Augmented Participation to Live Events through Social Network Content Enrichment

Marco Brambilla, Daniele Dell'Aglio, Emanuele Della Valle, Andrea Mauri,
Riccardo Volonterio

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico of Milano
P.za L. Da Vinci, 32. I-20133 Milano - Italy
{name.surname}@polimi.it

Abstract. During live events like conferences or exhibitions, people nowadays share their opinions, mutimedia contents, suggestions, related materials, and reports through social networking platforms, such as Twitter. However, live events also feature inherent complexity, in the sense that they comprise multiple parallel sessions or happenings (e.g., in a conference you have several sessions in different rooms). The focus of this research is to improve the experience of (local or remote) attendees, by exploiting the contents shared on the social networks. The framework gathers in realtime the tweets related to the event, analyses them and links them to the specific sub-events they refer to. Attendees have an holistic view on what is happening and where, so as to get help when deciding what sub-event to attend. To achieve its goal, the application consumes data from different data sources: Twitter, the official event schedule, plus domain specific content (for instance, in case of a computer science conference, DBLP and Google Scholar). Such data is analyzed through a combination of semantic web, crowdsourcing (e.g., by soliciting further inputs from attendees), and machine learning techniques (including NLP and NER) for building a rich content base for the event. The paradigm is shown at work on some past conferences in the CS domain (WWW 2013)

1 Introduction

During live events like conferences, exhibitions, and sports or fashion happenings, it has become common practice to share opinions, recommendations, materials, and reports through social media. Usually, the shared content refers to specific occurrences or objects related to the event, such as talks, speakers, exhibition stands, discussions, and so on. However, the mapping to such elements is often shallow or partial. This makes the social networking content an input not so valuable for the audience, especially if the social stream is very crowded and thus one has to deal with a big information overloading problem.

The problem tackled by this work is to enrich and classify the social media content related to a live event, in a way that makes it valuable for (local or remote) attendees. In particular, we focus on determining which contents are associated to which sub-event, and on enriching those contents with links to relevant entities (speakers, sessions, papers, and so on) in a domain-specific knowledge base. We then provide appropriate visualization to the enriched content, in a way that will make people able to understand what are the hot topics or sub-events and thus get guidance on what to do while attending the event.

In our approach, we select Twitter as the main social source for event-specific content. Twitter is indeed one of the most adopted platforms for social sharing, especially in the context of professional events: it can easily reach a large amount of interested people, messages are very short and require only few seconds to be shared. Furthermore, typically participants share their thoughts through event-specific hashtags, which are more or less officially related to the event itself, which makes it easy to associate them to the event.

We implement our solution in framework called ECSTASYS (Event-Centered Stream Analysis System) which combines semantic web, crowdsourcing (e.g., by soliciting further inputs by the attendees through social network invitations), natural language processing, named entity recognition and machine learning techniques for building a rich content base for the event. The application works in real time, processing the tweets as soon as they are available: in this way, attendees can have an updated and holistic view on what is happening and where, so as to get help when deciding what sub-event to attend. The application consumes data from different data sources: in addition to the afore mentioned Twitter, inputs include the official event schedule, plus domain specific content (for instance, in case of a computer science conference, DBLP and Google Scholar). The data processing determines the relevant entities described in the tweets and, consequently, the sub-events they relate to. The result of the analysis is shown to the attendees by room/sub-event, thus highlighting the interest and engagement of each sub-event, by means of appropriate user interfaces. The work is validated against a set of past conferences in the computer science field (for instance the WWW conference).

The paper is organized as follows: Section 2 describes the proposed solution, the ECSTASYS system and its components. Section 3 discusses the work done in order to apply our solution to the conference scenario. Finally, Section 4 describes possible future extensions and concludes.

2 The ECSTASYS framework

This section delves into the technical description of the ECSTASYS framework for augmented participation to live events through social network content enrichment and linking. Figure 1 provides an overview of the approach, covering both the used data sources and the main processing steps implemented by different components. In the following we describe each data source and processing component in detail.

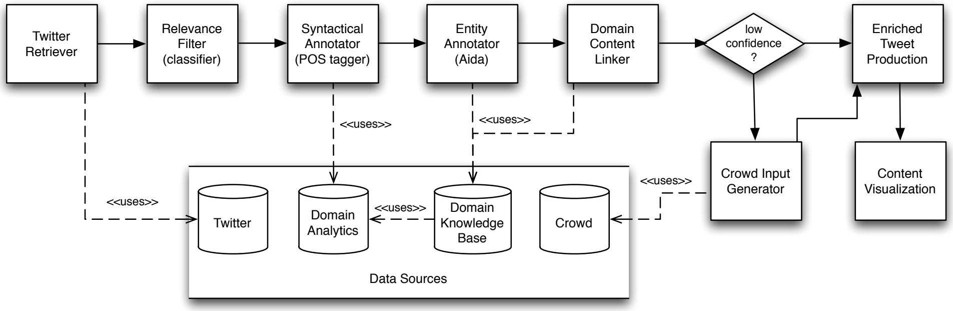


Fig. 1. Overview of the ECSTASYS framework, highlighting the processing flow and the datasources involved in each step.

2.1 Data Sources

Twitter. Twitter is the starting point of the whole approach: social feeds are retrieved by querying the Twitter stream based on hashtags, keywords, geographical locations, and people relevant to the event.

Domain Knowledge Base. The ECSTASYS knowledge base is the location on where the relevant data processed by the ECSTASYS components is stored. The knowledge base is exposed as a SPARQL endpoint and is built on the top of OpenRDF Sesame framework; as repository, we use OWLIM-Lite with the OWL 2 RL profile. Additional information about the schema and the data that it stores are provided in Section 3.

Domain Analytics. Analytics on the domain of interest are collected based on frequency of terms found in the social stream and of entities in the knowledge base. This aspect is important for reducing the impact of very frequent terms in the selected domain, which would not be considered as stop words in general sense but would actually generate noise in the specific domain. For instance, terms such as *framework*, *solution*, *Web* and so on would be too frequent in the domain of computer science conferences.

Crowd. The crowd is the source of input from human agents solicited by ECSTASYS. Typical collected information comprises confirmation of relevance of some entities for a tweet and selection of entities not automatically identified.

2.2 Processing Components

Tweet Retriever. It is the component that retrieves the tweets that are relevant for the current event from Twitter. It uses the Twitter Stream API[1] in order to connect itself to the public stream of tweets. This API allows to follow streams

[1] Cf. `https://dev.twitter.com/docs/streaming-apis`.

that match different predicates such as: users, keywords and location. All of these aspects are relevant for real world events, as they are typically identifiable by official hashtags, relevant people involved, and geographical coordinates of the venue.

Relevance Filter. The purpose of this component is to filter out the non-relevant tweets that have been extracted by the Twitter Retriever but do not provide valuable information on the event. Typical examples include: tweets written in non-English language, tweets emitted in the prescribed geographical area or containing relevant keywords but not pertaining to the event, and so on. The component immediately discards the tweets not written in English by looking at the *lang* field provided by Twitter as part of the tweet data structure. Furthermore, for selecting the relevant tweets we apply a classification approach, by exploiting a classifier based on Conditional Random Fields[10], in particular we use the *CRF++* implementation[2], trained on datasets coming from past events similar to the considered one.

Syntactical Annotator. Once the relevant tweets are selected, they are annotated through a Part Of Speech (POS) tagger. The component provides as output the annotated tweet, plus a customized set of syntactical elements extracted from the text which will be useful for the extraction of entities. Such elements consist in set of words that are good candidates for becoming named entities. On this we propose a set of heuristic solutions aimed at increasing the recall of candidate terms for the extraction of entities, as opposed to classical off-the-shelf Named Entity Extractors, which feature very high precision but also limited recall. Some examples of heuristics we apply include: generation of all the possible aggregation of contiguous nouns, contiguous nouns and adjectives, and so on.
For instance the tweet "Ingenious way to learn languages: duolingo #keynote #www2013 #gwap" is tagged in the following way:

> Ingenious‘JJ way‘NN to‘TO learn‘VB languages:‘NN duolingo‘NN #keynote‘NN #www‘NN 2013‘CD #gwap‘NN

Then the following aggregation is produced: ["way"]["languages","duolingo"]

Entity Annotator. This component processes the data produced by the Syntactical Annotator so as to determine which are the entities discussed in the text. Among the existing named entity recognition (NER) tools, we selected one based on the following requirements:

- capability of performing real-time processing of content;
- capability of linking the text items to entities in an ontology. In the recent years, several entity annotators were built on the top of open data and public knowledge bases (e.g. DBpedia and freebase) [6, 7].
- support of customization of the reference knowledge base to be used by the tool.

[2] Cf. `http://crfpp.googlecode.com/svn/trunk/doc/index.html`

The last requirement is extremely critical in our setting because usually entity annotators are only able to process generic textual content and to extract the generic entities (e.g. entities described in Wikipedia). However, in our case every event typically focuses on a very specific setting or domain, for which generic knowledge bases would contain only generic terms and very famous entities, while they would miss most of the less famous people and subjects. As an example, Dr. Jong-Deok Choi, keynote speaker at the upcoming WWW 2014 conference, does not have a page on Wikipedia (and consequently, does not appear in DBpedia).

To cope with those requirements, we decided to use AIDA [9], an open-source entity detector developed at the Max Planck Institute. It takes as input a text, it detects the set of mentions, i.e., relevant portions of the text, and associates each of them to an entity. To do it, it exploits an internal entity base and it performs two kinds of analyses: on the one hand, it selects the set of potential candidate entities for each mention; on the other hand, it performs entity-to-entity analysis to determine the coherence among the candidates. The default entity base of AIDA is built on the top of YAGO [8], but it can be customised (or replaced) with another one. In Section 3 we describe how we built our entity base out of the domain specific knowledge base of the experimental scenario.

The custom version of AIDA is wrapped in the Content Linker component: it takes as input a tweet, and it enriches it with a set of couples ($mention - entity$). The resulting tweet is pushed to the Rule-based Linker. Continuing the example introduced above, one of the mentions identified by the Syntactical Annotator is *Duolingo*; when the Entity Annotator processes the tweet, it associates the mention with the paper “Duolingo: learn a language for free while helping to translate the web” of Luis Von Ahn at the IUI 2013. As we explain above, this information comes by DBLP: we enriched the knowledge base with the recent papers of the people involved in the conference.

Domain Content Linker. This component aims at creating the relations between the tweets and the specific sub-events of the event, extracted from the official conference program (e.g., workshops, talks, sessions). As input, the component receives the tweets annotated by the Entity Annotator, i.e., a tweet with a list of related entities; as output, it enriches the tweets with the URI of the event it relates to.

This component infers two different relations: *discusses*, that indicates that a tweet talks about one of the sub-events (independently on the temporal relation between the two, i.e., the tweet could be talking about something that happened in the past or that will happen in the future); and *discusses during*, a sub-relation that states that the tweet talks about a sub-event while it is ongoing. This distinction is important for visualization purposes, as explained later.

The linkage among the tweets and the events is performed in two steps. First, the Linker retrieves the candidate events: this is done by combining the entities in the AIDA entity base that annotate the tweet, with the information in the ECSTASYS domain knowledge base. We encoded the rules that determine the candidates as continuous SPARQL queries [1] that are executed by the C-SPARQL engine; ECSTASYS runs a lifting operator of the tweet stream (from

JSON to RDF) to be able to process it. For example, one of the query is: select the *events* in which the creator of the work w is a participant, and w is an annotation of the tweet t. For instance, continuing the example, the component receives as input tweet is “Ingenious way to learn languages: duolingo #keynote #www2013 #gwap”, annotated with the paper “Duolingo: learn a language for free while helping to translate the web”. In this case, the query presented above is executed and returns all the events in which Luis Von Ahn participates.

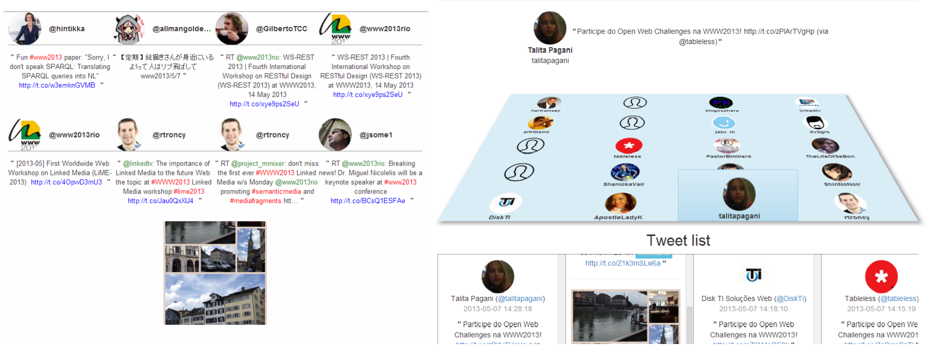
If a tweet has more than one annotation, the first step produces a set of candidate events; the second step works on it in order to derive an ordered list of candidates, associating to each of them a confidence value. The score is determined by the number of repetitions of the events in the multiset, and by their *temporal distance* to the tweet, i.e., it is more probable that a tweet discusses an event occurring temporally near. For instance, among the events on which Luis Ahn participated at the WWW 2013, the tweet was posted during the keynote, so it is the event with the highest rank in the output.

The time stamp of the tweet and the event scheduled time are also used to determine if the event can be related to the tweet through a *discusses during* relation: if the tweet is posted within 30 minutes before/after the event, the textitdiscusses during relation can hold.

Crowd Input Generator. This component is based on the CrowdSearcher framework [3, 4], which allows planning and control of crowdsourcing campaigns. The component is triggered by specific events (e.g., tweets that cannot be associated with any sub-event, or tweets for which the confidence of the association is low), and assigns them to the crowd for getting feedback. The invitation to respond is sent to people relevant to the event (e.g., the author of the tweet himself,or people who twitted about the event, or that are in the rooms of the possible sub-event).

Enriched Tweet Production. This component is a final aggregator that combines information from the crowd and the automated steps, and generates the data structure describing the enriched tweets, which can be used for any purpose.

Content Visualization. ECSTASYS provides two types of visualizations for the enriched stream of tweets, as shown in Figure 2. Both of them are web applications written in HTML5 and Javascript. The *Wall* visualization is meant to be used at the event venue on large panels (e.g., on screens or projectors in the lobby or outside the rooms of the sessions). It shows the tweets with highlighted author, mentions, hashtags and urls. Rich media content linked by the tweets is shown separately at the bottom. The *Room* visualization instead aims at personal use (e.g., on desktop browsers) and mimics the layout of a room where a sub-event is happening. It shows a 3D view of the audience (i.e., people that twitted something related to the current sub-event) in the center, with the last relevant tweet on top. The author of the tweet flips up in the audience layout. At the bottom, a continuous slider shows the tweet stream. Each tweet appears with related media, highlighted urls, mentions and hashtags.



@hintikka

" Fun #www2013 paper: "Sorry, I don't speak SPARQL: Translating SPARQL queries into NL" http://t.co/w3emknGVMB "

@allmangolde...

" 【定期】絵描きさんが身近にいるよ!って人はリプ飛ばして www2013/5/7 "

@GilbertoTCC

" RT @www2013rio: WS-REST 2013 | Fourth International Workshop on RESTful Design (WS-REST 2013) at WWW2013, 14 May 2013 http://t.co/xye9ps2SeU "

@www2013rio

" WS-REST 2013 | Fourth International Workshop on RESTful Design (WS-REST 2013) at WWW2013, 14 May 2013 http://t.co/xye9ps2SeU "

@www2013rio

" [2013-05] First Worldwide Web Workshop on Linked Media (LiME-2013) http://t.co/4OpwD3mU3 "

@rtroncy

" @linkedtv: The importance of Linked Media to the future Web: the topic at #WWW2013 Linked Media workshop #lime2013 http://t.co/Jau0QxXlJ4 "

@rtroncy

" RT @project_mmixer: don't miss the first ever #WWW2013 Linked Media w/s Monday @www2013rio promoting #semanticmedia and #mediafragments htt... "

@jsome1

" RT @www2013rio: Breaking news! Dr. Miguel Nicolelis will be a keynote speaker at #www2013 conference http://t.co/BCsQ1ESFAe "

Talita Pagani
talitapagani

"Participe do Open Web Challenges na WWW2013! http://t.co/zPlArTVgHp (via @tableless)"

http://t.co/Z1k3mSLw6a "

Talita Pagani (@talitapagani)
2013-05-07 14:28:18
" Participe do Open Web Challenges na WWW2013!

Disk TI Soluções Web (@DiskTi)
2013-05-07 14:18:10
" Participe do Open Web Challenges na WWW2013!

Tableless (@tableless)
2013-05-07 14:15:19
" Participe do Open We Challenges na WWW201

Fig. 2. *Wall* (left) and *Room* (right) visualizations of the enriched content.

3 Experimental Scenario

We apply the ECSTASYS approach to the experimental scenario of scientific conferences in the computer science domain, which are interesting complex events, with several parallel sub-events located in different rooms, typically in the same building. It follows that the precision error in geo-location would let infer wrong associations between tweets and sub-events. Moreover, people could discuss what happens in other rooms, so the geo-location is not enough to create the correct links. In the following, we provide an overview on the work devoted to the contextualization of ECSTASYS to the specific scenario of the World Wide Web conference (WWW) 2013[3].

Twitter retriever. We collected the tweets based on the hashtags of the conference (#WWW2013, #WWW), the location (i.e., the area around the conference building), and the official twitter account of the conference (@www2013rio). With these criteria, we collected more than 5000 tweets and we fed them in our experimental environment to replicate the event as if it were in real time.

Training of the classifier. The training of the classifier is built based on a dataset referring to a similar event happened in the past, namely the WWW 2012 conference. The training set comprises 500 tweets, each manually tagged as *relevant* or *not relevant*. A tweet was considered relevant if it referred to events occurring during the conference. For instance, the tweet "Sir Tim Berners-Lee @timberners_lee invented the #web about 20 years ago. Now sharing his vision at #WWW2012 in #Lyon. http://t.co/o6GRzaJP" is considered relevant, because it is about the keynote talk, while "The #www2012 wifi network supports 3000 simultaneous connections" is not, because it is a general comment regarding the conference. Our preliminary evaluation shows that the trained classifier achieves 81% precision and 97% recall when applied to the WWW 2012 content and 71% precision and 84% recall when applied to the WWW 2013 content. This shows

[3] Cf. http://www2013.org/.

that the training done on a similar past event is an acceptable starting point for solving the cold start problem of a new event.

Populating the ECSTASYS knowledge base. The knowledge base has been populated by: reusing some conference ontologies; importing the official data of the conference of interest; and importing bibliographic information about the people involved in the conference. We now report on this three aspects.

Ontology. To design the ontology for ECSTASYS knowledge base, we reused existing ontologies: *(i)* the *Semantic Web Conference Ontology*[4], currently used to describe the data stored in the Semantic Web Dog Food repository[5] and describing conferences, related sub-events (e.g., keynotes, workshops, tutorials), talks and involved people with the different roles; and *(ii)* the BOTTARI ontology [5] for describing the tweets, an extension of the SIOC vocabulary to take into account the Twitter concepts (e.g., retweets, followers and followings). We also defined as a set of custom concepts and properties to model the data produced by the Entity Annotator and the Domain Content Linker: the mentions in the tweets, their relation with the entities and, consequently, the relations between the tweets and the events they relate to.

Conference data. To describe the specific conference, we crawled the relevant information from the official Web site and we performed the lifting from HTML/XML to RDF through XSPARQL [2] (information about the WWW 2013 conference is not available as linked data). This required some manual work for setting up the crawler: in terms of effort, it costed one person day.

Bibliography. We use DBLP to enrich the ECSTASYS knowledge base with bibliographic information. We retrieved the list of the most recent papers written by each person involved in the conference (not only the authors, but also keynote speakers, organizers and chairs). This allows to enrich the keywords associated to each author while creating the AIDA entity base.

Involving the crowd. Since the events used for the experiments are located in the past, we could not involve the real crowd of participants. Therefore, the authors acted as the crowd for addressing the tasks proposed by the Crowd Input Generator. However, the governing rules have been designed and will be validated in the upcoming events.

4 Conclusions and Next Steps

In this paper we presented ECSTASYS, a system for improving the experience of conference attendees by exploiting and enriching the contents shared on the social networks. In the next months we plan three kinds of activities: evaluation (in terms of obtained precision and recall of each separate component and of the whole system); improvement of the components (e.g., a more sophisticate heuristic algorithm for the extraction of syntactical elements; and more precise crowd activation and control rules); and finally, validation of the approach during conferences and other events.

[4] Cf. `http://data.semanticweb.org/ns/swc/ontology`

[5] Cf. `http://data.semanticweb.org/`.

References

[1] D. F. Barbieri, D. Braga, S. Ceri, E. Della Valle, and M. Grossniklaus. C-sparql: a continuous query language for rdf data streams. *Int. J. Semantic Computing*, 4(1):3–25, 2010.

[2] S. Bischof, S. Decker, T. Krennwallner, N. Lopes, and A. Polleres. Mapping between rdf and xml with xsparql. *J. Data Semantics*, 1(3):147–185, 2012.

[3] A. Bozzon, M. Brambilla, and S. Ceri. Answering search queries with crowdsearcher. In *21st World Wide Web Conference (WWW 2012)*, 1009–1018, 2012.

[4] A. Bozzon, M. Brambilla, S. Ceri, and A. Mauri. Reactive crowdsourcing. In *22nd World Wide Web Conf.*, WWW '13, 153–164, 2013.

[5] I. Celino, D. Dell'Aglio, E. Della Valle, Y. Huang, T. K. il Lee, S.-H. Kim, and V. Tresp. Towards bottari: Using stream reasoning to make sense of location-based micro-posts. In *The Semantic Web: ESWC 2011 Workshops*, 80–87, 2011.

[6] M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, 249–260, 2013.

[7] A. Gangemi. A comparison of knowledge extraction tools for the semantic web. In *The Semantic Web: Semantics and Big Data, 10th International Conference (ESWC2013)*, 351–366, 2013.

[8] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.*, 194:28–61, 2013.

[9] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP2011)*, 782–792, 2011.

[10] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML '01: Proceedings of the 18th Int. Conf. on Machine Learning*, 282–289, 2001.