Web Document Clustering using Genetic Algorithm

Pyae Sandi Hein, May Aye Khine
University of Computer Studies, Yangon
pyaesandihein@gmail.com

Abstract

Clustering (or cluster analysis) is one of the main data analysis techniques and deals with the organization of a set of objects in a multidimensional space into cohesive groups, called clusters. Each cluster contains objects that are very similar to each other and very dissimilar to object in other cluster. Web page clustering is one of the major preprocessing step in web mining analysis. Clustering is also useful extracting salient features of related web document to automatically formulated queries and search for other similar document on the Web. Web page clustering faces with and many challenges due to the high dimensionality and due to heterogeneity nature of the web document. Efficient and scalable algorithm are need for web clustering. Genetic algorithm is a of the algorithm from evolutionary computing that can effectively search in the large search space by simulating the nature of evolution. This paper present the genetic algorithm for web page clustering that is scalable and efficient. Genetic algorithm with medoid representation was used because it provides shorter chromosome length and medoid based clustering is more tolerable to noisy data such as web document and employs a supervised features selection method for selection of appropriate features terms