XML Documents Classification using Composite SVM Kernel

Htet Khine Soe

University of Computer Studies, Yangon
soe.htetkhine@gmail.com

Abstract

More and more structured or semistructured data is stored and exchange in XML format. XML mining becomes increasingly important, especially the study of classification of XML documents. As the number of XMLdocuments on the WWW grows, there arises a need for a classification system for these XML documents that would make organization and querying more effective. Document categorization is the process of classifying text documents into a set of predifined classes. This system presents combination of structure and content information using composite support vector machine (SVM) kernels for XML document classification. Combination of structure and content features is necessary for effective retriveal and classification of XML documents. Composite Kernel classifier achieves significantly better performance as compared to complex and time consuming approaches. Consine Similarity is used to find similarity on terms and paths.