Web Page Classification based on Unsupervised Learning using MIME type Analysis

Luis Roberto Jiménez[(1)]
lrjp@ic.uma.es
[(1)]Communication Engineering Department, University of Málaga, Campus de Teatinos, s/n E-29071, Spain

The properties of a web page have a strong impact on the experience of web users. In this work, a classification method based on unsupervised clustering is proposed to group web pages into classes based on download content that may affect the Quality of Experience (QoE) perceived by the user. Groups are defined based on Multipurpose Internet Mail Extensions (MIME) content breakdown and external subdomain connections, obtained with a desktop personal computer (PC) running WebPageTest tool. The dataset is generated with a PC as a terminal, emulating the first access to 500 popular websites. The collected data is divided into groups with a classical unsupervised learning algorithm, namely K-means clustering. Results show how web pages are classified into six groups and their cluster characteristics.