Building a knowledge base using Microblogs: the case of festivals and location-based events

Hoang Thi Bich Ngoc, Josiane Mothe

Université de Toulouse, UT2J, IRIT, CNRS UMR5505, Toulouse, France
{thi-bich-ngoc.hoang, josiane.mothe}@irit.fr

RÉSUMÉ. *Les médias sociaux comme twitter sont très utilisés lors d'un évènement (conférence, catastrophe, évènement culturel...) pour collaborativement commenter ou donner des avis sur son déroulement. Les utilisateurs du réseau social sont alors avertis via les personnes qu'ils suivent ou en recherchant les tweets portant sur l'évènement. Cependant compte tenu de la taille d'un tweet, l'information obtenue par un seul post est souvent très partielle. L'utilisation d'un ensemble de tweets sur un évènement peut permettre d'avoir une vue plus complète en combinant toutes les informations postées. Dans cet article, nous proposons un modèle de représentation d'une collection de microblogs basé sur une ontologie de domaine. Nous indiquons également comment populer cette ontologie en se basant à la fois sur la collection de tweets mais également sur des collections externes. Nous appliquons notre modèle au cas des tweets sur les festivals (collection issue du challenge CLEF 2016) et montrons comment il pourra être utilisé pour faire des recommendations.*

ABSTRACT. *Social media like Twitter are used during an event (catastrophe, cultural events ...) to collaboratively comment or advise on that event. Social network users are then notified through the people they follow or by seeking tweets related to the event. However, given the size of a tweet, the information obtained by a single post is often very partial. Using a set of tweets about an event makes it possible to have a more complete view by combining all the information posted. In this paper, we propose a model to represent a collection of micro-blogs based on a domain ontology. We also show how to populate this ontology based on both the collection of tweets and external collections. We apply our model to the case of tweets on festivals (a collection of the CLEF 2016 challenge) and show how it can be used to make recommendations.*

MOTS-CLÉS : *Base de connaissance de domaine, Microblog, Extraction d'information à partir de tweets.*

KEYWORDS: *Domain knowledge base, Microblog, Information extraction from tweets.*

1. Introduction

The power of social networks can be illustrated by the number of worldwide users which is expected to reach some 2.5 billion by 2018 according to Statista [1]. According to the same source Twitter is one of the leading worldwide social networks based on active users. Twitter enables users to send short 140-character messages and to follow posts from other users. Live-tweeting events such as conferences or cultural events is very popular and is basically a community that engages online while sharing topical conversations and thoughts on current experiences (Nagarajan *et al.*, 2010). During an event, some Twitter users will discuss, comment, or advise on this event while their followers will be notified. Alternativelly, it is possible for a Twitter user to search for tweets related to some content using the Search API.

However, given the 140-character size of a tweet, the information obtained by a single tweet is often very partial. It is more likely that a user rather needs to read a set of tweets to get a clear picture of an event.

For example, the two following tweets all related to Cannes 2015 provide different and complementary pieces of information :

`Vincent Lindon & Gaspar Noé, guests of honour at #VentanaSur Festival de Cannes Film Week from 30/11/15 to 6/12/15! pic.twitter.com/slPVKflt24`

`Irina Shayk, somptueuse, lors du tapis rouge du 19 mai 2015 à Cannes, pinterest.com/pin/4530340437...`

The first tweet is about the film *Carol* directed by *Todd Haynes* to be presented at the *Cannes 2015* festival. While the second tweet provides the date of a related event in Buenos Aires (*VentanaSur*) along with two actors who were there ; it is an add for the Buenos Aires festival.

From these two tweets, it is obvious that some users will lack of context to understand them individually and that some information from various tweets help understanding a given tweet. If all the pieces of information from the set of tweets were used to build a knowledge base, it would then be possible to understand better each tweet individually by enriching it using additional knowledge. For example, understanding a tweet would be easier if each entity mentioned in a tweet was explicitly associated with its entity type such as film, festival, player, ... It is possible to populate different parts of the knowledge base by using information extracted from tweets. Even if each tweet taken individually provides partial information, the sum of them could give a better picture of the information. Moreover, some parts of the knowledge base can rely on existing resources such as geographical hierarchies to model the location part for example.

[1] `http://www.statista.com/topics/1164/social-networks/`

In this paper, we propose a model to represent a collection of micro-blogs that allows to understand better information from a set of tweets on events. There are various ways to represent knowledge, but the semantic web has made ontology a very popular way to encode domain knowledge. The reason is that it enables the implementation of algorithms that include reasoning, inferring new knowledge from existing data (Margara *et al.*, 2014). In the case of the incomplete data from the tweets, an ontology can help inferring new data. This is specifically interesting when considering incomplete stream data.

Our model is based on a domain ontology which is populated using the micro-blogs as well as other Internet-based information. We focus on a specific domain which is the festivals domain that has been chosen by the CLEF 2016 challenge. This paper is centered on the domain representation and on the ontology population ; but we also mention some ways this knowledge base could be used in many applications including for localization-based recommendation.

The rest of the paper is organized as follows : Section 2 presents the related works. Section 3 details the model we suggest to represent the festival domain. Section 4 explains how the knowledge base is populated. Finally, section 6 concludes this paper and presents future works.

2. Related work

Due to the rising popularity of social media, many studies have been conducted to study ways to effectively extract information from this plentiful resource. Prior works relevant to our knowledge base are grouped into three categories : ontology-based information extraction, event detection, and location estimation in microblog.

In recent years, a number of papers have addressed the ontology-based information extraction. (Narayan *et al.*, 2010) suggest an approach to populate an ontology with the events retrieved from Twitter. Data is parsed and mined for various properties such as name, date, time, location, type and URL that are later used to populate the ontology. The authors use existing ontology of (Hobbs *et al.*, 2004) to identity Time and Alexandria Digital Library Gazetteer (1999) to recognize Location and Name. Using these ways, they omit NE when it is not explicitly mentioned in the tweet content. Our work also recognizes NE in tweets (artist, time, location, festival...) but we combine Stanford Named Entitied Recogition (NER) [2] and mining Twitter users' profile. To detect festivals in tweets, we compare string similarity of tweet content to the list of festivals accompanied by some properties such as festival names, twitter accounts, twitter hashtags and keywords retrieved from Wikipedia and tourism websites. Different from our approach, (Kontopoulos *et al.*, 2013) present the deployment of ontology-based sentiment analysis of tweets. They first identify the subject discussed in tweets and then give each tweet the sentiment score for each distinct aspect relevant to the subject.

[2] `http://stanfordnlp.github.io/CoreNLP/`

In the area of event detection, (Weng *et al.*, 2011) build signals for individual words and filter out trivial words based on their corresponding auto correlations signal. They extract events by clustering signals together using modularity-based graph partitioning. Similarly, (Zhao *et al.*, 2007) propose the text-based clustering and temporal segmentation combined with information flow-based graph. Using a different approach, (Quack *et al.*, 2008) detect local events by analyzing community photo collections while (Lee *et al.*, 2010) and (Watanabe *et al.*, 2011) analyze the geographical distribution of geotagged micro-blogs to detect events. Our work addresses the event detection in a different way. We use external resources from Wikipedia and official websites to get a list of festivals and accompanied properties retrieve tweets related to events based on the string comparison.

Finally, location estimation from microblog documents has been widely studied in recent years. NE recognition (NER) systems have addressed the problem of retrieving location specified in documents ; however they do not perform very well on informal texts (Huang *et al.*, 2015). The literature proposes some methods to improve this limitation. (Liu *et al.*, 2011) combine a K-Nearest Neighbors classifier with a linear Conditional Random Fields model under a semi-supervised learning framework to tackle the lack of information in microblog. Another location estimation approach based on analyzing geo-location by content analysis either with terms in gazetteer (Fink *et al.*, 2009), with probabilistic model (Cheng *et al.*, 2010), or users' networking (Chandra *et al.*, 2011). In our approach, we solve the problem of location identification in tweets combining two techniques : 1) using Stanford NER ; 2) extracting user hometown. These techniques complement each others in the location detection process. In the cases when location is not detected by Stanford NER, we mine Twitter user's profile to extract his hometown and consider this location as the one a tweet is about.

3. Knowledge base model : the geographical-festival ontology

Events have several dimensions, the main are :

– Location information which indicates *where* the event took, takes or will take place ;

– Temporal information that indicates *when* the event takes place ;

– Entity-related information which indicates what the event *is about* ;

In addition to their content, tweets contain additional dimensions that can be considered as meta-information and which are related to the user who posted the tweet. These dimensions can also be separated into location (from where the tweet was posted), temporal information (when the tweet has been posted) and entity-related information (who posted the tweet).

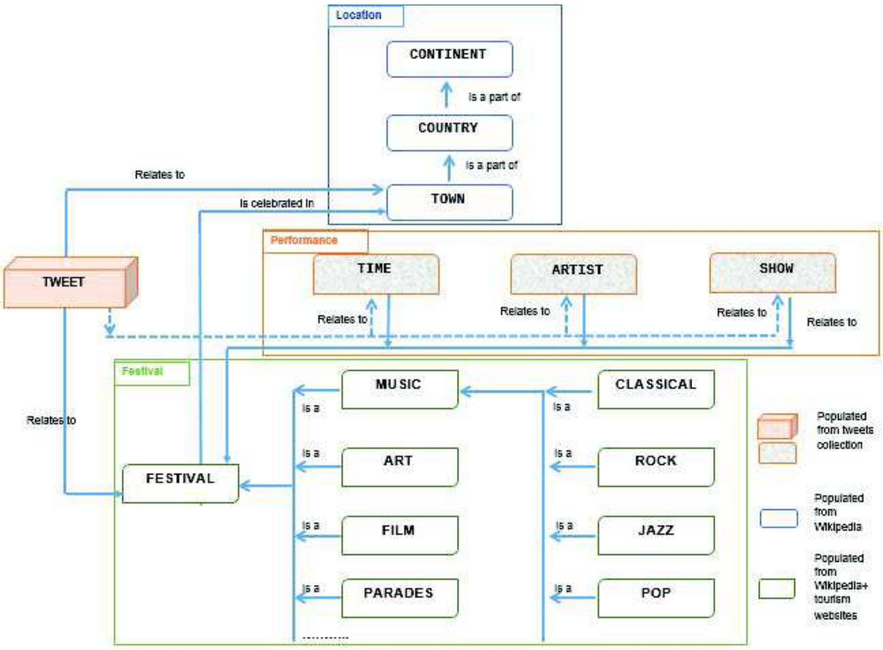

Figure 1. *Model to represent events - the case of the Festival ontology*

In the case of festival-related events, entities can be depicted in a more specific way since the types of entities can be set such as the type of festival for example, as we will see in details in the rest of this section.

Figure 1 provides the model of the knowledge base that represents the events associated to festivals. In this ontology, at this stage, we only represent the events, that is to say the tweet content, not the additional information or meta-information related to tweets such as the users.

Our geographical-festival ontology includes four sub-parts : the first part (top part of the Figure 1 - Location) represents the locations of the events, the second part (Performance) presents performance information related to each event while the third part (Festival) concerns the festivals in general. Finally, Tweet class includes the tweets related to festivals or locations. We make this splitting mainly to ease the description of the ontology.

The location part of the ontology is a hierarchy. Countries over the world are constituted in different ways, for example the United-States is divided in States, then in counties or county-equivalents, then in towns, while France is divided into regions, departments, towns. Towns can in turns be divided in parts. Considering the domain we are interested in, the town level looks appropriated as the deeper level. We then simplify the hierarchy so that it works for any part of the world. We finally kept a three level hierarchy : Town, Country, Continent, related by Is-part-of relationships. The way this part of the ontology is populated presented in section 4.2.

The second part of the ontology gathers the Performance information with three classes : Time, Artist and Show. There is not a unique resource to consider to populate this part of the ontology ; rather, it can be based both on festival or related web sites and on the tweets them-selves.

The third part of the ontology represents the Festivals. Wikipedia classifies many festivals into a set of categories and provides a hierarchy : some classes, in turn, are composed of a set of other classes. For instance, the *Music* class consists of Classical, Rock, Jazz, Pop... We re-use this set of categories to contribute to the Festival part of our ontology including a number of classes such as Music, Art, Film, Parades and so on which are types of Festivals. It might not be complete but it is appropriate to start with, and it can be completed later on, considering tweets contents.

Finally, Tweet class contains tweets relates to a festival or a specific location. From a tweet, it is possible to extract information to populate entities in the Performance part of the ontology.

4. Populating the domain ontology

In this section, we first provide the general principles of the knowledge base population (section 4.1), then we detail the various steps of the ontology population (section 4.2).

4.1. *Principles*

The domain ontology is populated considering complementary resources. As mentioned in the previous section, we use both a flow of tweets that match the information need *festival* and which can be seen as our main resource for fresh information, and external resources such as Wikipedia or web sites that contains more stable information even if they can be frequently up-dated (specifically considering future festivals).

Figure 2 depicts the overall principle of the ontology population : Wikipedia and tourism websites are used to first populate the ontology. Wikipedia provides some general information about existing locations, festival categories and even most of recognized festivals ; websites that advertise festivals provide more specific information about some festivals and some hub such as the Syndicats d'initiative web sites can also provide some additional links to other festivals. Then the domain collection consisting of tweets on festivals is used to extract additional information. The ontology and the collection are used in a process that combines the information : from the ontology, we know locations and types of festivals that help analyzing the tweets and allows us to extract new information to populate the ontology.

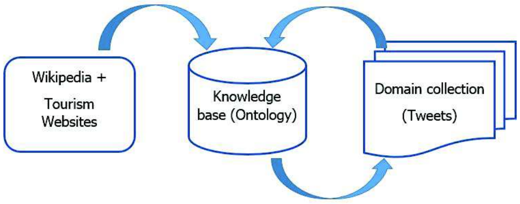
Wikipedia + Tourism Websites
Knowledge base (Ontology)
Domain collection (Tweets)

Figure 2. *The arrows show how a resource is used. Wikipedia and tourism websites are used to populate the ontology ; the ontology is used to help information extraction from the tweet collection and the additional extracted information is used to populate the ontology.*

Thus, the ontology population using Wikipedia and official websites resources can be seen as resources for background ontology population while tweet collection is a resource for providing complementary views about the events

In our approach, we choose Protégé [3] to build the ontology that implements the knowledge base and use a part of CLEF 2016 collection which includes 38,686,650 tweets about festivals in the world collected from May to October 2015.

4.2. *Detailed process*

Populating the festival ontology is achieved using the following steps :

1) The Location part of the ontology is populated using (Ngo *et al.*, 2012) results. They extract the geographic data from Wikipedia [4] which proposes adequate information of all cities by countries in the world. Data is structured in RDF format before being populated to the ontology.

2) We extract data of festivals by categories from Wikipedia and official tourism websites to populate the Festival part of the ontology. Although the information from these resources changes, the update rate from these resources is not necessarily very high to keep the ontology accurate. In our approach, we have not yet automated the festivals collecting process, but this structured information can be extracted using SPARQL on DBpedia [5].

3) The location entities from the ontology are then used to retrieve all tweets related to a specific location (we consider the tweets that explicitly mention a location and possibly tweets for which location can be inferred from the tweet content). We use Stanford (NER) to identify locations in a tweet content if locations are explicitly

[3] `http://protege.stanford.edu/` is an open-source ontology editor and framework

[4] `https://en.wikipedia.org/wiki/Lists_of_cities_by_country`

[5] `http://dbpedia.org/snorql/` : BDpedia presents structured data of Wikipedia pages

mentioned. Otherwise, we mine Twitter user profile to extract hometown and consider it as the location this tweet is about.

4) From above identified tweets, we then use the list of festivals (resulted in step 2) with some properties such as Twitter account, hashtag, keywords collected from festivals official websites to detect festival events by string comparison to tweet contents. The relationships between tweets and festivals are established in this step.

5) Finally, we automatically populate the Performance part by time, artist... extracted from tweets (results of step 3) using Standford NER.

5. Some applications

Our knowledge base model could be applied in a broad range of applications in several domains such as tourism, transportation, marketing and advertisement.

In the field of tourism, a graphical event recommender system could be developed on top of the festival ontology. Tourists would be proposed information time, famous people and related activities of events celebrated surrounding them or any specific place without spending time to search and process information. In addition, users could find latest news, opinions and feedback of attendees on tweets about events rather than browsing many tourism websites.

In the transportation domain, a system suggesting suitable route or transportation mean to avoid crowds, traffic jam or other problems linked to events could be developed using the information stored in the knowledge base for travelers.

Besides, festival events would be perfect places for companies to market their brand. They could communicate with thousands of participants and engage participants through targeted campaigns. Knowing the type of festivals, type of participants as well as the time, artists, shows ; companies could propose and implement effective advertisement campaigns for their products.

6. Conclusions and Future work

In this paper, we introduce an approach of building a knowledge base using Twitter and other external resources for the case of festivals and location-based events.

The model considers festivals organized in a specific location and related information such as time, artist or show in order to be able to provide short highly informative suggestions by analyzing tweets.

For this purpose, we define a geographical – festival ontology. As a background task, a first population of the ontology is based on various resources such as Wikipedia and official tourism websites. In addition, retrieved tweets of a specific location and events are used to populate the ontology and are analyzed to extract concerned data.

In future work, our knowledge base can be developed to be a visualized event recommendation system for users basing on their current location and other aspects such as their profile, interest and festivals that their friends participate.

7. Bibliographie

Chandra S., Khan L., Muhaya F. B., « Estimating twitter user location using social interactions–a content based approach », *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 2011.

Cheng Z., Caverlee J., Lee K., « You are where you tweet : a content-based approach to geolocating twitter users », *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010.

Fink C., Piatko C. D., Mayfield J., Finin T., Martineau J., « Geolocating Blogs from Their Textual Content. », *AAAI Spring Symposium : Social Semantic Web : Where Web 2.0 Meets Web 3.0*, 2009.

Hobbs J. R., Pan, Feng, « An ontology of time for the semantic web », *ACM Transactions on Asian Language Information Processing*, vol. 3, n^{o} 1, p. 66-85, 2004.

Huang Y., Liu Z., Nguyen P., « Location-based event search in social texts », *Computing, Networking and Communications (ICNC), 2015 International Conference on*, 2015.

Kontopoulos E., Berberidis C., Dergiades T., Bassiliades N., « Ontology-based sentiment analysis of twitter posts », *Expert systems with applications*, 2013.

Lee R., Sumiya, Kazutoshi, « Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection », *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*p. 1-10, 2010.

Liu X., Zhang S., Wei F., Zhou M., « Recognizing named entities in tweets », *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, 2011.

Margara A., Urbani J., van Harmelen F., Bal H., « Streaming the web : Reasoning over dynamic data », *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 25, p. 24-44, 2014.

Nagarajan M., Purohit H., Sheth A. P., « A Qualitative Examination of Topical Tweet and Retweet Practices. », *ICWSM*, 2010.

Narayan S., Prodanovic S., Elahi M. F., Bogart Z., « Population and Enrichment of Event Ontology using Twitter », *Information Management SPIM 2010*, 2010.

Ngo Q.-H., Doan S., Winiwarter W., « Using Wikipedia for extracting hierarchy and building geo-ontology », *International Journal of Web Information Systems*, 2012.

Quack T., Leibe B., Van Gool L., « World-scale mining of objects and events from community photo collections », *Proceedings of the 2008 international conference on Content-based image and video retrieval*p. 47-56, 2008.

Watanabe K., Ochi M., Okabe M., Onai R., « Jasmine : a real-time local-event detection system based on geolocation information propagated to microblogs », *Proceedings of the 20th ACM international conference on Information and knowledge management*p. 2541-2544, 2011.

Weng J., Lee, Bu-Sung., « Event Detection in Twitter », *ICWSM*, vol. 11, n^{o} 3, p. 401-408, 2011.

Zhao Q., Mitra P., Chen B., « Temporal and information flow based event detection from social text streams », *AAAI*, vol. 7, p. 1501-1506, 2007.