DC Proposal: Enriching Unstructured Media Content About Events to Enable Semi-Automated Summaries, Compilations, and Improved Search by Leveraging Social Networks

Thomas Steiner*

Universitat Politècnica de Catalunya
Department LSI
08034 Barcelona, Spain
tsteiner@lsi.upc.edu
–
Advisors:
Joaquim Gabarró Vallés (UPC)
Michael Hausenblas (DERI)

Abstract. Mobile devices like smartphones together with social networks enable people to generate, share, and consume enormous amounts of media content. Common search operations, for example searching for a music clip based on artist name and song title on video platforms such as YouTube, can be achieved both based on potentially shallow human-generated metadata, or based on more profound content analysis, driven by Optical Character Recognition (OCR) or Automatic Speech Recognition (ASR). However, more advanced use cases, such as summaries or compilations of several pieces of media content covering a certain event, are hard, if not impossible to fulfill at large scale. One example of such event can be a keynote speech held at a conference, where, given a stable network connection, media content is published on social networks while the event is still going on.

In our thesis, we develop a framework for media content processing, leveraging social networks, utilizing the Web of Data and fine-grained media content addressing schemes like Media Fragments URIs to provide a scalable and sophisticated solution to realize the above use cases: media content summaries and compilations. We evaluate our approach on the entity level against social media platform APIs in conjunction with Linked (Open) Data sources, comparing the current manual approaches against our semi-automated approach. Our proposed framework can be used as an extension for existing video platforms.

Keywords: Semantic Web, Linked Data, Multimedia Semantics, Social Networks, Social Semantic Web

* Thomas Steiner is partly funded by the European Commission under Grant No. 248296 for the FP7 I-SEARCH project.

1 Introduction

Official statistics [15] from YouTube, one of the biggest online video platforms, state that more than 13 million hours of video were uploaded during 2010, and 35 hours of video are uploaded every minute. The mostly text-based video search engine behind YouTube works mainly based on textual descriptions, video titles, or user tags, but does not take semantics into account: it does not get the meaning of a video, for example whether a video tagged with "obama" is about *the* Obama, or about a person that just happens to have the same name. We speak of the semantic gap in this context. In [11], Smeulders et al. define the semantic gap as *"The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation"*. Our thesis presents an approach for bridging the semantic gap for media content published through social networks by adding proper semantics to it, allowing for summaries, compilations, and improved search.

The remainder of this paper is structured as follows: Section 2 lists related work, Section 3 presents the proposed approach, Section 4 provides an evaluation of our work so far and gives an outlook on the evaluation plan for future work, which is detailed in Section 5. We conclude this paper with Section 6.

2 Related Work

We refer to previous work in a number of related areas, including enriching unstructured media content, event illustration, and video summarization.

The W3C Ontology for Media Resources [7] defines a set of mappings for many existing metadata formats. It aims to foster the interoperability among various kinds of metadata formats currently used to describe media resources on the Web. Time-based semantic annotations are possible using the `relation` property by linking to an RDF file or named graph containing annotations for a media resource (or fragment [14]), but currently there is no solution for embedding a set of RDF triples directly into one of the properties of the ontology. Similarly, the MPEG-7 [5] standard deals with the storage of metadata in XML format in order to describe multimedia content. Several works like for example [1] by Celma et al. have already pointed out the lack of formal semantics of the standard that could extend the traditional text descriptions into machine understandable ones. The authors explain that semantically identical metadata can be represented in multiple ways. Efforts [3] have been made to translate MPEG-7 into an ontology to enhance interoperability. These efforts, however, did not gain the traction their authors had hoped for.

In [8], Liu et al. present a method combining semantic inferencing and visual analysis for automatically finding photos and videos illustrating events with the overall objective being the creation of a Web-based environment that allows users to explore and select events and associated media, and to discover connections between events, media, and people participating in events. The authors

scrape different event directories and align the therein contained event descriptions using a common RDF event ontology. The authors use Flickr and YouTube as media sharing platforms and query these sources using title, geographic coordinates, and upload or recording time. In order to increase the precision, the authors either use (title and time), or (geographic coordinates and time) as combined search queries. Liu et al. prune irrelevant media via visual analysis. While the authors of [8] start with curated event directory descriptions and the two media sources Flickr and YouTube, we focus on leveraging social networks and a broad range of media content sharing platforms specialized in live-streaming. Limiting the event scope to concerts, Kennedy et al. describe a system for synchronization and organization of user-contributed content from live music events in [6]. Using audio fingerprints, they synchronize clips from multiple contributors such that overlapping clips can be displayed simultaneously. Furthermore, they use the timing and link structure generated by the synchronization algorithm to improve the representation of the event's media content, including identifying key moments of interest. In contrast to us, Kennedy et al. focus mainly on content analysis, without revealing the origin of the considered media content, where we focus on semantically enriching media content coming from social networks.

Shaw et al. present a platform for community-supported media annotation and remix in [9], and describe how community remix statistics can be leveraged for media summarization, browsing, and editing support. While our approach is semi-automatic, their approach is manual.

3 Proposed Approach

Our objective for this thesis is to create a framework that allows for the semi-automatic generation of summaries or compilations of several pieces of media content covering a certain event and leveraging social networks. The term "event" is defined[1] by WordNet as *"something that happens at a given place and time"*. Our proposed process of generating a summary or compilation for an event includes the following steps:

Event-Selection: Decide on an event that shall be summarized.

Micropost-Annotation: Find relevant microposts on social networks containing links to media content about the event, and annotate these microposts one by one using RDF, leveraging data from the LOD cloud[2].

Media-Content-Annotation: Retrieve the pieces of media content and the accompanying metadata one by one, and annotate them using RDF, leveraging data from the LOD cloud.

Media-Content-Ranking: Rank and order the pieces of media content by relevance, creation time, duration, and other criteria.

Media-Content-Compilation: Based on the ranking, suggest a summary or compilation taking into account user constraints such as desired duration, composition (videos only, images only, combination of both), etc.

[1] `http://wordnetweb.princeton.edu/perl/webwn?s=event`

[2] `http://lod-cloud.net/`

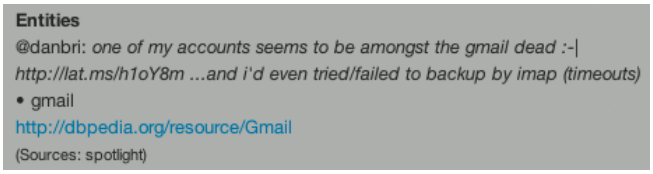
4 Evaluation

In the following we introduce our evaluation plan and present already existing evaluation for each of these steps.

Event-Selection For this task we currently have preliminary results only. We have selected two recent events, the Semantic Technology Conference 2011 (SemTech) and the Apple Worldwide Developers Conference 2011 (WWDC), both located in San Francisco. For WWDC there is a DBpedia page (`db:Apple_Worldwide_Developers_Conference`) besides the official event website (`http://apple.com/wwdc`), for SemTech there is just the event website (`http://semtech2011.semanticweb.com/`). We have disambiguated both event names with their locations using an API from previous work [12], [13]. The plaintext labels of the extracted named entities for WWDC are "Apple", "San Francisco", "Apple Inc.", "iPhone OS", and "Apple Worldwide Developers Conference". Each named entity is uniquely identified by a URI in the LOD cloud, which allows for ambiguity-free exploration of related entities, and is also indispensable in cases where no unique event hashtag is known. Having links to the LOD cloud is very important for the discovery of microposts on social networks that might have links to media content for the task *Micropost-Annotation*.

Micropost-Annotation We have implemented a generic framework based on several Natural Language Processing (NLP) Web services in parallel for the on-the-fly enrichment of social network microposts [13]. This framework has been successfully tested on overall 92 seven-day active users. The context of the tests was the detection of news trends in microposts, however, the general contribution of this framework is the extraction and disambiguation of named entities from microposts. Figure 1a shows an example. By using Google Analytics, named entity occurrences can be easily tracked over time. As an example, Figure 1b shows the occurrences graph generated by Analytics of the named entity "tsunami" (`db:tsunami`). Japan was hit by a tsunami on March 11, exactly where the peak is on the graph. In general the occurrences graphs also for other examples indeed correspond to what we would expect from the news headlines of the considered days, which implies the correct functioning of our micropost annotation framework.

Media-Content-Annotation For this task we currently have preliminary results only. At present we have implemented an interactive Ajax application called SemWebVid [12] that allows for the automatic annotation of videos on YouTube with RDF. Based on the same API that powers SemWebVid, we have implemented a command line version of the annotation mechanism that in the future can be used to batch-process videos. For now, we have annotated videos with the interactive Ajax application and reviewed the annotations manually, however, at this point, have not yet compared the annotations to a gold standard. We use the Common Tag [2] vocabulary to annotate entities in a temporal video fragment [14]. An example can be seen in Figure 2. This simple and consistent annotation scheme will make the comparison with a gold standard easier. Using Common Tag, both the video per se, as well as video fragments of the whole video can be annotated in the same way.



(a) Screenshot of a tweet and the thereof extracted named entity "gmail" with its representing DBpedia URI.

(b) Popularity of the named entity "tsunami" from March 10 - 14 in tweets. Japan was hit by a tsunami on March 11, at the peak.

Fig. 1: Tweet annotation and popularity of a named entity over time.

```
<http://www. youtube.com/watch?v=hzFp3rovfY0#t=171,177>
  a ma:MediaFragment ;
  ctag:tagged
    [ a ctag:Tag ;
      ctag:label "Commodore 64" ;
      ctag:means <http://dbpedia.org/resource/Commodore_64>
    ] .
```

Fig. 2: Annotated named entity in a video fragment.

Whole Video Annotation From our experiences so far, video annotation of the whole video works accurately. We have tested our approach with keynote and conference session videos (for examples from the Google I/O events in 2010 and 2011), political speeches (for example Obama's inauguration address), but also more underground video productions such as a video[3] about the music artist Timbaland being accused of stealing a tune from the Commodore 64 scene.

In-Video Annotation Results for the subtask of annotating in-video fragments are currently still sparse in most test cases. We found that sending smaller input texts increases the recall of the NLP Web services that we use without lowering the precision. In consequence we are now considering splitting up the to-be-analyzed data into smaller pieces, at the cost of processing time. We found the sweet spot between recall, precision, and processing time to be around 300

[3] `http://www.youtube.com/watch?v=hzFp3rovfY0`

characters. This figure was also publicly confirmed by Andraž Tori, CTO of Zemanta, at a keynote speech.

Media-Content-Ranking We have no results yet for this task. We plan to let the user tweak the ranking criteria interactively and see the effect on the ranking immediately. This could happen via sliders, where a user could change the weights of criteria like view count, duration, recency, etc. The evaluation will happen based on user feedback from a test group. We will have to test whether video genre-specific weights have to be introduced, or whether common weights across all genres already reveal satisfactory results.

Media-Content-Compilation We have no results yet for this task. Similar to the previous task *Media-Content-Ranking*, we plan to let the user adjust media composition criteria on-the-fly. For the desired duration, a slider seems adequate UI-wise, however, we have to do some experiments whether the video compilation can work fast enough to make the slider's reaction seem interactive. The same speed constraint applies to the video composition selection (videos only, images only, combination of both). Obviously the quality of the final video summaries needs to be evaluated by a test group, ideally against a gold standard of human-generated *event x in n seconds* videos. A typical example is the YouTube video `http://www.youtube.com/watch?v=skrz1JsxnKc`, which has a 60 seconds summary of Steve Jobs' keynote at WWDC. The problem, however, with these user-generated summaries is that they usually use official professionally produced video footage and not user-generated content. This allows for high quality audio and video quality, whereas user-generated content from mobile devices typically suffers from problems like noisy environments when recorded from the middle of an audience, overexposure when recorded against stage lighting, or a lack of detail when recorded from too far away. We will see in how far this is an issue once we have a working prototype of the whole framework.

5 Future Work

We present future work for each of the previously introduced steps.

Event-Selection We will keep an eye on a wide variety of events, such as concerts, conferences, political demonstrations, elections, speeches, natural disasters, festivities, but also non-publicly announced events such as private parties, always given there is enough social media coverage and media content available. We will archive social network communication produced around these events for further analysis.

Micropost-Annotation So far we have implemented a solid framework capable of annotating microposts. Future work in this task will be to further increase recall and precision by incorporating more Natural Language Processing engines both for English and non-English languages. While English is covered quite well by the existing engines, other key languages such as the so-called FIGS

languages (French, Italian, German, Spanish) are still not optimally covered. Our work here will focus on the integration and the alignment of the output formats of the various NLP services, both commercial and non-commercial. The main constraint here will be the processing time, and depending on the event the sheer amount of potentially available microposts within a short period of time (compare nation-wide elections with a private party). We will also work on improving entity consolidation and ranking algorithms when different NLP services have agreeing or contradicting results for the same input text.

Media-Content-Annotation At present we have implemented both a command line and an interactive Ajax version of the media content annotation mechanism tailored to the YouTube video platform. Future work in this task will be to improve precision and recall by the same improvement steps as in the *Micropost-Annotation* task. In addition to these steps, a major improvement will come from piece-wise rather than all-at-once analysis of the available unstructured metadata, taking into account the 300 characters sweet spot mentioned before. The constraint with this task is processing time, especially the more NLP services are involved in processing the data. Our current approach will have to be broadened to support other popular social media video and photo sharing platforms, some of them covered in [10]. In addition to that we will work to support Facebook's and Twitter's native photo and video sharing features.

Media-Content-Ranking This task has not started yet. It will consist of development efforts in order to create a testing framework for the interactive ranking and re-ranking of user-generated media content.

Media-Content-Compilation This task has not started yet. In a first step, the task consists of application development using JavaScript, HTML5, and CSS in order to generate media content compilations. We will make heavy use of the HTML5 media elements interface [4] for the `video` and `audio` elements as defined in the HTML5 specification. In a second step, an evaluation framework has to be developed in order to objectively judge the generated results. We will also investigate in how far existing third party manually generated summaries can be used as a gold standard.

6 Conclusion

We have had a look at related work from the fields of enriching unstructured media content, event illustration, and video summarization. In continuation we have introduced the required steps for our proposed approach and have evaluated our work so far, considering the tasks *Event-Selection*, *Micropost-Annotation*, *Media-Content-Annotation*, *Media-Content-Ranking*, and *Media-Content-Compilation*. Finally, we have provided an outlook on future work for each task.

Keeping in mind our objective for this thesis, we have the basic bricks in place, both for media content annotation, and for social network communication enrichment. Now we need to put the two pieces together in order to get a working

product. The main research question is *"how can semantically annotated media content linked to from semantically annotated microposts be ranked, key moments of interest be detected, and media content fragments be compiled in order to get a compelling summary?"*. We are envisioning both a stand-alone Web application where a user can select an event and get a custom-made video summary, but also Web browser extensions for existing media content sharing platforms where users can start off with one piece of media content and see its broader context.

References

1. Ò. Celma, S. Dasiopoulou, M. Hausenblas, S. Little, C. Tsinaraki, and R. Troncy. MPEG-7 and the Semantic Web, 2007. `http://www.w3.org/2005/Incubator/mmsem/XGR-mpeg7-20070814/`.
2. Common Tag. Common Tag Specification, January 11, 2009. `http://commontag.org/Specification`.
3. R. García and O. Celma. Semantic integration and retrieval of multimedia metadata. In *Proc. of the ISWC 2005 Workshop on Knowledge Markup and Semantic Annotation*, volume 185, pages 69–80, 2005.
4. I. Hickson. HTML5 W3C Editor's Draft, Media elements, October 25, 2007. `http://www.w3.org/TR/html5/video.html#media-elements`.
5. IEEE MultiMedia. MPEG-7: The Generic Multimedia Content Description Standard, Part 1. *IEEE MultiMedia*, 9:78–87, April 2002.
6. L. Kennedy and M. Naaman. Less talk, more rock: automated organization of community-contributed collections of concert videos. In *Proc. of the 18th Int. Conference on World Wide Web*, pages 311–320, New York, NY, USA, 2009. ACM.
7. W. Lee, T. Bürger, F. Sasaki, V. Malaisé, F. Stegmaier, and J. Söderberg. Ontology for Media Resource 1.0, June 2009.
8. X. Liu, R. Troncy, and B. Huet. Finding media illustrating events. In *Proc. of the 1st ACM Int. Conference on Multimedia Retrieval*, pages 58:1–58:8, New York, NY, USA, 2011. ACM.
9. R. Shaw and P. Schmitz. Community annotation and remix: a research platform and pilot deployment. In *HCM '06: Proc. of the 1st ACM Int. Workshop on Human-centered Multimedia*, pages 89–98. ACM Press, 2006.
10. Sheldon Levine. How People Currently Share Pictures On Twitter, June 2, 2011. `http://blog.sysomos.com/2011/06/02/`.
11. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:1349–1380, December 2000.
12. T. Steiner. SemWebVid - Making Video a First Class Semantic Web Citizen and a First Class Web Bourgeois. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.
13. T. Steiner, A. Brousseau, and R. Troncy. A Tweet Consumers' Look At Twitter Trends. Workshop Making Sense of Microposts at ESWC 2011, Heraklion, Crete, 30 May 2011. `http://research.hypios.com/msm2011/posters/steiner.pdf`.
14. R. Troncy, E. Mannens, S. Pfeiffer, and D. V. Deursen. Media Fragments URIs. W3C Working Draft, December 8, 2010. `http://www.w3.org/2008/WebVideo/Fragments/WD-media-fragments-spec/`.
15. YouTube.com. Official Press Traffic Statistics, June 14, 2011. `http://www.youtube.com/t/press_statistics`.