Aggregating Social Media for Enhancing Conference Experiences

Houda Khrouf, Ghislain Atemezing, Giuseppe Rizzo, Raphaël Troncy
EURECOM, Sophia Antipolis, France
<firstName.lastName>@eurecom.fr

Thomas Steiner
Universitat Politècnica de Catalunya, Barcelona, Spain
tsteiner@lsi.upc.edu

Abstract

A scientific conference is a type of event where attendees have a tremendous activity on social media platforms. Participants tweet or post longer status messages, engage in discussions with comments, share slides and other media captured during the conference. This information can be used to generate informative reports of what is happening, where (which specific room) and when (which time slot), and who are the active participants. However, this information is locked in different data silos and platforms forcing the user to monitor many different channels at the same time to fully benefit from the event. In this paper, we propose a framework named *Confomaton* that aggregates in real-time social media shared by conference attendees and aligns it with event descriptions. Developed with Semantic Web technologies, this framework enables to relive past events and to follow live conferences. A demonstrator is available at http://eventmedia.eurecom.fr/confomaton.

Introduction

Just like any other popular event, scientific conferences trigger an ever-growing amount of activities on social media. These activities consist of slides, photos and videos posted by authors, physical participants and followers, but also status messages published on social networks such as Twitter, Google+, or Facebook, and media platforms such as Flickr, YouTube, or SlideShare. The problem is that this data is largely unstructured, spread over multiple platforms, and just weakly associated with a conference event as a whole as opposed to its fine-grained sub-events. Therefore, the physical participants or the ones who try to follow the event online are forced to monitor multiple channels to fully benefit from the conference. In contrast, a scientific conference is highly structured consisting generally of workshops and tutorials, parallel sessions composed of talks, keynotes, panels, posters, and demos that all have planned schedules, topics, and allocated rooms.

There have been several attempts to model the fine-grained structure of events. For instance, the Semantic Web community publishes conference information in RDF in the so-called Semantic Web Dog Food corpus[1] (Möller et al. 2007). Lanyrd[2] exposes information about past and upcoming conferences via a REST API. The service holds a considerable number of events, ranging from large conferences such as TED[3] to smaller ones. Both services are different in terms of data models, but also from the data creation and data granularity point of view. The Semantic Web Dog Food server hosts RDF archives prepared by a small set of people who are usually the conference organizers. It provides fine-grained information detailing the set of sub-events, such as sessions, tutorials, talks, as well as referencing the publications and their authors. The Lanyrd items are created and maintained by the wisdom of the crowd but the descriptions are kept at a very general level: only some aspects related to conference are described, such as location, date, speakers, and attendees without precision about sub-events and presented papers.

Exploring the intrinsic connection between structured events and media shared on the Web has been the focus of numerous studies (Becker, Naaman, and Gravano 2010; Troncy, Malocha, and Fialho 2010; Liu, Troncy, and Huet 2011). They propose different techniques in the area of media classification, data interlinking and event detection, trying to leverage the wealth of user generated content. However, most of these works have targeted a specific social service such as Twitter or Flickr, without any guarantee that they can be valid for others services. Furthermore, the processing is performed after the completion of the events. In contrast, we propose to aggregate media items and align them with events in real-time.

In the context of *Confomaton*, we aim at creating a real-time environment that enables users to browse events and sub-events as well as their various representative media such as pictures, slides and microposts. A typical usage is to gather data about a scientific conference and investigate the added value of collecting scientific related media. A non trivial task in such application is to connect structured data with extremely noisy content, especially in the case of a major conference. In this paper, we describe a general framework composed of an event collector, a media collector and

[1] http://data.semanticweb.org
[2] http://lanyrd.com
[3] http://www.ted.com/

a real-time reconciliation module. We present a generic user interface for visualizing the social media items that have been associated with sub-events.

Framework Description

The *Confomaton* framework we propose is based on Semantic Web technologies, a key enabler for interlinking heterogeneous information coming from multiple sources. More precisely, our framework produces and consumes Linked Data. Starting from a feed of event descriptions, an item will be identified by a URI, represented with various RDF vocabularies and published following the Linked Data principles. For instance, we mint the new URI http://data.linkedevents.org/event/a8573785-4882-4265-b842-a9684803c8f9, which describes a `TalkEvent` that happened during the ISWC 2011[4] conference.

The framework is modular and composed of four main components: (i) an Event Collector that extracts events descriptions coming from either the Semantic Web Dog Food corpus or from Lanyrd feeds; (ii) a Media Collector that collects social media content and represents them in RDF using various vocabularies; (iii) a Real-Time Reconciliation Module playing the role of associating social media with sub-events and external knowledge, using named entity detection when textual data is associated to media items; (iv) a User Interface powered by an instance of the Linked Data API as a logical layer connecting all the data in the triple store with the front-end visualizations.

Event Collector

It takes as input up-to-date data from Lanyrd feeds[5] that provides conferences information serialized using a simple data model. We do not collect upcoming events since they are provided only through specific filters such as topic and place[6]. The event collector takes also as input the Semantic Web Dog Food corpus described using the SWC ontology[7]. Both types of data coming from Lanyrd and the Semantic Web Dog Food corpus are converted to RDF according to the LODE ontology[8], a minimal model that encapsulates the most useful properties for describing events. An explicit relationship between an event and its representative media (photo, slide, tweet, etc.) is realized through the `lode:illustrate` property. For describing those media, we re-use two popular vocabularies: the W3C Ontology for Media Resources[9] for photos and videos, and SIOC[10] for tweets, status updates, posts, comments and slides.

Media Collector

It has the purpose to search for event-related media items such as photos, videos and slides from various social networks and media platforms. We currently support 4 social networks (Google+, MySpace, Facebook, and Twitter) and 7 media platforms (Instagram, YouTube, Flickr, MobyPicture, img.ly, yfrog, SlideShare, and Twitpic). Our approach being agnostic of media providers, we offer a common alignment schema for all of them containing information such as the deep link of the media, the media type, the story URL, the story content, the author profile URL, the timestamp, etc. In order to retrieve data from media providers, we use the particular media provider's search Application Programming Interfaces (API) where they are available, and fall back to Web scraping the media provider's website if not. The media collector can be tested at http://webmasterapp.net/social/.

Listing 1: Sample output of the media collector showing Google+ and Flickr results using `#iswc2011` as query.

```
{
  "GooglePlus": [
    {
      "mediaurl": "http://software.ac.uk/sites/default/
              files/images/content/Bonn.jpg",
      "storyurl": "https://plus.google.com/
              107504842282779733854/posts/6ucw1Udb5NT",
      "message": {...}
    }],
    "Flickr": [
    {
      "mediaurl": "http://farm7.staticflickr.com/
              6226/6290782640_e8a1ffdcc2_o.jpg",
      "storyurl": "http://www.flickr.com/photos/
              96628098@N00/6290782640/",
      "message": {..}
    }]
 }
```

We collected social media data in real-time using the main tags advertised by the organizers of ISWC 2011 conference or provided by the Lanyrd website. Table 1 shows some statistics about the different media services used by the attendees during ISWC 2011, along with the number of items from a number of distinct users. As expected, Twitter is by far the most used service: we have been able to collect 3,390 tweets from 519 different users. A significant proportion of tweets contains hyperlinks that we have further analyzed. Hence, we extracted 384 different websites indexed by so-called URL shorteners such as Bitly found in 1,464 tweets (43% of tweets). These links represent a rich source of media as they point to various Web resources categories such as blogs, slides, photos, publications and projects.

Media Service	Items	Users
Twitter	3390 tweets	519
pic.twitter	12 photos	6
yfrog	10 photos	9
Twitpic	10 photos	6
Flickr	47 photos	6
Google+	30 posts	26
Slideshare	25 slides	20

Table 1: Media services used during ISWC 2011 conference

[4] http://iswc2011.semanticweb.org/
[5] http://api.lanyrd.com/conferences/
[6] http://lanyrd.com/blog/2010/feeds/
[7] http://data.semanticweb.org/ns/swc/swc_2009-05-09.html
[8] http://linkedevents.org/ontology/
[9] http://www.w3.org/TR/mediaont-10/
[10] http://rdfs.org/sioc/spec/

Real-Time Reconciliation Module

It aims at aligning the incoming stream of social media with their appropriate events and to interlink some descriptions with general knowledge available in the LOD cloud[11] (e.g. people and institutions descriptions). Attaching social media to fine-grained events is a challenging task. We rely on two levels of reconciliation: one uses a simple keyword and the second one exploits the description (if available) of each media item. These two levels allow to be independent from the events granularity, which differs from one provider to another.

Algorithm 1 real-time reconciliation algorithm

```
tags ← set of tags used for lookup resource from a media server
MS ← media servers
while true do
  C ← fresh data from the conference provider
  mediaList ← retrieve media item list from MS using tags
  for media ∈ mediaList do
    Intialize namedEntityList, publicationList
    namedEntityList ← extract named entities
    for namedEntity ∈ unique(namedEntityList) do
      if type of e = Person then
        publicationList ← list of publications of the author e from C
      else
        publicationList ← list of publications in which title contains e from C
        publicationList ← list of publications in which topics contain e from C
      end if
    end for
    relevantPublication ← overlap(publicationList)
    relevantEntities ← the named entities retrieving the relevantPublication
    relevantTypes ← the types of relevantEntities
    matchedEvent ← the event related to relevantPublication
  end for
end while
```

In the first level of reconciliation, we pre-process the data with two successive filters for reducing the noise: the first relies on keyword search applied to some fields such as title and tag (i.e. Twitter hashtag), while the second filters data based on temporal clues. We exploit the hashtags already provided by Lanyrd, and we manually detect those which are used during ISWC 2011 conference. The reconciliation is then ensured through a pre-configured mapping between a set of hashtags and their associated events. This task is performed when an explicit relationship between events and media is materialized by tags. Otherwise, we enhance the reconciliation by a second level using the NERD framework (Rizzo et al. 2012) to extract named entities potentially contained in the textual messages that go with social media items. We use this information to infer common features between the item and the set of events and sub-events. A feature can be the speaker's name, the title of the paper presented and its related topics. This second level is mainly targeting the ISWC 2011 conference, aiming at reconciling the fine-grained sub-events with their associated media. This approach is formalized in the algorithm described in the listing 1.

We obtain an average precision of 61% for the named-entity-based reconciliation algorithm. On the one hand, these results show a relative good performance considering the lack of useful metadata in the Dog Food corpus. Surprisingly, true alignments have been detected, mainly due the performance of named entity extractors for classifying Persons. For example, the tweet *"@pmika: RDF indexing via MapReduce, triples are grouped into documents by subjects (see paper on SemSearch from 2009), #iswc2011"* has been correctly aligned with the talk event *Effective and Efficient Entity Search in RDF data*. Actually, the class `Person` is the most efficient type (occurring 465 times) for matching publications, compared to 295 times for `Organization` and 124 for the `Technology` concept. On the other hand, most of the unmatched or false aligned tweets are due to two reasons: either there is no mention of a particular event (the tweets express a general opinion) or there is a lack of information in the Dog Food corpus (e.g. keynote speakers, *death match* event and publications related to *outrageous ideas* event).

Web User Interface

The Web User Interface (UI) is powered by the Linked Data API[12], which provides a configurable way to access RDF data using simple RESTful URIs that are translated into queries to a SPARQL endpoint. More precisely, we use the Elda[13] implementation developed by Epimorphics. Elda comes with some pre-built samples and documentation, which allows us to build the specification to leverage the connection between the back-end (data in the triple store) and the front-end (visualizations for the user). The API layer helps to associate URIs with processing logic that extract data from the SPARQL endpoint using one or more SPARQL queries and then serialize the results using the format requested by the client. A URI is used to identify a single resource whose properties are to be retrieved or to identify a set of resources, either through the structure of the URI or through query parameters. The main demo is available at http://eventmedia.eurecom.fr/confomaton reflecting the up-to-date conferences coming from Lanyrd feeds. A second demo corresponding to the archived ISWC 2011 conference is available at http://eventmedia.eurecom.fr/iswc2011.

The user interface is built around four perspectives (tabs in the UI) characterizing an event: (i) *"Where does the event take place?"*, (ii) *"What is the event about?"*, (iii) *"When does the event take place?"*, and finally (iv) *"Who are the participants of the event?"*. In addition, the UI offers full text search for these four dimensions. On the left side of the

[11] http://lod-cloud.net/

[12] http://code.google.com/p/linked-data-api/wiki/Specification

[13] http://code.google.com/p/elda

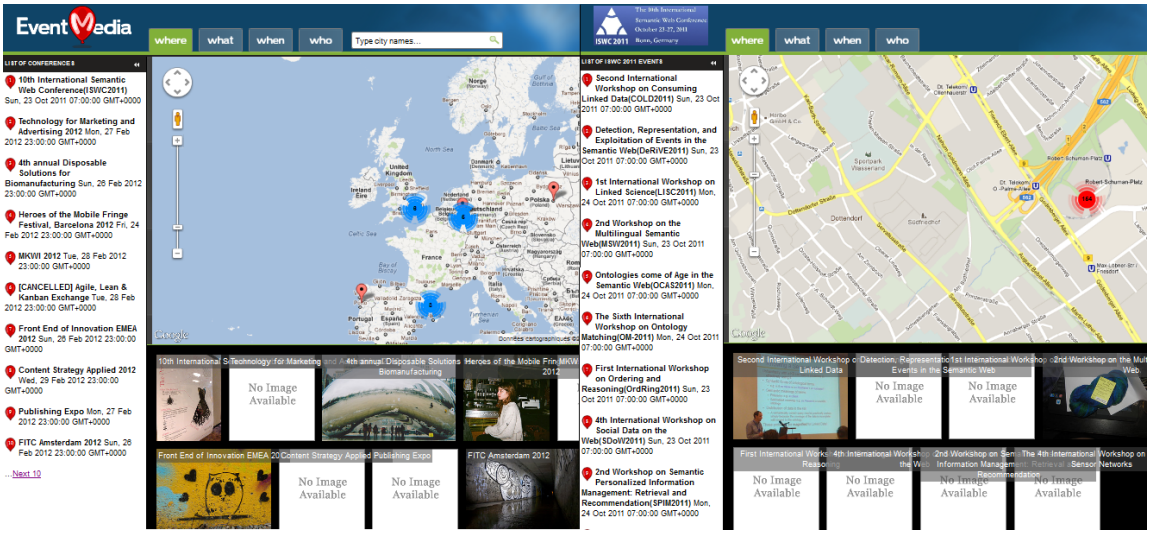

Figure 1: A showcase of *Confomaton* with Lanyrd and Dog Food data

main view, the user can select the main conference event or one of the sub-events if available as provided by the Dog Food metadata corpus. In the center, the default view is a map centered on where the event took place and the user is also encouraged to explore potential other type of events (concerts, exhibitions, sports, etc.) happening nearby, this data being provided by EventMedia (Troncy, Malocha, and Fialho 2010). The *What* tab is media-centered and allows to quickly see what illustrates a selected event (tweets, photos, slides). Zooming in an event triggers a popup window that contains the title and timetable of the event, the precise location and a slideshow gallery of all the media items collected for this event. For the *When* tab, a timeline is provided in order to filter events according to a day time period. Finally, the *Who* tab aims at showing all the participants of the conference. This is intrinsically bound to a social component, aiming not only to present relevant information about participants (their affiliations, homepages, or roles at the conference) but also the relationships between participants themselves and with events. Figure 1 shows two screenshots: one (on the left) describes up-to-date conferences located in different regions, and the second one (on the right) describes the ISWC 2011 conference.

Conclusion

In this paper, we presented a Semantic Web application, which uses the diversity of media resources generated by users that can potentially be linked with more structured metadata such as a detailed program of a scientific conference as exposed by services such as the Semantic Web Dog Food corpus and Lanyrd. We show that gathering media items from many services (Twitter, SlideShare, Flickr, Google+) in real-time while reconciling this information with an event program using pattern matching (tags) and natural language analysis (named entity) enable to provide a better conference experience including visual conference summarization or explorative search during and after the event. *Confomaton* is a framework that collects and aligns media items with events while offering a user interface that allows end-users to explore the rich semantic structure of events. We have deployed a demonstrator where past and upcoming conferences are archived. Linking media to fine-grained events is a tremendous challenge that is not yet solved. We plan to analyze more social media activities, such as the Facebook *Like* button, the *+1* button from Google+, or view counts to detect the popularity of some parts of the events and better reflect its highlights in the visual summary.

Acknowledgments

This work is partially supported by the project AAL-2009-2-049 “Adaptable Ambient Living Assistant” (ALIAS). Thomas Steiner is partially supported by the European Commission under Grant No. 248296 FP7 I-SEARCH project.

References

Becker, H.; Naaman, M.; and Gravano, L. 2010. Learning similarity metrics for event identification in social media. In *3rd ACM International Conference on Web Search and Data Mining*, 291–300.

Liu, X.; Troncy, R.; and Huet, B. 2011. Using Social Media to Identify Events. In *3rd Workshop on Social Media (WSM’11)*.

Möller, K.; Heath, T.; Handschuh, S.; and Domingue, J. 2007. Recipes for Semantic Web dog food - The ESWC and ISWC metadata projects. In *6th International Semantic Web Conference (ISWC’07)*, 802–815.

Rizzo, G.; Troncy, R.; Hellmann, S.; and Bruemmer, M. 2012. NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud. In *(WWW’12) 5th Workshop on Linked Data on the Web (LDOW’12)*, 1–10.

Troncy, R.; Malocha, B.; and Fialho, A. 2010. Linking Events with Media. In *6th International Conference on Semantic Systems (I-SEMANTICS’10)*.