International J.of Multidispl.Research & Advcs. in Engg.(IJMRAE),
ISSN 0975-7074, Vol. 6, No. I (January 2014), pp. 33-54

REFINEMENT OF CLUSTERS BASED ON DISSIMILARITY MEASURES

V. S. DIXIT[1] AND SHVETA KUNDRA BHATIA[2]
[1] Computer Science Department, Atma Ram Sanatan Dharma College,
University of Delhi, New Delhi 110021, India
[2] Research Scholar, Computer Science Department, University of Delhi,
New Delhi, India

Abstract

Session clustering is one of the ways to improve web site structure, providing recommendations and web personalization implemented by creating a pool of web pages from similar sessions within a cluster. Session clustering depends on the various similarity/dissimilarity measures used to compare the sessions. It is one of the most sought research problem in Web Usage Mining these days. In this paper we propose a refinement technique for session clusters with the purpose of solving the problem of scalability of log data for reducing the domain of recommendations to the end user. This technique considers the page access, access time and session weight dissimilarity within a cluster for refinement. We also present an analysis of different available dissimilarity measures like Simple Difference, Jaccard, Variance, Pattern Difference and BLWMN based on the above defined features for evaluating dissimilarity. The results show that clusters generated by refinement using a combination of web page access (Jaccard Dissimilarity) and time (Cosine Dissimilarity) features are of a good quality as compared to other feature combinations. The results are compared with the available dissimilarity measure based on the feature of number of hits on a page in a session. Based on the experimental results the proposed approach resulted in better quality clusters. Quality analysis of refined clusters is done using internal cluster quality measures.

Keywords : Web Usage Mining, Web Session Clustering, Page Access Matrix, Dissimilarity Measures, Modified Knockout Refinement Algorithm (MKRA), Daveis Bouldin Index, Dunn's Index.