Combining Linked data and Data Mining Techniques to improve Clustering of Large Scale Media Repositories: A case study with BBC

Ross Fenning
British Broadcasting Corporation (BBC)
MediaCityUK
Salford, United Kingdom
ross.fenning@bbc.co.uk

Dhavalkumar Thakker
University of Bradford
Bradford
United Kingdom
dhavalkumar.thakker@gmail.com

Tejal Shah
UNSW
Australia
tejals@cse.unsw.edu.au

ABSTRACT

Media companies produce ever larger numbers of articles, videos, podcasts, games, commonly collectively known as "content". A successful content producing organisation not only has to develop systems to aid producing and publishing content, but there are also demands to engineer effective mechanisms to aid consumers in finding that content. Linked Data technologies provide data enrichment beyond attributes and keywords explicitly available within content data or metadata. In our work we experiment with all the plausible mapping of an RDF graph representing a content item to an attribute set suitable for data mining. With such a mapping, we have explored the application of machine learning, particularly unsupervised learning, across an organisation's whole content corpus. We have created an innovative pipeline of linked data and data mining using Apache Spark that allows clustering of media content items on the fly. We have evaluated such system in the context of British Broadcasting System (BBC) and present the optimum combination of RDF graphs and data mining with qualitative and quantitative results.

Keywords

Semantics; Linked Data; Semantic Web; Machine Learning; Clustering; Data Mining; RDF Graph; Media repositories; Content Management