A Webpage Similarity Measure for Web Sessions Clustering Using Sequence Alignment

Mozhgan Azimpour-Kivi
School of Engineering and Science
Sharif University of Technology, International Campus
Kish Island, Iran
mojgan_az@kish.sharif.edu

Reza Azmi
Department of Computer Engineering
Alzahra University
Tehran, Iran
azmi@alzahra.ac.ir

***Abstract*—Web sessions clustering is a process of web usage mining task that aims to group web sessions with similar trends and usage patterns into clusters. This process is crucial for effective website management, web personalization and developing web recommender systems. Accurate clustering of web sessions is highly dependent to the similarity measure defined to compare web sessions. In this paper, we propose a similarity measure for comparing web sessions. The sequential order of web navigations in sessions is considered using sequence alignment method. Furthermore, we propose to consider the usage similarity of two web sessions based on the time a user spends on a webpage, and also the frequency of visit of each page within the session. The proposed method is validated by clustering a collection of web sessions using an agglomerative clustering technique and comparing the results with available methods. The experimental results show effectiveness of the proposed method to capture the properties of web session data.**

Keywords-interestingness of webpage; webpage simialrity measure; sequence alignment; web sessions clustering

I. INTRODUCTION

The World Wide Web (WWW) is considered as the largest distributed collection of information. By rapid growth of this information resource, it has become a difficult task for users to acquire their desired information even in a particular website. Hence, a need for developing techniques that can facilitate this issue has been highlighted. Knowing the users and making profiles of them can be helpful for websites to present relevant information to particular visitors. To address this issue, web usage mining has recently attracted many attentions [1]. Web usage mining is an application of data mining which tries to extract useful patterns from data that are obtained from the interaction of users with the web. Deploying web mining techniques are crucial for any application that aims to ease the use of the web such as creating adaptive websites, web personalization, web recommender systems, etc. Web usage mining from web access log files has three steps [1]: (1) data pre-processing; (2) pattern discovery by applying various techniques such as clustering, classification, association discovery, and sequential pattern discovery to the data; and (3) pattern analysis which aims at eliminating irrelevant patterns from the discovered patterns in previous step.

In our study, we are interested in the web sessions clustering which is the problem of grouping web sessions with similar usage patterns into groups. Any clustering process tends to maximize the intra-group similarity and minimize the inter-group similarity of cluster objects [2]. One of the most challenging issues in clustering web session is how to measure the similarity of two web sessions. A more precise similarity measure can definitely be more helpful for investigating the nature of the data. The most popular measures that are used for web sessions clustering are Euclidean distance, Cosine similarity measure, and Jaccard coefficient. We should keep in mind that a web session contains a sequence of URLs accessed by a user. Therefore, a good similarity measure should be defined so that it does not ignore the sequential nature of web navigations in sessions. On the other hand, not all of the URLs visited in a session are equally important to the user.

In this paper, we borrow the idea of sequence alignment [3] from bioinformatics in order to find the best match of two session sequences. Sequence alignment is one of the fundamental operations in bioinformatics in order to capture the relationships between DNA sequences. Similar to each session which consists of a sequence of web pages, each DNA contains a sequence of amino acids. Consequently, techniques used in DNA sequences alignment can be applied to measure the similarity of web sessions. Similar to the DNA sequence alignment, the problem of computing the similarity between web sessions can be facilitated by using dynamic programming techniques [3]. In addition to the sequential nature of web sessions, we consider the time a user spends on a webpage and also the frequency of the visitation from a particular webpage in a web session in order to estimate the importance of that page to the user. Using the Silhouette coefficient [2] of the obtained clusters as the evaluation measure, we compare our method with available methods. The results show that our method is more effective in investigating the similarity of web sessions for web sessions clustering task.

The remainder of this paper is organized as follows. In Section 2, a review on some available methods for clustering web sessions is presented. In Section 3, we introduce a new method for estimating the similarity of two sessions using sequence alignment. The experimental results for evaluating the proposed method are presented in Section 4. Finally, Section 5 concludes our study. Also, future works are presented in this section.

978-1-4244-9834-5/11/$26.00

II. RELATED WORK

Different similarity measures have been proposed for capturing web sessions' similarities. Also, various clustering algorithms have been introduced in literature for grouping web users with common practices. Most of these works represent a session using a vector defined over the space of web pages within a particular website: a vector dimension corresponds to specific URL within the website. Depending on the values assigned to these dimensions, different user behavior analysis can be performed. The most common method is to associate binary values to a dimension, i.e. the value 1 for pages which the user has visited them in the session and 0 for the others. Nasraoui et al. [4] used this representation for web sessions and deployed the normalized cosine of the angle between the two vectors as the similarity of them for web sessions clustering. Some other methods have been proposed to use feature weights based on the time a user spends on a particular webpage (perhaps normalized by the size of the webpage) or the frequency of occurrence of a URL within the user session instead of binary weights [5], [6]. Yan et al. [5] applied Euclidean distance in web user clustering task and then suggested some links to web users according to their corresponding cluster. None of the mentioned methods captures the sequential nature of web navigations within the sessions.

In a newer attempt, Banerjee and Ghosh [7] used the relative time spent on the longest common sub-sequence between two sessions, found through dynamic programming, as the similarity between two sessions. The authors then built an abstract similarity graph for the set of sessions and applied the graph partitioning methods in order to cut the abstract graph into clusters. Wang and Zaiane [8] introduced a new method to measure similarities between web sessions based on sequence alignment in computational biology. In this method, they first defined a similarity between two web pages using hierarchical structure of URLs in the website. Then, they utilized dynamic programming to find the best match between two session sequences. In the method presented in [9], the accurate viewing time of the accessed pages are considered in addition to the URL of pages for defining the similarity of two web pages. Similar to [8], a dynamic programming process is then applied to find the best mach for two sessions. In [10], an algorithm for Web Session Clustering Based on Increase of Similarities (WSCBIS) is presented using the method proposed in [9]. This algorithm decreases the time and space complexity of clustering compared to k-means and Robust Clustering using links (ROCK) [11]. Hay et al. [12] have clustered web users using two different similarity measures: Sequence Alignment Method (SAM) and Association measure (Euclidean distance-based measure). In SAM, the sequential order of requests is taken into consideration and not the position of them. The results proved that SAM retrieves sequences not only with similar pages, but the order of pages is also considered compared to the associative measure.

The method proposed in this paper has similar basic idea to the methods presented in [9]. Compared to this work, we consider not only the time a user spends on a webpage, but also the frequency of the visitation from a particular webpage in order to estimate the interestingness of that page to the user in a session. Then, we define the similarity of two web pages based on the conjunction of the similarity of the web pages' URLs and the similarity of their interestingness to the users. Finally, we employ sequence alignment in order to find the best match of two sessions and estimate the similarity of two sessions.

A. Web Page Similarity Based on the URLs

Wang and Zaiane [8] proposed a method to measure the similarity of two different web pages. This method does not consider the content of web pages but simply the paths leading to a webpage in the hierarchical structure of the URLs of the website. The detail information of this method is discussed in [8]. Briefly, in order to measure the similarity of two web pages, we first represent each level of their URLs by a token. As a result, the token string of the full path of a URL is the concatenation of all the representative tokens for each level. The token for each level is assigned based on the hierarchical structure of the website. Marking the tree structure of a nominal website is illustrated in Fig.1. In order to compute the similarity of two web pages, we first determine the length of the longest token string among the two. Then, we give a weight to each level of tokens from the last to the first. The last level of the longest token string is given a weight equal to 1, the second to the last is given weight 2, and so on. Finally, the similarity between two token strings (*Token Sim*) is defined as the sum of the weights of those matching tokens divided by the sum of the total weights. The obtained similarity for pair of web pages ranges from 0, for the pages without any identical token in identical place, to 1, for pages that are exactly the same. An example of this process for two nominal URLs (example-website/A/C.html and example-website/A/B.html) is presented in Fig. 2. For this example, the similarity between two token strings is (3+2) / (3+2+1) = 0.83.

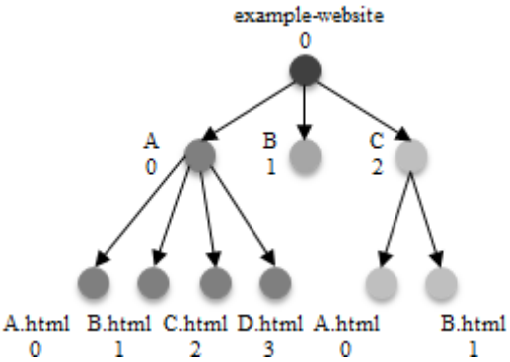

Figure 1. Marking the URL tree of a nominal website.

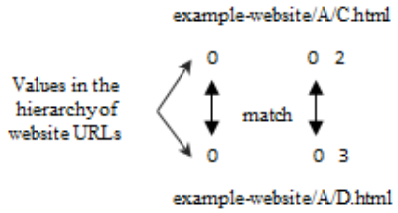

Figure 2. An example of token string comparison.

B. Web Page Similarity Based on the Importance to the User

As mentioned, we use the similarity of two web pages with conjunction of the similarity between the interests of their users in visiting those pages. In a particular session, the normalized frequency of the visit of the i^{th} webpage (P_i) and the time spent on this page can be represented using (1) and (2) respectively. In (1), the *Frequency(P_i)* is simply the number of times the webpage P_i has been visited in the session. In (2), the *Time Spent on(P_i)* is the difference between the exact time of the request of page P_i and the time of the request for the next webpage in the session from the access log file. Consider that, we cannot compute this value for the last webpage requested in the session. Here, we define the time spent on the last webpage of a session as the average time spent on other web pages of the session. Considering the fact that the length of a webpage (in bytes) can have impact on the time that is needed to visit that page, we have normalized the spent time on pages by dividing this value by the length of the corresponding webpage as shown in (2).

$$\text{Freq}(P_i) = \frac{\text{Frequency}(P_i)}{\sum_{P \in \text{All pages in session}} \text{Frequncey}(P)} \quad (1)$$

$$\text{SpentTime}(P_i) = \frac{\frac{\text{Time Spent on}(P_i)}{\text{Length}(P_i)}}{\sum_{P \in \text{All pages in session}} \frac{\text{Time Spent on}(P)}{\text{Length}(P)}} \quad (2)$$

In (2) and (3), frequency of the visitation of a webpage and the spent time on that page are normalized by their denominator which is the sum of these values for the whole requests in the session. These two measures should be combined to describe the interestingness of a webpage to a user. In mathematics, the harmonic mean is one of the several kinds of average. In our case, we use the harmonic mean of the *Freq* and *SpentTime* for page P_i as the measure of interestingness of this page to a user in one session. This value can be described using (3).

$$\text{Interest}(P_i) = \frac{2}{\frac{1}{\text{Freq}(P_i)} + \frac{1}{\text{SpentTime}(P_i)}} \quad (3)$$

Finally, we define a measure to estimate the similarity of the interestingness of two visitors from i^{th} page (P_i) and j^{th} (P_j) page in a session. This value can be estimated using (4). Consider that P_i and P_j can also belong to different sessions.

$$\text{Interest Sim}(P_i, P_j) = \frac{\min\{\text{Interest}(P_i), \text{Interest}(P_j)\}}{\max\{\text{Interest}(P_i), \text{Interest}(P_j)\}} \quad (4)$$

Considering the similarity between pair of pages based on their URLs (*Token Sim*) and their interestingness to the user (*Interest Sim*), we can define the similarity of two web pages as shown in (5). In this equation, the parameter α is a scale factor which should be associate with a value between 0 and 1.

$$\text{Similarity}(P_i, P_j) = \alpha\ \text{Token Sim}(P_i, P_j) + (1-\alpha)\ \text{Interest Sim}(P_i, P_j) \quad (5)$$

C. Similarity of Web Sessions

As mentioned earlier, we consider each session as a sequence of URLs that are requested by the user. For estimating the similarity of two web sessions, we apply the sequence alignment method in order to find the best match between two sequences. In aligning two sequences, not only characters that match identically are considered, but also spaces or gaps (or conversely, insertions in the other sequence) and mismatches, both of which can correspond to mutations. In sequence alignment, we want to find an optimal alignment that, loosely speaking, maximizes the number of matches and minimizes the number of spaces and mismatches.

For applying the sequence alignment method, we need to define a scoring function which helps find the optimal matching between two session sequences. Consider that, the similarity of two web pages discussed in previous section plays the role of a page matching goodness function. The scoring function deployed in the method of this paper is as follows. For each identical matching, i.e. a pair of pages with similarity 1, the score is 20; for each mismatching, i.e. a pair of pages with similarity 0, or matching a page with a gap, the score is –10; for a pair of pages with similarity $\beta \in (0,1)$, the score for their matching is between -10 and 20. Hence, the scoring function for the similarity between two pages P_i and P_j can be calculated using (6). Consider that, the parameter β is actually the similarity between pages P_i and P_j which was calculated in previous section.

$$\text{Score}(Pi, Pj) = -10 + 30\beta \quad , \quad 0 \le \beta \le 1 \quad (6)$$

As mentioned earlier, the estimation of the similarity between two web sessions is calculated using the sequence alignment method in order to find the best match between two sequences. The final similarity between the two sequences is obtained based on their optimal matching and the length of the sequences. An optimal matching is an alignment with the highest possible score. As mentioned earlier, the problem of finding the optimal matching of two sequences can be facilitated using dynamic programming, i.e. the similarity of two sessions sequences can be computed by considering the contribution of the similarity of pages in the head of each sequence and the maximum similarity in the remaining sub-sequence. This process can be described using a matrix in which one sequence (session) is placed along the top and the other sequence (session) is placed along the left side of the matrix. An example of such matrix is illustrated in Fig. 3. In this figure, each webpage in a session is shown using corresponding token to each level of the webpage URL in the structure of the website. Also, the time the user has spent on each webpage is shown in parentheses for each webpage.

	-	/1/// (5)	/1/2// (15)	/1/2/7/ (8)	/1/3// (24)	/2/// (38)	/2/1// (18)
-	0	-10	-20	-30	-40	-50	-60
/1/// (4)	-10	18.33	8.33	-1.66	-11.66	-21.66	-31.66
/1/2// (11)	-20	8.33	37.08	27.08	17.08	7.08	-2.91
/1/2/3/ (25)	-30	-1.66	27.08	48.18	38.18	28.18	18.18
/2/// (16)	-40	-11.66	17.08	38.18	46.22	57.67	47.67
/2/5// (14)	-50	-21.66	7.08	28.18	36.75	58.17	69.17

Figure 3. An example of session matching matrix.

In order to calculate the optimal matching using the sequence alignment matrix, a gap is added to the start of each sequence which indicates the starting point of the matching. The goal is to find an optimal path from the top left corner to the bottom right corner of the matrix. In each step, we can only have a right, down or diagonal move. A right move corresponds to inserting a gap to the sequence in the left and matching the sequence on top with a gap, while a down move corresponds to inserting a gap to the sequence on top and matching the sequence on left with a gap. In each step, the score for each three moves is calculated and the maximum of them is added to the current score, which had been obtained from previous moves. In other words, the direction which provides maximum score is chosen in each step. The optimal path is then achieved through back propagating from bottom right corner to the starting point. In the given example, the optimal path found through back propagating is shown by arrows in Fig. 3. The score that is put in the lower right corner is the optimal sequence alignment score. In our scoring system, the optimal score cannot be bigger than the length of the shorter session multiplied by 20. Also, it cannot be smaller than the length of the longer session multiplied by -10. Therefore, the final similarity measure can be calculated by normalizing the optimal score with respect to these maxima and minima. In the case of our example, the similarity of two web sessions is $[69.17 - (-10 \times 6)] / [(20 \times 5) - (-10 \times 6)] = 0.81$. Using this definition, the similarity value for two web session will always be between 0 and 1.

III. Experimental Evaluation

Many attempts have been made to evaluate the clustering goodness and to find rules to quantify the quality of a clustering result. Cluster validation for large datasets of categorical data such as web session data is a very hard task. Previous works that proposed to use the sequence alignment method to cluster web session data, validates their experimental results manually rather than quantitatively [8], [9]. Consider that, due to the large number of web pages in a typical website, usually a large number of web sessions are needed to fully represent possible usage patterns over that website. However, as the sequence alignment method is a time consuming process, dealing with a large dataset increases the processing time for constructing clusters.

The empirical evaluation reported in this paper concerns the question whether the proposed method, which considerers both the similarity of web pages based on their URLs and the usage similarity of them in order to define a scoring function for sequence alignment method, can properly reflect the nature of the session data. Web usage data that are used for this experiment are collected from the Music Machine website by Perkowitz and Etzioni [13]. The original data, used in our experiment, contains 62668 requests from log files of the web server for five random days in 1998. After preprocessing task on the original data using methods described in [14], 2664 web sessions were extracted. We have compared our method to the one presented in [9], which we refer to as the extended Sequence Alignment (SA) method. As described earlier, in the extended SA method, only the accurate viewing time of the accessed pages are taken into consideration for defining the usage similarity between two web sessions. This method is claimed to perform better in revealing the nature of data compared to the normal SA method presented in [8]. Also, it has been proved that the SA method can effectively reflect the sequential nature of web navigations in web sessions clustering compared to other similarity measures such as Jaccard coefficient [8].

To compare our method with extended SA, two distance matrices holding pairwise sequence alignment distance measures between web sessions are obtained using our proposed method and the extended SA. Considering the fact that the similarity measures obtained from both methods always range from 0 to 1, the distance between two sessions can be calculated by subtracting the similarity of two web sessions from 1. Finally, the obtained distance matrices can be used for clustering the web sessions. For our experiment, we have applied the agglomerative clustering using average linkage method [1] to construct a linkage tree. The average linkage method is not very susceptible to noise and outliers in the input data. After constructing the linkage tree, we cut off the tree in order to generate desired number of clusters from the session data. In this experiment the parameter α used in (5) is set to 0.7.

For evaluating the effectiveness of the similarity measure in this paper, the average Silhouette Coefficient (SC) [2] is calculated for web sessions data in their clusters. The Silhouette coefficient considers both cohesion and separation of data points for evaluating clusters. The Silhouette coefficient value for i^{th} web session (s_i) can be calculated using (7).

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{7}$$

In (7), a_i is the average distance of i^{th} sessions from the other sessions in its corresponding cluster. The parameter b_i is the minimum of the average distances from the i^{th} session to the sessions in other clusters. The value of Silhouette coefficient can vary from -1 to 1; the closer the value to 1, the better the clustering result. The average Silhouette coefficient values for different number of clusters (range from 2 to 25) for the two distance matrices are shown in Fig. 4.

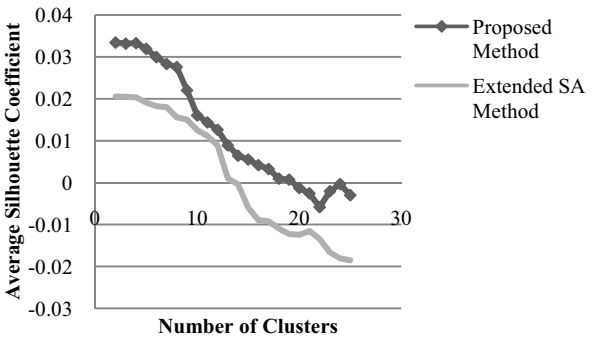

Figure 4. The comparison of the average SC values for different number of clusters, generated by using the pairwise distance measure matrices of proposed method and the extended SA method.

As we can see in Fig. 4, the average SC values for the clusters generated by the distance measure of the proposed method has higher values compared to the SC values of the clusters generated by the distance measure of the extended SA method. This result can show our approach can properly reflect the nature of the session data since, not only we have considered the sequential nature of web sessions, but also we have taken into account the similarity of the usage patterns for comparing two web sessions. As we can see in Fig. 4, the average SC value reduces by increasing the number of clusters. The reason is that, by increasing the number of clusters we cut off the linkage tree in lower values of distance between clustering objects. The result would be smaller clusters which may have higher inter-cluster similarities.

IV. CONCLUSIONS AND FUTURE WORK

As described in this paper, web session clustering is an important task to group web sessions with similar trends. This is an essential process for effective website management, web personalization, and web recommender systems. Accurate clustering of web sessions is highly dependent to the similarity (or dissimilarity) measure defined to compare web sessions. In this paper, we proposed a new similarity measure for web sessions clustering. We considered the time a user spends on a webpage and also the frequency of the visitation from a particular webpage within a session in order to estimate the interestingness of that page to the user. Then, we defined the similarity of two web pages within two sessions based on the conjunction of the similarity of the web pages and the similarity of their interestingness to the users. Finally, we employed the sequence alignment method in order to find the best match of two session sequences and estimate the similarity of them. We compared our method with a case in which only the spent time on a webpage was considered for defining the usage pattern similarity. The evaluation was performed by measuring the average Silhouette coefficient of the sessions which were clustered using agglomerative clustering with average linkage method. Experimental results verify the effectiveness of our method. However, the time complexity of SA methods is still high.

As described, we estimated the similarity of two web pages based on the similarity of their hierarchical structure of URLs while ignoring the content of web pages. To have a better estimation of the similarity of two web pages, we can use other methods proposed for web content mining such as Information Retrieval or semantic web approaches. Furthermore, for having a more general evaluation, we can use a larger collection of web sessions data and apply different clustering algorithms on these data.

REFERENCES

[1] T. Hussain, S. Asghar, and S. Fong, “A hierarchical cluster based preprocessing methodology for Web Usage Mining,” in 6th International Conference on Advanced Information Management and Service (IMS), 2010, pp. 472-477.

[2] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, Nov. 1987.

[3] K. Charter, J. Schaeffer, and D. Szafron, “Sequence alignment using FastLSA,” in International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS), 2000, p. 239–245.

[4] O. Nasraoui, C. C. Uribe, C. R. Coronel, and F. Gonzalez, “TECNO-STREAMS: tracking evolving clusters in noisy data streams with a scalable immune system learning model,” in Third IEEE International Conference on Data Mining, ICDM, 2003, pp. 235-242.

[5] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, “From user access patterns to dynamic hypertext linking,” Computer Networks and ISDN Systems, vol. 28, no. 7-11, pp. 1007-1014, May. 1996.

[6] R. Forsati, M. R. Meybodi, and A. Rahbar, “An efficient algorithm for web recommendation systems,” in IEEE/ACS International Conference on Computer Systems and Applications, Los Alamitos, CA, USA, 2009, vol. 0, pp. 579-586.

[7] A. Banerjee and J. Ghosh, “Clickstream clustering using weighted longest common subsequences,” in Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, 2001, p. 33–40.

[8] Weinan Wang and O. R. Zaiane, “Clustering Web sessions by sequence alignment,” in In Proceedings 13th International Workshop on Database and Expert Systems Applications, 2002, pp. 394-398.

[9] Chaofeng Li and Yansheng Lu, “Similarity Measurement of Web Sessions by Sequence Alignment,” in IFIP International Conference on Network and Parallel Computing Workshops, NPC Workshops, 2007, pp. 716-720.

[10] C. Li, “Research on Web Session Clustering,” Journal of Software, vol. 4, no. 5, Jul. 2009.

[11] S. Guha, R. Rastogi, and K. Shim, “Rock: A robust clustering algorithm for categorical attributes,” Information Systems, vol. 25, no. 5, p. 345–366, 2000.

[12] B. Hay, G. Wets, and K. Vanhoof, “Segmentation of visiting patterns on web sites using a sequence alignment method,” Journal of Retailing and Consumer Services, vol. 10, no. 3, pp. 145-153, May. 2003.

[13] M. Perkowitz and O. Etzioni, “Adaptive sites: Automatically learning from user access patterns,” in Proceedings of 6th International World Wide Web Conference, Santa Clara, California, 1997.

[14] V. Sathiyamoorthi and V. M. Bhaskaran, “Data Preparation Techniques for Web Usage Mining in World Wide Web-An Approach,” International Journal of Recent Trends in Engineering, vol. 2, no. 4, 2009.