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 a 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 that allows clustering of media content items on the fly. We have evaluated such system in the context of website content produced by the British Broadcasting Corporation (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

1. INTRODUCTION

Media companies produce ever larger numbers of articles, videos, podcasts, games, etc. – commonly collectively known as "content". A successful content-producing website not only has to develop systems to aid producing and publishing that content, but there are also demands to engineer effective mechanisms to aid consumers in finding that content.

Approaches to aid discovery of content used in industry include providing a text-based search, hierarchical categorisation (and thus navigation thereof) and even more tailored recommended content based on past behaviour or content enjoyed by friends (or sometimes simply other consumers who share your preferences).

1.1 Problem

In order to build systems that operate across the full corpus of their content, organisations face several problems:

- Large organisations can have content across multiple content management systems, in differing formats and data models.
- Many content items are in fairly opaque formats, e.g. video content may be stored as audio-visual binary data with minimal, textual metadata to display on a containing web page.
- Content is being published continuously, which means any search or discovery system needs to keep up with content as it is published and process it into the appropriate data structures. Any analytical process that operates over all content (e.g. machine learning) may need to be run periodically or in an incremental fashion.

1.2 Hypothesis

The following hypotheses are proposed for gaining new insights about an organisation's diverse corpus of content:

- Research and software tools around the concept of *Linked Data* can aid us in rapidly acquiring a broad view (perhaps at the expense of depth) of an organisation's content whilst also providing a platform for simple enrichment of that content's metadata.
- We can establish at least a naïve mapping of an RDF graph representing a content item to an attribute set suitable for data mining. With such a mapping, we can explore applying machine learning across an organisation's whole content corpus.
- Linked Data and Semantic Web *ontologies* and models available can provide data enrichment beyond attributes and keywords explicitly avaiable within content data or metadata.

- Many content-producers currently enrich their web pages with small amounts of semantic metadata to provide better presentation of that content as it is shared on social media. This enables simple collection of a full breadth of content with significantly less effort than direct integration with content management systems (i.e. Enterprise Integration).

2. BACKGROUND

3. DESIGN

4. IMPLEMENTATION

5. ANALYSIS

6. EVALUATION

7. CONCLUSIONS