ABSTRACT

Customized web services are offered to users by grouping them according to their access patterns. Clustering techniques are very useful in grouping users and analyzing web access patterns. Clustering can be an object clustering performed on feature vectors or relational clustering performed on relational data. The relational clustering is preferred over object clustering for web users’ sessions because of high dimensionality and sparsity of web users’ data. However, relational clustering of web users depends on underlying dissimilarity measures used. Therefore, correct dissimilarity measure for matching relational web access patterns between user sessions is very important. In this chapter, the various dissimilarity measures used in relational clustering of web users’ data are discussed. The concept of an augmented user session is also discussed to derive different augmented session dissimilarity measures. The discussed session dissimilarity measures are used with relational fuzzy clustering algorithms. The comparative performance binary session similarity and augmented session similarity measures are evaluated using intra-cluster and inter-cluster distance-based cluster quality ratio. The results suggested the augmented session dissimilarity measures in general, and intuitive augmented session (dis)similarity measure, in particular, performed better than the other measures.
INTRODUCTION

The access data generated by Web applications are growing with phenomenal rate, and this web usage data is used to gain extra knowledge about the users’ navigation patterns. The knowledge extracted from the accesses patterns of users may be utilized to serve the needs of the user in a better way. The access patterns of Web users are stored in the server logs and logs are partitioned into user sessions for analysis.

The user sessions are represented as a feature vector of accessed page URLs. The number of accessed page URLs represented the dimension and assigned binary values such as zero and one. Where one and zero denotes the accessing and not accessing respectively of web page (URL) in an individual session (Nasraoui, Hichem, Krishnapuram, & Joshi, 2000).

Let’s consider \( n \) number of web pages in any website then Eq. (1) may be used to represent the \( i^{th} \) user session as \( n \)-dimensional binary vector space of web pages.

\[
S_i^j = \begin{cases} 
1, & \text{if the user accessed the } j^{th} \text{ web page in } i^{th} \text{ session} \\
0, & \text{Otherwise,}
\end{cases} \quad j = 1, 2, 3 \ldots n \tag{1}
\]

The similarity measures are defined between any two user sessions using binary feature vector representation of Eq. (1) and incorporating accessed URLs along with their For better understanding and clear distinction conventional web user session similarity measures are renamed as binary session similarity (BSS), binary URL syntactic similarity (BUSS) and combined binary session similarity (CBSS) (Sisodia, Verma, & Vyas, 2016d). A simple binary session similarity measure between web user sessions \( S_i^k \) and \( S_i^l \) is given by Eq. (2). In this similarity measure individual accessing web page URLs in any user session are completely ignore the syntactic structure of URL and assumed totally independent.

\[
BSS(S_i^k, S_i^l) = \frac{\sum_{j=1}^{n} S_i^k \times S_i^l}{\sqrt{\sum_{j=1}^{n} S_i^k} \times \sqrt{\sum_{j=1}^{n} S_i^l}} \tag{2}
\]

The main limitations of BSS measure are that it completely ignores the syntactic structure of accessing URLs. In literature (Nasraoui et al., 2000) an alternative URL based syntactic similarity measure is also reported to overcome the limitations of BSS which is renamed as BUSS. The BUSS measure employed the syntactic similarity between any pair of URLs given by Eq. (3).

\[
BUSS(S_i^k, S_i^l) = \text{Min} \left\{ 1, \frac{\left| \text{LoP}\left(P_{(k,i)}\right) \cap \text{LoP}\left(P_{(l,j)}\right) \right|}{\text{Max}\left(1, \text{Max}\left(\text{LoP}\left(P_{(k,i)}\right), \text{LoP}\left(P_{(l,j)}\right)\right) - 1\right)} \right\} \tag{3}
\]