Building a Multimedia Web Observatory Platform

Jonathon S. Hare, David P. Dupplaw, Wendy Hall, Paul H. Lewis, and Kirk Martinez

Electronics and Computer Science,
University of Southampton, Southampton, UK
{jsh2,dpd,wh,phl,km}@ecs.soton.ac.uk

Abstract. The data contained within the web is inherently multimedia; consisting of a rich mix of textual, visual and audio modalities. Prospective Web Observatories need to take this into account from the ground up. This paper explores some uses for the automatic analysis of multimedia data within a Web Observatory, and describes a potential platform for an extensible and scalable multimedia Web Observatory.

1 Introduction

The web is inherently multimedia (MM) in nature, and contains data and information in many different audio, visual and textual forms. To fully understand the nature of the web and the information contained within it, it is necessary to harness all modalities of data. More generally in the context of Web Science, multimedia not only gives us a more complete picture of the Web but also gives us a more complete picture of society. With this in mind, Web Observatories need to incorporate MM data from the ground-up, from harvesting through to analysis. This paper describes some of the needs and opportunities for harvesting and analysing MM data within a Web Observatory, and describes the approach of the EU ARCOMEM project[1].

2 Example uses for multimodal multimedia analysis

There are many specific use-cases for the combined analysis of different modalities of web-data. Some of these are described below:

Understanding meaning of content. Fully understanding the meaning of the content of a web-resource is challenging, with intricacies due to context and culture. Tools for the automatic analysis of meaning is something that target end-users of web-observatories would like to have available. Automatic prediction of aspects of sentiment, attractiveness and privacy all benefit from combined analysis of textual and MM data.

Interlinking content through MM features. MM analysis can be used to determine whether two resources are the same, similar, or related. Detection of near-duplicates using MM analysis can interlink content across languages, domains and groups. By exploiting these relationships it becomes possible to interlink sets of resources from the

[1] http://www.arcomem.eu

web and social web, from disjoint producers and user networks. In the case of social media analysis, this could be used to identify all users that are discussing the same event/activity/object thereby identifying new groups within the social network. Other use-cases are in areas like business or political analytics, where an organisation or entity wants to measure something about their presence on the web; in these cases, MM analysis can help link and expose relevant resources based on techniques such as face recognition or logo detection.

Mining entities, trends, topics and events. Entities, events, topics and trends manifest themselves in many ways on the web. References to entities such as people occur in all forms of web-data. Trends can be detected by analysing the text of tweets on Twitter, but equally, trends can be detected from the MM data increasingly contained within tweets. The results of MM analysis can complement pure text analysis, and equally detect *things* that cannot be determined from text analysis alone.

3 The ARCOMEM approach

Within the ARCOMEM project, we are building a system for crawling and analysing samples of web and social-web data at scale. Whilst the project is ostensibly about issues related to web-archiving, the software has features that make it ideal for use as a platform for a scalable Multimedia Web Observatory.

Fundamentally, the software has four goals; firstly it provides a way to intelligently harvest data from the web and social web around specific entities, topics, and events. Secondly, it provides a scalable and extensible platform for analysing harvested datasets, and includes modules for state-of-the-art MM (textual, visual and audio) content analysis. Thirdly, it exposes the results of the analysis in the form of a knowledge base that is interlinked with standard linked-data resources and accessible using standard semantic technologies. Finally the software provides the ability to export the harvested and analysed dataset in standardised formats for preservation and exchange. In terms of scalability, the ARCOMEM software is designed to work with small (tens/hundreds of gigabytes) to medium (multi-terabyte) web datasets.

3.1 Intelligently harvesting and sampling the Web.

ARCOMEM has developed web-crawlers specifically designed for harvesting MM data from both standard web-pages (the Large Scale Crawler) as well as social media sources via their APIs (the API crawler). In addition the web crawler is able to crawl the *deep web* through a set of modules that know how to extract information from certain kinds of web-site. Both crawlers run in a distributed fashion across a cluster of machines and store the raw content in an HBase[2] column-oriented database.

The crawlers can operate in a number of different ways, depending on the requirements of the user who defines what they want in their dataset. There are three common crawling strategies that can be used individually or together:

[2] `http://hbase.apache.org`

Standard crawling. A standard web-crawl starts with a seed list of URLs, and crawls outwards from the seed pages by following the outlinks. Constraints might be added to limit the number of hops the crawler is allowed to make from a seed page, or to limit the crawler to specific internet domains or IP addresses.

API-directed crawling. In an API-directed crawl, the user provides keywords that describe the domain of the dataset they want to create. These keywords are then fed to the search APIs of common social media sources (e.g. Twitter, Facebook, YouTube, etc), and the returned posts are examined for outlinks that are then used as seeds for a web crawl.

Intelligent crawling. In an intelligent crawl, the user provides a detailed *intelligent crawl specification* (ICS) consisting of keywords, topics, events and entities that describe the target domain. A standard crawl and/or api-directed crawl is then started, and as new resources are harvested they are scored against the ICS (using pattern matching and machine learning techniques). Scores for the outlinks of each resource are then created (combining the resource score with specific scores computed based on the link) and these scores are fed back to the Large Scale Crawler, which prioritises the next URL to crawl based on the score. URLs with low scores will not be crawled.

3.2 Advanced, scalable multimodal multimedia content analysis.

Once the data has been harvested it must be processed and analysed in order to allow the end-users to explore and query the data. The ARCOMEM system is designed such that most processing occurs as scalable, distributed Map-Reduce tasks using Hadoop[3] performed over the HBase database. The final output of these processes stored as RDF triples in a knowledge base which is queryable through SPARQL. Interfaces are built on top to allow the end users to explore the dataset without the need to know SPARQL.

The content analysis modules, provided by GATE[4] [1] and OpenIMAJ[5] [2], are primarily based around the detection of Entities, Topics, Opinions and Events (ETOEs) in both textual and visual media. Additional information such as textual terms (words with high TF-IDF values) and near duplicates of images and videos are also recorded. In visual media specifically, face (and object) detection and recognition techniques are used to detect Entities. Visual opinion analysis takes two forms; facial expression analysis, and global sentiment/attractiveness/privacy classification (e.g. [4]). High-level semantic enrichment is used to disambiguate entities and semantically link them to concepts in standard external knowledge bases such as DBpedia.

Temporal aspects. The system architecture is specifically designed to allow resources to be crawled multiple times if required. This allows changes over time to be detected and analysed.

[3] `http://hadoop.apache.org`
[4] `http://www.gate.ac.uk`
[5] `http://www.openimaj.org`

3.3 Interoperability, reusability and provenance.

The web is very dynamic and constantly changing, so it is unlikely that any dataset harvested by a Web-Observatory could ever be re-collected from scratch and end up the same. That being said, it is important that any given dataset has associated provenance that describes the exact strategy that was used to create it. In ARCOMEM this is achieved by storing the crawl specification and system configuration in the knowledge base along with information about exactly what was harvested and when.

ARCOMEM allows data to be exported in two forms that can be used together. The raw resource data can be exported as standard ISO 28500 WARC (Web ARChive) files [3], and can be used with standard tools for handling WARC files such as the Wayback Machine (which allows the dataset to be visually reconstructed and explored). The results of the processing and analysis can be exported directly in RDF. Whilst the ARCOMEM system uses its own ontology for describing the analysis, for interoperability the ontology is provided with mappings to several other standard ontologies.

4 Conclusions and Outlook

The web is inherently multimedia in nature, and this must be taken into account in the design of a Web Observatory platform, both from the point of view of dataset collection and in terms of dataset analysis. In ARCOMEM we are building an exemplar platform for harvesting and analysing the MM web. Our experiments using the ARCOMEM platform show that analysis of multimedia content facilitates enhanced interlinking and understanding of the content of the web that text-analysis alone cannot provide. Looking forwards, it seems that the technologies and techniques developed in ARCOMEM could provide a good foundation for a scalable multimedia web observatory.

Acknowledgements

This work was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 270239 (ARCOMEM).

References

1. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE (Version 6) (2011), http://tinyurl.com/gatebook
2. Hare, J.S., Samangooei, S., Dupplaw, D.P.: OpenIMAJ and ImageTerrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In: Proceedings of ACM Multimedia 2011. pp. 691–694. MM '11, ACM (2011)
3. ISO TC 46/SC 4: Information and documentation – WARC file format (ISO 28500:2009). ISO (2009)
4. Zerr, S., Siersdorfer, S., Hare, J., Demidova, E.: Privacy-aware image classification and search. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp. 35–44. SIGIR '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2348283.2348292