELS University,Chennai, India
[2]Professor,Department of MCA, School of Computing Sciences, VELS University,Chennai, India

Abstract*— we present a new method to improve the web Navigation Usage Pattern to discover the web data based on similarity between two cluster points. The web usage patterns can be extracted from Web server logs regularly verified for working websites by first handling the log data to find users, user sessions, and user task-oriented transactions, and then applying a Web usage mining algorithm to determine patterns among web usage paths. In conventional Web usage mining, semantic information of the Web page content does not take part in the pattern generation process. The web navigation usage pattern including information about both the path and time essential for user-oriented tasks. It is taken by our ideal user communicating path models. It can be measure to distance between similar web usage patterns. In this approach, the user visited pages are subdivided into clusters using a non-Euclidean distance measure called the Sequence Order Method (SOM) and Euclidean method measure called Association Distance Measure (ADM). In this paper mainly focus to identify page path similarity, and implementing KHM clustering algorithm. The minimum number of pages in a session and similarity of usage path were calculated.

Keywords*— Web Data Mining; Pattern Discover; Web Log data; Classification of Users, Association Rules, clustering algorithm(KHM).

1.INTRODUCTION

The World Wide Web has developed the biggest and the most popular way of communicating, retrieving and circulating information. The number of Web pages available is increasing very rapidly adding to the hundreds of millions pages already on-line. The rapid and chaotic growth has resulted into more complex structure of Web sites. Web mining are classified in three categories 1) Web Structure Mining 2) Web Content Mining 3) Web Usage Mining.

Web Structure Mining is the task for discovering knowledge from the structure of hyperlinks within Web pages and given useful information for the relationship among Web pages. Web Content Mining is the task of discovering different kinds of information contents and improving efficient mechanisms to organize and grouping (clustering) multimedia content to the search engines for accessing these contents by using keywords, categories, related contents etc. When a web user visits a website, for one request ordered by the user one or more than one record of the server is stored in the web access log [2]. The analysis of such data can be used to understand the user preferences and behavior in a process commonly referred to as Web Usage Mining.

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [5]. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. SOM is compared to a commonly used distance measure within cluster analysis called association distance measure, which does not incorporate structural information[3].

2.PRINCIPLES OF WEB USAGE MINING

Web usage mining Process

Web usage mining is the task of applying data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of users navigating on the Web. As every data mining task, the process of Web usage mining also consists of four main steps: (i) data gathering, (ii) preprocessing, (iii) pattern discovery and (iv) Pattern analysis[2].

- Data gathering or information gathering: This is done mostly by the web servers; however there exist Methods, where client side data are collected as well.

- Preprocessing: (i) Data Cleaning. As in all information discovery processes, in web usage mining can also be happen that such data is recorded in the log file that is not useful for the further process, or even ambiguous or faulty. These records have to be corrected or removed. (ii)User identification. In this step the unique users are distinguished, and as a result, the different users are identified. This can be done in various ways like using IP addresses, cookies, and direct authentication and so on. (iii)Session identification. A session is understood as a sequence of activities performed by a user when he is navigating through a given site. To identify the sessions from the raw data is a complex step, because the server logs do not always contain all the information needed. There are Web server logs that do not contain enough information to reconstruct the user sessions; in this case (for example time-oriented or structure-oriented) heuristics can be used. (iv)Feature selection. In this step only those fields are selected, that are relevant for further processing [7]. (v) Data transformation, where the data is transformed in such a way that the data mining task can use it. For example strings are converted into integers, or date Fields are truncated etc [1].

- Pattern discovery: Executing the data mining task. This can be for example frequent item set mining, sequence mining, graph mining, clustering and so on.
- Pattern Analysis: Result understanding and visualization. Last step involves representing knowledge achieved from web usage mining in an appropriate form

Web Access Log

Each access to a Web page is recorded in the web access log of web server that hosts it. Each entry of web access log file consists of fields that follow a predefined format [2]. The fields of the CLF(common log format) are

remotehost rfc931 authuser date “request” status bytes

In the following a short description is provided for each field:

- *remotehost*. Name of the computer by which a user is connected to a web site. In case the name of computer is not present on DNS server, instead the computer's IP address is recorded.
- *rfc931*. The remote log name of the user.
- *authuser*. The username as witch the user has authenticated himself, available when using password protected WWW pages.
- *date*. The date and time of the request.
- *request*. The request line exactly as it came from the client (the file, the name and the method used to retrieve it).
- *status*. The HTTP status code returned to the client, indicating whether or not the file was successfully retrieved and if not, what error message was returned.
- *Byte*. The content-length of the documents transferred.

3.METHODS USING WEB USAGE MINING

3.1. Sequence Order Method

SOM is a non-Euclidean distance measure reflecting the order of elements. Basically, a non-Euclidean distance measure is a parallel measure that goes outside the Euclidean straight line drawn between two Points. First point denotes the authentic Usage Pattern. Second point denotes the probable Usage Pattern. In general, the distance or similarity between sequences is reflected by the amount of work that has to be done to convert one sequence to another. As a result, SOM distance measure is represented by a score. The higher/lower the score, the more/less effort it takes to equalize the sequences.

In addition, SOM scores for the following operations during the equalization process: *insertion (I), deletion (D) and reordering (R).* The three process used to swap the usage pattern in authentic and probable. In particular, SOM distance measure between two sequences A and P is calculated using the following formula

$$d_1(A, P) = (X_d D + X_i I) + \eta R \qquad (1)$$

d → is the distance between two sequences A(Authentic) and P(Probable)

X_d → is the weight value for the deletion operations, a positive constant not equal to 0, determined by the User ($X_d > 0$);

X_i → is the weight value for the insertion operations, a positive constant not equal to 0, determined by the User ($X_i > 0$);

D → is the number of deletion operations;

I → is the number of insertion operations;

R → is the number of reordering operations;

η → is the reordering weight, a positive constant not equal to 0,determined by the User ($\eta > 0$).

The above formula indicates that the score, represented by SOM distance measure between two sequences A and P, consists of the costs for deleting and inserting unique elements and the costs for reordering common elements. For example when sequences are compared with regard to pages and other information types such as time spent on pages, optimal trajectories of operation sets are found using heuristics[8]. For the purpose of this project, sequences are represented by server sessions and elements stand for visited web pages. A server session or a visit is defined as the click-stream of page views for a single visit of a user to a web site.

To illustrate SOM, consider the following sequences: Suppose $X_d = X_i = 1$ and $\eta = X_d + X_i$:

(A, 1, 3, 25, 37, 48, 50, 123) (P, 1, 25, 37, 3, 48, 50, 3)

Server sessions A and P have six common elements (1, 3, 25, 37, 48 and 50) and one unique element (123). Element 3 needs to be reordered from position 4 to 3 in P or from position 3 to 4 in A. The results of this reordering operation are six identities and two unique elements (123 and 3). Element 3 is now considered as a unique element because identities have been formed between every element of the source and target sequence, except for elements 123 (at position 7 of A) and 3 (at position 7 of P). The deletion (3,−) in sequence 1 and insertion (−, 123) in sequence 3 are necessary to equalize sequence 1 with sequence 3. Applying formula (1). we obtain a distance measure of 4. The Sequence Order method used to reorder the position of web pages that is authentic usage pattern and probable web page. First calculate the distance authentic and probable web page position in web site. And then reordering the probable usage pattern is first position in web page [1].

3.2. Association Distance Measure

A commonly used distance measure between sequences for segmentation studies is association distance measure. The method is Euclidean based and does not take into account the order of elements within sequences [5]. A Euclidean-based distance measure is a similarity measure calculating the length of a straight line drawn between two objects [1]. A simple form of association measure for analyzing data in non-metric terms will transform each sequence into a vector and counts the number of dissimilarities at each position of the sequence. Missing values in either one of the compared sequences are treated as dissimilarity[7]. In particular, the distance between two sequences based on the association measure is presented with the following formula:

$$d_2(A, P) = \sum_{i=1}^{n} f_i \qquad (2)$$

with

fi =1 if A$(i) \neq$ P(i)
fi = 0 otherwise

where:

$d_2 \longrightarrow$ is the distance between two sequences, A and P, based on Association distance measure;

$\sum_{i=1}^{n} f_i \longrightarrow$ is the sum of dissimilarities between sequences A and P from positions i to n;
$n \longrightarrow$ is the number of positions of $S1$ or $S2$ if the sequences are of equal length; otherwise n is equal to the number of positions of the longest sequence.

In this way, the distance between sequences A and P, sequence order method will be based on the association measure.

3.3. Page path similarity Matrix

Website developers usually store pages which are related either in structure and content is same subdirectory, or create links between two related pages. Due to our lack of knowledge about links between pages on web access logs, to realize the developer's opinion on conceptual relation between pages, the website's pages storage path is employed. For example, two pages s1 and s2 which are located in the following paths.

Directory1/Subdir1/subdir2/s1.html
Directory1/Subdir1/subdir2/s2.html

Are more related than two pages which are on the following paths

Directory1/Subdir1/subdir2/s1.html
Directory2/Subdir3/subdir2/s2.html

Hence, a new matrix called ‘page path similarity matrix’ can achieved. To compute path similarity matrix, first the function similarity PS (s1, s2) is defined. This function proceeds the number of common sub-directories of two similarity pages, i.e. s1 and s2. To compute path similarity matrix elements, the following equation is used:

$$Si_j = \frac{2 \times PS}{No.ofdirectory(Path(s1)) + No.ofdirectory(Path(s2))} \qquad (3)$$

Where number of directory(path(si)) is the number of subdirectories of storage path in si. When two paths of two pages are close to each other, the value of each element of this matrix get closer to 1, and if there is no similarity in storage path, it becomes 0. The similarity of two pages authentic page and probable page.

3.4. The K-Harmonic Mean (KHM) Clustering Algorithms

KHM clustering algorithm uses the harmonic means of the distances from data points to the cluster centers in its cost function. The k-harmonic mean algorithm is a method similar from a different objective function. The KHM objective Function uses the harmonic mean of the distance from each data point to all clusters.

$$KHM(X, C) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\left\|d_i - c_j\right\|^p}} \qquad (4)$$

Where di is the i_{th} data point, C_j is the j_{th} cluster center. k is the number of desired clusters. a is a positive number. Here p is a input parameter and typically p ≥ 2. KHM clustering algorithm has following steps for data having n data points and k desired clusters;

1. KHM algorithm starts with random cluster centers.
2. The distances between each data point to all the centers are calculated.
3. The new cluster centers are calculated.

The harmonic mean gives a good score for each data point to any one center. This is a property of the harmonic mean. It is similar to the minimum function used in k-mean, but is a smooth differentiable function [9].

3.5. Co-occurrence Matrix

Combining these two matrixes, the new matrix CA is created which shows relation between dissimilar pages of site based on a mix of users and developers opinions [2]. To combine these two matrixes whose elements of each differs between zero and 1,

$$CA_{ij} = \alpha + C_{ij} + (1-\alpha)\times S_{ij} \qquad \text{□□□}$$

Where C is Co-Occurrence Matrix and S is the Page Similarity Matrix. To arrive at clusters of related pages, the graph corresponding to the achieved matrix is divided into strong partitions. When the value of Cij is higher than the MinFreq, two corresponding nodes are considered connected, and in other case they are taken disconnected. Each node which is visited is labeled with a visited label. If all nodes bear visited labels, the algorithm ends, otherwise the node not visited is selected and DFS algorithm id performed[8].

4.EXPERIMENTAL RESULT

Experiments result in this paper concern the request whether structural information embedded in web-click stream data is well reflected by equation (1) SOM and whether the incorporation of the structural information affects the final cluster solution. Hence, we first calculate pair-wise distances between sequences representing sequentially ordered visited web pages using SOM as a distance measure. SOM is used to retrieve the distance between authenticated and probable pages, and the Association distance measure method also retrieve the authentic and probable page. The above two methods are used to retrieve the position of web navigation usage Pattern.

The web-click stream data are implemented the SOM, the data positions are reordering and match the data in clusters. When the click stream data are grouped in to the web, then the pages were partitioned to same clusters. Once the user request the authentic page were associated. In this time the clusters were associated in distance measure of A and P.

In order to compare SOM with a method that does not incorporate structural information of web pages, pair-wise distances between the same sequences are calculated using association distance measure. This is a commonly used, non-Euclidean distance measure in clustering. Equation (2) is Association-based pair-wise distances are inserted into a second similarity matrix. Finally, we examine which cluster solution better separates between the structural characteristics, represented by the order of visited pages, of navigation patterns [4].

Equation (3) is find the Page path similarity of authentic and probable page. In this path were identify the same path in another directory. Equation (3) calculate the similarity of pages.

Equation (4) the KHM clustering algorithm find the web navigation usage pattern. And then Equation (5) is calculated co-occurrence matrix for clustering web log data.

The following table using data set at Shivaqua paddle wheel aerator web log file. In this web log files are calculated page similarity. The features of the web log file at Shivaqua paddle wheel aerator dataset.

TABLE I. DATASET USED IN EXPERIMENT

Data Set	*Size(MB)*	*No. Of Pages*	*Periods(days)*
Shivaqua Paddel Wheel aerator	1005.64	2,043	30

TABLE II. REMOVED EXTRA ENTRIES

Page Extension	*Hits*
gif	273
.js	2961
.pdf	42
.bmp, .wav, …, web bots entries	9800
Total	13076

After removing unwanted entries, different web users are identified. This step is conducted based on remote host field. After identified distinct web users, users' sessions are reconstructed. As sessions with one page length are free from any useful information, they are removed too. In Table 3, characteristics of web access log file is represented after performing pre-processing phase.

TABLE III. WEB ACCESS LOG FILE AFTER PERFORMING PRE-PROCESSING PHASE

Data Set	*Size(MB)*	*Number Of Hits*	*Number of Distinct Users*	*Number of Visits*

Data Set	*Size(MB)*	*Number Of Hits*	*Number of Distinct Users*	*Number of Visits*
Shivaqua Paddel Wheel aerator	1005.64	15,801	464	825

As shown in Figure 1, the percentage of sessions formed by a predefined number of pages quickly decreases when the minimum number of pages in a session increases. First all the uninteresting entries from the web access log file are removed. For example, samples of these extra inputs are cited in Table 2 along with the number of their repetition shivaqua Paddel wheel aerator web access log. Once the users sessions are reconstructed based on clustering algorithm .clustering operation is calculated based on varying values of MinFreq and α, the percentage of pages clustered is calculated.

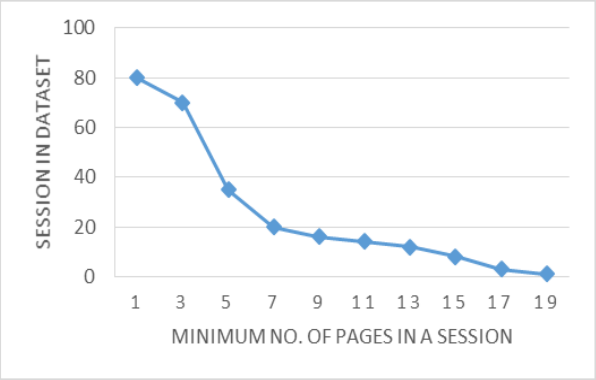

Fig. 1. Minimum number of pages in session

The tests showed that the percentage of participated pages for value α = 0.8 is at the best status. In Figure 2, the percentage of clustered pages is represented as a function of the parameter MinFreq and for two values α = 1.0 and α =0.8.

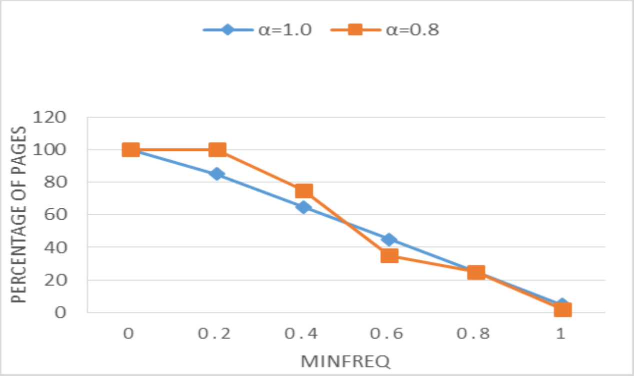

Fig. 2. Percentage of pages in session

Figure 3 show the number of achieved clusters for two values α=1.0 and α=0.8 as a function of the MinFreq parameter.

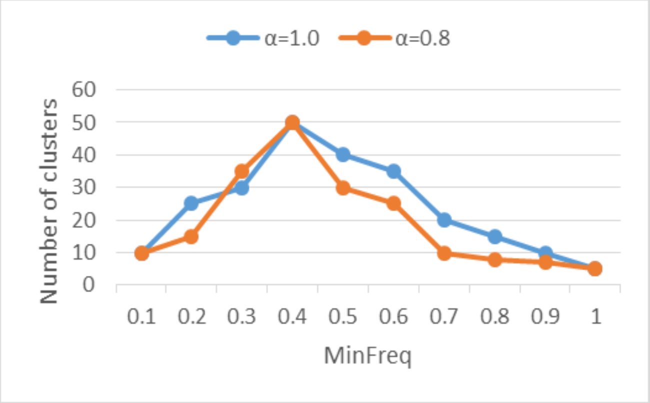

Fig. 3. Number of clusters found

To calculate the quality of clusters found for varying values of α, the Visit-similar index. The number of clusters are found in Minfreq.

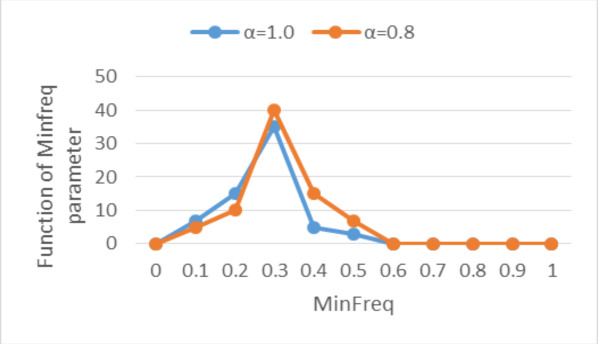

Fig. 4. Similarity or parallel visit

In Figure 4, the value of p, is represented as a function of the MinFreq parameter for two values α. As shown in the figure 4, using our proposed clustering algorithm has enhanced clusters' quality.

5.CONCLUSION

In order to extract web log data that can be effectively used for User support on the Internet, profiles of navigation patterns on web sites are identified. In this paper, the sequence order method and association distance measure to identify the distance measure between similar page. The KHM algorithm Calculate the page path similarity and co-occurrence of clusters using the Shivaqua Paddel wheel Aerator Data sets.

REFERENCES

[1] Birgit Hay, GeertWets and Koen Vanhoof "Mining Navigation Patterns Using a Sequence Alignment Method"

Knowledge and Information Systems (2004) 6: 150–163, Springer, 2004.

[2] Heidar Mamosian, Amir Masoud Rahmani, Mashalla Abbasi Dezfouli “A New Clustering Approach based on Page's Path Similarity for Navigation Patterns Mining” (IJCSIS) International Journal of Computer Science and Information Security,Vol. 7, No. 2, 2010.

[3] Ruili Geng and Jeff Tian “Improving Web Navigation Usability by Comparing Actual and Anticipated Usage” IEEE Transactions On Human-Machine Systems, Vol. 45, No. 1, February 2015

[4] Mini T. V. Nedunchezhian R and Vijayakumar V “A Hybrid Pre-processing Approach for Temporal Associative Rule Classification” IJCTA, 10(07), 2017, pp. 195-203

[5] Dharmarajan K., and Dr M A Dorairangaswamy. "Web Usage Mining: Improve The User Navigation Pattern Using Fp-Growth Algorithm." Elysium journal of

[10] 2013).

engineering research and management (EJERM) 3.4 (2016).

[6] Sampada Khorgade1, Praful Sambhare “Web Recommendation System Based on Approach of Mining Frequent Sequential Patterns” IJLERA) ISSN: 2455-7137 Volume – 02, Issue – 01, January – 2017.

[7] Patil, Ruchika, and Amreen Khan. "Bisecting K-Means for Clustering Web Log data." International Journal of Computer Applications 116.19 (2015).

[8] Hirudkar, Arpita M., and S. S. Sherekar. "Comparative analysis of data mining tools and techniques for evaluating performance of database system." Int J Comput Sci Appl 6.2 (2013): 232-237.

[9] Garg, Kanwal, and Deepak Kumar. "Comparing the performance of frequent pattern mining algorithms." International Journal of Computer Applications 69.25 (