Enhancing Web intelligence with the content of online video fragments

Lyndon Nixon[1], Matthias Bauer[1], and Arno Scharl[12]

[1] MODUL University, Vienna, Austria
lyndon.nixon@modul.ac.at
[2] webLyzard technology gmbh, Vienna, Austria
scharl@webLyzard.com

Abstract. This demo will show work to enhance a Web intelligence platform which crawls and analyses online news and social media content about climate change topics to uncover sentiment and opinions around those topics over time to also incorporate the content within non-textual media, in our case YouTube videos. YouTube contains a lot of organisational and individual opinion about climate change which currently can not be taken into account by the platforms sentiment and opinion mining technology. We describe the approach taken to extract and include the content of YouTube videos and why we believe this can lead to improved Web intelligence capabilities.

1 Introduction

Web intelligence refers to use of technologies to extract knowledge from data on the Web, in particular, the learning of how opinions or sentiment towards specific topics or concepts change over time by analyzing the content of time-specific Web data such as activity streams on social media platforms, news stories from trustworthy newsfeeds and press releases from relevant organisations. The webLyzard Web intelligence platform [3], which has been in development since many years of university R&D, collects and analyzes big data repositories gathered from a range of electronic sources and uses state-of-the-art Web intelligence tools to reveal flows of relevant information between stakeholders through trend analyses, benchmarks and customized reports. One key domain to which the platform has been applied is climate change [1], and the insights provided by webLyzard are being used by the NOAA in the US to inform their online communication.

A public demonstrator "Media Watch on Climate Change" is available at http://www.ecoresearch.net/climate. This demonstrator analyses news articles from British and US news sources, social media and the (RSS) PR feeds of Fortune 1000 companies. For example, for any searched term, the frequency of mentions of the term over a selected time period can be seen, the level of positive or negative sentiment expressed around that term, and extent of disagreement across sources. The individual sources can be explored and the content displayed,

[3] http://webLyzard.com

e.g. text of a news article or company press release, or in social media a tweet or a YouTube video. While mentions of terms within the textual sources is being used to provide deep analytics on frequency, sentiment and disagreement over time for that term, any use of the same term within the YouTube videos which are crawled continually by the platform is disregarded, as details of the video content is not available to the internal analytics tools of the platform.

In the MediaMixer project `http://mediamixer.eu`, whose goal was to promote innovative media technology supporting fragments and semantics to industry use cases [2], a collaboration with webLyzard led to a prototype platform where the content of fragments of crawled YouTube videos could be exposed to the platforms analytics capabilities and hence video fragments could be made available to platform search and data visualisation components. This demo paper describes the approach taken and the resulting implementation (under the name v̈ideoLyzard)̈[4] as well as how we believe this work can help lead to improved Web intelligence capabilities for stakeholders such as the NOAA.

2 Technical process and workflow

A new server-side processing pipeline has been created which takes a batch of YouTube video URLs from the webLyzard crawl component and processes them by getting transcripts of each YouTube video, performing Named Entity Recognition (NER) over the transcripts, and generating on that basis an annotation for each YouTube video which identifies the temporal fragments of the video and the entities which occur in each fragment. These annotations are exported back into the webLyzard platform and on that basis access to video fragments matching search terms is made possible. videoLyzard makes use of different semantic and multimedia technologies to:

- split videos into distinct temporal fragments (generally corresponding with the sentence level in the accompanying speech), using the Media Fragment URI specification to refer to the fragments by URL [5]
- extract distinct entities from the textual transcript of the video, using the aggregation of Named Entity Recognition, or NER, Web services called NERD [6], and attaching entity annotations to a temporal fragment of the video
- normalizing entity identifications to DBPedia URIs, thus using Linked Data to provide a Web-wide unique identification for each concept, disambiguating terms which are ambiguous in natural langauge and connecting annotations to additional metadata about each entity
- create machine-processable annotations of the video in RDF, using the LinkedTV Ontology [7] which follows the Open Annotation Model [8] with specific extensions for multimedia annotation

[4] `http://webLyzard.com/video`

[5] `http://www.w3.org/TR/media-frags/`

[6] `http://nerd.eurecom.fr`

[7] `http://linkedtv.eu/ontology`

[8] `http://www.openannotation.org/`

– enabling the computer-aided ‘semantic search’ over video at the fragment level by storing the generated RDF in a triple store (Sesame) where it can be queried using SPARQL in combination with queries to complementary Linked Data repositories.

3 Implementation and results

The public demonstrator [9] initially incorporated 297 annotated YouTube videos (the videos crawled by MediaWatch in September and October 2013). Now more YouTube videos are returned for a search (listed in the Documents tab) based on locating the search term within the video transcript, and the search can be expanded to show each video fragment containing the search term (listed in the Quotes tab). For the purpose of easily browsing through transcribed videos, a new Video Tab (cf. top right corner, Figure 1) plays back the entire video (from the Documents tab) or just the fragment which matched the search (from the Quotes tab). An administrator interface is also available to allow admins to launch new video processing batches via videoLyzard as well as monitor running processes through to the successful export of the generated video annotations. Considering the small video dataset which has already been analyzed, which is much lower than the total number of YouTube videos in webLyzards crawl index, it is noteworthy how much new relevant content can be uncovered now that the platform search is able to include video fragments in its search index. For example, the search term “hydroelectricity” in the live site returns a total of 3 YouTube videos for the period November 2013 to July 2014. On the other hand, the prototype with videoLyzard annotations finds 6 YouTube video fragment matches for the 2 month period alone, all matched against semantically similar terms (hydropower, hydro geothermal, hydro-electric) which have been normalized to the DBPedia term for hydroelectricity in the NER step.

4 Future work and conclusion

Our next step is that the number of annotated YouTube videos will be scaled up, with the goal to reach to near real time accessibility to annotated YouTube content. Based on knowledge of common terms in the specific domain (climate change), we also plan to research means to clean up YouTube’s automatic transcriptions prior to performing a more domain-specific NER over them. Since a video fragment at sentence level typically does not contain enough context for viewer understanding, we also want to explore less granular fragmentation of the videos for playback, as well as use of semantic and sentiment information attached to fragments to drive exploration of related (or opposing) fragments around a topic. In conclusion, given the growing use of audiovisual content to

[9] Available at `http://link.weblyzard.com/video-showcase` with an increasing number of annotated videos.

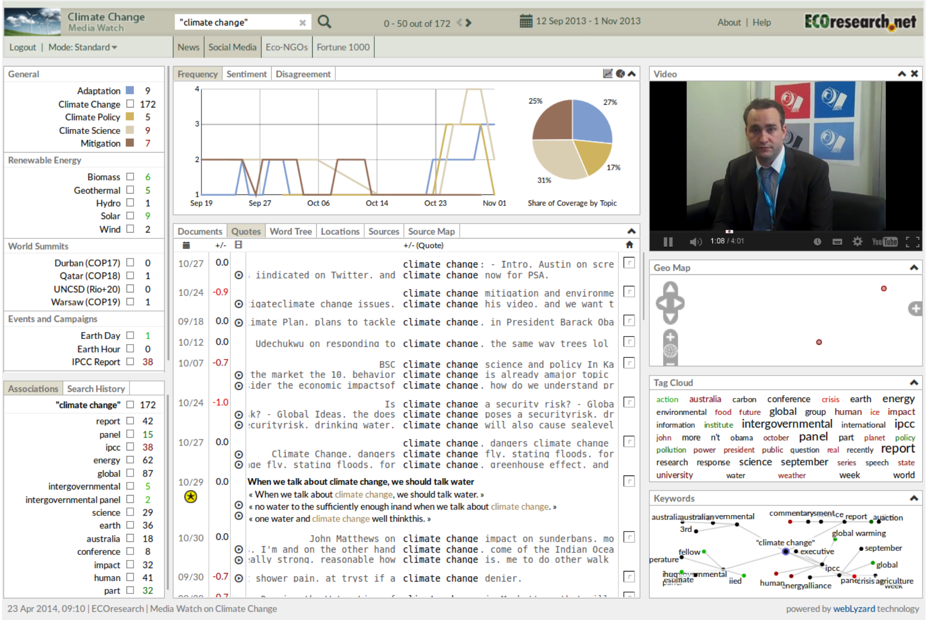

Fig. 1: Video fragment search and playback in webLyzard

communicate about topics online as opposed to just text, Web intelligence platforms miss out on significant amounts of information when they do not consider video material such as that being shared by organisations and individuals on YouTube. The videoLyzard prototype shows that even a small amount of video analysis can uncover additional intelligence for stakeholders, with semantic technologies playing a key role in associating content to distinct entities.

Acknowledgments

This work is supported by the EU FP7 funded Support Action MediaMixer (www.mediamixer.eu)

References

1. "Media Watch on Climate Change - Visual Analytics for Aggregating and Managing Environmental Knowledge from Online Sources". A Scharl, A Hubmann-Haidvogel, A Weichselbraun, H-P Lang and M Sabou. In 46th Hawaii International Conference on Systems Sciences (HICSS-46), Maui, USA, January 2013
2. "Second Demonstrators", L Nixon et al., MediaMixer Deliverable D2.2.3, April 2014