Clustering Homogeneous and Heterogeneous XML Documents by Summarizing Edge

May Myat Thu, Khin Nwe Ni Tun
University of Computer Studies, Yangon
maythu321@gmail.com, knntun@ucsy.edu.mm

Abstract

Extensible Markup Language(XML) is a markup language that defines a set of rules for encoding documents in format that is both human-reliable and machine-reliable. Large amount of XML documents on the web require the developed clustering techniques to group.In this system,X-Edge clustering algorithm is applied for clustering of the homogeneous and heterogeneous XML documents in order to utilize in search engine.LevelStructure and LevelEdge are generated from tree and then calculate similarity and distance metrics.X-Edge provides a structure representation of XML documents based on edges summaries.Finally, the outputs of the system are clusters for homogeneous and heterogeneous XML documents. The advantage of this system is that the output clusters can be applied in the search engine.