Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 279 (2015) 40–63

FUZZY sets and systems

www.elsevier.com/locate/fss

Mountain density-based fuzzy approach for discovering web usage clusters from web log data

Zahid Ansari [a,*], Syed Abdul Sattar [b], A. Vinaya Babu [c], M. Fazle Azeem [d]

[a] *Department of Computer Science, P.A. College of Engineering, Mangalore, India*
[b] *Department of Computer Science, Royal Institute of Technology, Hyderabad, India*
[c] *Department of Computer Science, J.N.T.U. College of Engineering, Hyderabad, India*
[d] *Department of Electrical Engineering, Aligarh Muslim University, Aligarh, India*

Received 8 July 2014; received in revised form 30 January 2015; accepted 31 January 2015

Available online 4 February 2015

Abstract

Due to the continuous proliferation of e-businesses, there is intense competition among organizations to attract and retain customers. Analyses of the web server logs of these organizations are critical for obtaining insights into web usage behavior, which can support the design of more attractive web structures. In this study, we propose a mountain density function (MDF)-based fuzzy clustering framework for discovering user session clusters in web log data. The major steps in this framework include web log preprocessing, MDF-based discovery of fuzzy user session clusters, and validation of these clusters. To consider the high dimensionality of user session data, we propose a fuzzy approach for assigning weights to user sessions. Fuzzy c-means (FCM) and fuzzy c-medoids (FCMed) algorithms are used to cluster the user sessions. The selection of suitable initial cluster centers is a major challenge for these methods, so we propose MDF-based FCM (MDFCM) and FCMed (MDFCMed) algorithms to overcome this problem. MDF-based clustering is also used to estimate the number of clusters. Our results clearly indicate that the quality of the clusters formed using the proposed algorithms is much better in terms of various validity measures compared with the FCM and FCMed algorithms.

Keywords: Fuzzy cluster validation; Fuzzy clustering; Mountain density function; User session clustering

1. Introduction

The World Wide Web is a huge source of information and a great resource for the application of data mining. Web usage mining involves the automatic discovery and analysis of usage patterns based on data collected from user interactions with web resources on one or multiple web sites [1,2]. The patterns discovered are usually represented as collections of web resources that are accessed frequently by groups of users with common interests [3–5]. Clustering

[*] Corresponding author. Tel.: +919341623210.
E-mail address: zahid_cs@pace.edu.com (Z. Ansari).

http://dx.doi.org/10.1016/j.fss.2015.01.021
0165-0114/© 2015 Elsevier B.V. All rights reserved.

algorithms have been applied widely to discover user session clusters that represent similar URL access patterns. Details of various clustering techniques can be found in [6–8]. Partitioning-based user session clustering starts with a set of user session vectors, which are partitioned into c clusters using similarity distance measures, such as Euclidean distance. The extracted clusters represent the access patterns of users based on their common navigational behavior. In [9–11], a framework was presented for user session clustering based on hard clustering algorithms where the results obtained were compared using various performance and validity measures. Chaofeng [12] provided a framework for web user session clustering by c-means and robust clustering with links algorithms. Lee and Fu [13] clustered web user browsing features by hierarchical agglomerative clustering.

Fuzzy clustering leads to the formation of overlapping clusters, where each data object can belong to several clusters with different degrees of membership, thereby allowing the clusters to grow into their natural shapes [14]. Arotaritei and Mitra [15] provided a web mining survey of various fuzzy sets-based clustering techniques. One of the most popular fuzzy clustering techniques is fuzzy c-means (FCM), which was proposed by Dunn [16] and modified by [17–21]. The FCM algorithm performs fuzzy clustering such that a given object may belong to several clusters where the degree of belongingness is specified by membership grades. The algorithm computes the cluster centers and assigns a membership value to each user session that corresponds to every cluster within a range of 0 to 1. Lingras et al. [22] applied the FCM clustering algorithm to discover the usage patterns in web log data. In [23], an FCM-based model was proposed for obtaining user profiles from web usage data. Suresh et al. [24] proposed an FCM clustering algorithm that utilizes information entropy for cluster center initialization. Another category of fuzzy clustering algorithms known as fuzzy c-medoids (FCMed) algorithms [25] extend the hard c-medoids algorithm [26] by incorporating fuzzy set concepts to produce fuzzy clusters [27,28]. These algorithm have been used to discover fuzzy clusters of web user sessions in a given set of user sessions. Each cluster is represented by a representative user session object as the medoid of that cluster [25,29]. Yager and Filev proposed a mountain clustering technique to find the cluster centeres based on a density measure called the mountain function [30,31].

There are several reasons for selecting FCM and FCMed algorithms to discover interesting usage patterns in web user session data, as follows.

- Web usage data are unlabelled so they do not contain any class information. The FCM/FCMed algorithms can cluster similar user sessions in an efficient manner based on the frequencies at which URLs are accessed during user sessions [32,33].
- Web user session data usually contain inaccurate, inconsistent, and missing information. These weaknesses have a negative impact on the cluster discovery process. Thus, the clusters formed might not be reliable and trustworthy. However, the FCM/FCMed techniques utilize the fuzzy membership concept in fuzzy sets, which are more robust against imperfections, so they are more suitable than traditional hard clustering techniques for pattern discovery in imperfect data [14].
- Due to the non-deterministic browsing patterns of various web users, user session data do not have crisp boundaries and they often form overlapping clusters [34]. Because of the overlapping nature of web user session data, FCM/FCMed clustering techniques can be applied very well to form overlapping clusters, where each user session object can belong to several clusters with different degrees of membership.
- Moreover, FCM/FCMed algorithms are simple, efficient, easy to implement, and they have been used widely for mining web usage data [25].
- The FCMed algorithm has the added advantage that it is more robust against noise compared with FCM [35].

One of the major problems associated with the FCM and FCMed methods is their sensitivity to the initial selection of the cluster centers. Therefore, estimating suitable values for the initial cluster centers is a major challenge associated with these methods. In this study, we propose the use of a mountain density function (MDF)-based initialization strategy to find a suitable set of initial cluster centers. MDF-based clustering is also used to estimate suitable values for the number of clusters parameter c. A novel user session clustering framework is proposed, which integrates the MDF cluster center initialization strategy with the FCM and FCMed algorithms. Fig. 1 provides an overview of the proposed MDF-based fuzzy clustering framework for discovering web user session clusters, which aims to improve the quality of the clusters extracted. There are three main steps in the proposed framework: i) web log data preprocessing and the assignment of fuzzy weights to the extracted user sessions, ii) discovering user session clusters with MDF-based