Abstract

Clustering web usage data is useful to discover interesting patterns related to user traversals, behavior and their characteristics, which helps for the improvement of better Search Engines and Web personalization. Clustering web sessions is to group them based on similarity and consists of minimizing the Intra-cluster similarity and maximizing the Inter-group similarity. The other issue that arises is how to measure similarity between web sessions. There exist multiple similarity measures in the past like Euclidean, Jaccard, Cosine and many. Most of the similarity measures presented in the history deal only with sequence data but not the order of occurrence of data. A novel similarity measure named SSM (Sequence Similarity Measure) is developed that shows the impact of clustering process, when both sequence and content information is incorporated while computing similarity between sequences. SSM (Sequence Similarity measure) captures both the order of occurrence of page visits and the page information as well, and compared the results with Euclidean, Jaccard and Cosine similarity measures. Incorporating a new similarity measure, the existing Density clustering technique DENCLUE is enhanced and the new named as SSM-DENCLUE for Web personalization. The Inter-cluster and Intra-cluster distances are computed using Average Levensthein distance (ALD) to demonstrate the usefulness of the proposed approach in the context of web usage mining. This new similarity measure has significant results when comparing similarities between web sessions with other previous measures, and provided good time requirements of the newly developed SSM-DENCLUE algorithms. Experiments are
performed on MSNBC.COM website (free online news channel), in the context of Density based clustering in the domain of Web usage mining.

References


Index Terms

Computer Science

Data Mining

Keywords

Data Mining Clustering similarity measures Web Personalization DENCLUE SSM-DENCLUE.