

Building a Knowledge Base using Microblogs: the Case of Cultural MicroBlog Contextualization Collection

Thi Bich Ngoc Hoang, Josiane Mothe

▶ **To cite this version:**

Thi Bich Ngoc Hoang, Josiane Mothe. Building a Knowledge Base using Microblogs: the Case of Cultural MicroBlog Contextualization Collection. Conference and Labs of the Evaluation forum (CLEF 2016), Sep 2016, Evora, Portugal. pp. 1226-1237. hal-01671370

HAL Id: hal-01671370
https://hal.science/hal-01671370

Submitted on 22 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Open Archive TOULOUSE Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/
Eprints ID : 18771

The contribution was presented at CLEF 2016 :
http://clef2016.clef-initiative.eu/

To cite this version : Hoang, Thi Bich Ngoc and Mothe, Josiane *Building a Knowledge Base using Microblogs: the Case of Cultural MicroBlog Contextualization Collection.* (2016) In: Conference and Labs of the Evaluation forum (CLEF 2016), 5 September 2016 - 8 September 2016 (Evora, Portugal).

Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

Building a Knowledge Base using Microblogs: the Case of Cultural MicroBlog Contextualization Collection

Thi-Bich-Ngoc Hoang (1)(2) and Josiane Mothe (1)(3)

(1) IRIT, UMR5505 CNRS, Université de Toulouse, Toulouse, France
(2) University of Economics, the University of Danang, Vietnam
(3) ESPE, UT2J
{thi-bich-ngoc.hoang,josiane.mothe}@irit.fr

Abstract. The Cultural MicroBlog Contextualization (CMC) Workshop provides a collection of tweets on cultural events related to festivals. Given the size of a tweet, the information obtained by a single post is often very partial. We develop the idea that using a set of tweets about an event could enable having a more complete view of that event by combining all information posted. In this paper, we propose a model to represent the collection of microblogs into a knowledge base. Considering the set of tweets on festival events from CMC, we define a domain ontology and show how to populate this ontology based not only on the tweet collection but on external data too. We detail how the knowledge base could be used to provide a complete view of an event. This paper presents the preliminary results.

Keywords: Information retrieval, Tweet analysis, Cultural MicroBlog Contextualization, Knowledge base from tweets, Microblog, Information extraction from tweets

1 Introduction

The Cultural MicroBlog Contextualization (CMC) Workshop aims at discussing applications and tools based on a collection of tweets on cultural events related to festivals [30, 31]

A tweet is composed of 140-characters and in the case of CMC, it corresponds to a twitter's post that contains the "festival" term. Such a collection can be used to analyze what has been said about a given festival, what happens during the festival, ... However, when considering a single tweet, the information is often very partial and it is more likely that a human rather needs to read a set of tweets to get a clear picture of an event.

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

> Ouverture de la route des Golden Globes avec Carol de Todd Haynes, Le fils de Saul et Mustang! A suivre! #Cannes2015 pic.twitter.com/YKd43HORmk

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. Finally the third tweet gives the information about a specific date at festival de Cannes 2015 where the model *Irina Shayk* showed up.

When considering these three individual tweets, it is obvious that some users will lack of context to understand them individually. However, some pieces of information from various tweets could help understanding a given tweet. For example, given the second tweet, if the user does not know the VentanaSur festival, he may mismatch festival de Cannes and VentanaSur festival. When considering both the second and the third tweets, he will find that festival of Cannes is in May and not at the end of the year, which was not obvious when considering the second tweet only. Each tweet taken individually provides partial information; but the sum of them could give a better picture of the information or of an event. If all pieces of information from the tweet set could be used to enrich a knowledge base, it would then be possible to understand better each tweet individually by contextualizing it using additional knowledge.

Moreover, some parts of the knowledge could rely on existing resources such as geographical hierarchies or domain knowledge rather than on tweets only. For example, understanding the second tweet would be easier if the user knew the entity types "Vincent Lindon" and "Gaspar Noé" belong to (V. Lindon is a player and G. Noé a director) and that "VentanaSur" is a "Festival".

In this paper, we propose a model to represent a collection of microblogs by a domain ontology. By combining the festival tweet collection (the CMC collection of CLEF 2016) with other Internet resources, we aim at bringing a complete picture of the collection content that can make a complete view of festival events referenced in this collection.

To populate the skeleton of the ontology, we use Wikipedia (or rather DBPedia[1]) as well as websites which provide official pieces of information about geography, list of festivals and related details. This information is quite stable in time. Next, the tweets related to each festival are selected using information retrieval methods. They are analyzed to recognize and extract named entities (NE) such as locations, artists, festival names, time. This extracted information can be used to populate instances of the corresponding classes in the ontology.

The knowledge base we design could be used in applications where the users (1) would choose a specific festival name and have a picture of that festival through the tweets (2) would choose a location and would get a list

[1] BDpedia structures the information from Wikipedia pages; it can be queried using SPARQL to extract structured information

of corresponding festivals, etc. For example, from the three tweets mentioned above, our ontology could help a user inferring from `Ouverture de la route des Golden Globes avec Carol de Todd Haynes #Cannes2015` and `Irina Shayk, somptueuse, lors du tapis rouge du 19 mai 2015 à Cannes` that the film *Carol* was presented in May 2015 at the Cannes festival.

Currently, there are various ways to represent knowledge, but we believe that ontology (e.g. OWL-based) is an appropriate and efficient solution because of the following reasons. Firstly, it makes our system an easily accessible knowledge base. The ontology-based knowledge represents data in a common language platform which can be shared and retrieved by Resource Description Framework (RDF) query language. Moreover, it allows inferring new knowledge from existing data that make users understand more about incomplete data in tweets. Finally, it could provide complete and updated information about festivals by combining Internet resources and the tweet collection.

This paper does not cover the entire project, rather it focuses on the domain representation and on the ontology population. We also mention some ways this knowledge base could be used in some applications.

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 details the model we suggest to represent the festival domain. Section 4 explains how the knowledge base is populated. Finally, section 5 concludes this paper, discusses about applications and future work.

2 Related work

Due to the rising popularity of social media, many studies propose ways to extract information from this resource. Prior works related to ours are grouped into three categories: ontology-based information extraction, event detection, and location estimation in microblogs.

2.1 Ontology-based information extraction

In recent years, a number of papers have addressed the ontology-based information extraction. Narayan *et al.* [5] suggest an approach to populate an ontology with the events retrieved from Twitter. Data is parsed and mined for various features such as name, date, time, location, type and URL that are later used to populate the ontology. The authors use the existing ontology from [25] to identify *time* and use Alexandria Digital Library Gazetteer (1999) to recognize Location and Name. Using these methods, they are not able to detect NE when it is not explicitly mentioned in a tweet content.

Kontopoulos *et al.* [6] present a method for sentiment analysis of tweets based on an ontology. They first identify the topic discussed in tweets and then give each tweet the sentiment score for each distinct aspect relevant to the topic. Another study is from [3], the authors propose an ontology-based information extraction for recognizing and semantically disambiguating NE in tweets. They solve the problem of entity disambiguation by using syntactical context and Linked Data as Freebase.

2.2 Event detection

In the area of event detection, Weng *et al.* [27] build signals for individual words and filter out trivial words based on their corresponding auto correlations signal. They extract events by clustering signals and using modularity-based graph partitioning. Similarly, Zhao *et al.* [16] propose a text-based clustering and temporal segmentation combined with information flow-based graph analysis. Besides, by aggregating information across multiple messages, Benson *et al.* [24] present a graphical model to detect entertainment events while Sakaki *et al.* [12] use a probabilistic spatio-temporal model to detect earthquake and use Kalman and particle filtering to estimate location. Using a different approach, Quack *et al.* [13] detect local events by analyzing community photo collections while Lee *et al.* [15] and Watanabe *et al.* [14] analyze the geographical distribution of geo-tagged microblogs to detect events.

2.3 Location extraction

A location is either explicitly mentioned or should be inferred from content. NE recognition (NER) systems have addressed the problem of retrieving location specified in documents; however they do not perform very well on informal texts [19]. The literature proposes some methods to improve this limitation. Liu *et al.* [29] 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 microblogs, while Krishnan *et al.* [26] propose a two-stage approach to handle non-local dependencies in NER. By aggregating information garnered from the World Wide Web to build local and global contexts from tweets, Li *et al.* [20] target the error-prone and short nature challenges. Another location estimation approach is to rely on analyzing geo-location by content analysis either with terms in gazetteer [9], with probabilistic model [7], or users' networking [8].

In the next sections, we present the knowledge base we promote as well as the way we populate it. We also present some preliminary results based on the CMC CLEF 2016 festival tweet set.

3 Knowledge base model: the geographical-festival ontology

Events have several dimensions, the main ones are:

- Location information which indicates *where* the event takes place;
- Temporal information that indicates *when* the event takes place;
- Entity-related information which indicates what the even *is about*.

In the case of festival-related events, we can have a more specific representation.

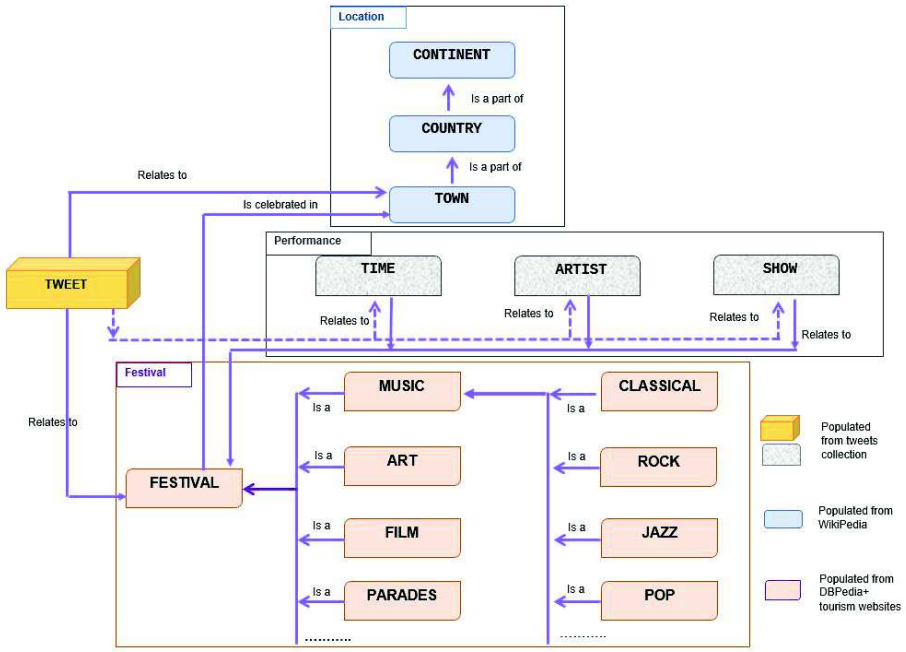


Fig. 1. Model to represent events - the case of the Festival ontology

Figure 1 depicts the model of the knowledge base that represents the events associated to festivals. The festival ontology we build includes four sub-parts. The classes and relationships between them are presented in the Figure 1. We make this splitting in four parts mainly to ease the description of the ontology. Each part of the ontology is described in details in the next paragraphs while the ontology populating is presented in the section 4.

- The first part (top part of the Figure 1 - *Location*) represents the locations of the events. 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, then towns. Towns can in turns be divided in arrondissements. Considering the domain we are interested in, the town level looks appropriated as the deeper level. We thus simplify the hierarchy so that it works for any part of the world. We finally kept a three-levels hierarchy: *Town*, *Country*, *Continent*, related by Is-part-of relationships.
- The second part (*Performance*) presents performance information related to each event; it gathers information related to each festival with three classes: *Time*, *Artist*, and *Show*.
- The third part of the ontology concerns the Festivals in general. Festivals can be classified into a set of categories that can be hierarchical. For instance, the

Music class consists of *Classical*, *Rock*, *Jazz*, *Pop*... We use a set of categories to contribute to the *Festival* part of our ontology including a number of classes such as *Music*, *Art*, *Film*, *Parades*, which are types of festivals. This hierarchy of categories is proposed by DBPedia. It might not be complete but it is appropriate to start with and it can be completed later on, considering tweets contents.

– Lastly, the *Tweet* class contains tweets which relate either to a specific festival or a location. Tweets that cannot be related to either a festival or a location are not stored and considered as useless. One tweet might be about entities from the Performance part of the ontology such as *Time*, *Artist*, *Show* ...or contain fresh information of a festival or a location such as traffic, weather, stories and feedback of attendees. We do not store this type of information in various classes but keep the tweets that can be associated to either a location, or a festival (or both) to be able to retrieve fresh information on atmosphere, twitters' comments.

4 Populating the domain ontology

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

4.1 Principles

The domain ontology is populated considering complementary resources. We use both a flow of tweets that match the information need *festival* and which can be seen as our main resource for fresh (and possibly subjective) information, and external resources such as DBPedia or tourism websites that contain more stable information even if they can be frequently up-dated (specifically considering festivals to come). Figure 2 depicts the overall principle of the ontology

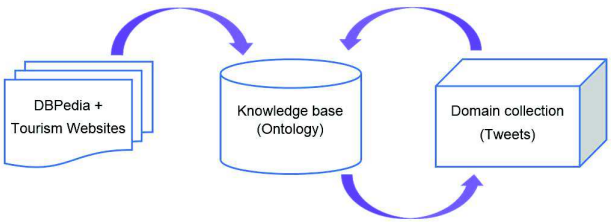


Fig. 2. The arrows show how a resource is used. DBPedia and tourism websites are used to populate the ontology; the ontology is used to help information extraction from the tweet collection and the additional extracted information is used to populate the ontology.

population: Web and DBPedia resources are used to first populate the skeleton of the ontology. DBPedia provides general information about existing locations,

festival categories and even most of well-known festivals in the world; official festival and tourism websites provide more specific information about some festivals (for example for the Jazz festival in Marciac, the official festival website can be analyzed) and some hubs such as the Syndicats d'initiative websites can also provide some additional links to other festivals.

Then the ontology and the tweet collection are used in a process that combines the information: from the ontology, we know festivals and locations that help analyzing the tweets which in turns can be used to extract new information to populate the ontology. For instance, from the ontology, it is possible to know that in *Cannes*, there is a event named *Cannes film festival*. Then, *Cannes film festival* is used to detect all tweets related to this event. These tweets, in turn, are used to extract time, artists, and shows to populate the ontology.

The ontology population using DBPedia 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.

To begin with, we chose Protege [2] to build the ontology that implements the knowledge base. We created the ontology structure as described in Figure 1 including classes such as Continent, Country, Town, Tweet, Festival.... The Location and Festival parts are to be created by data extracted from resources such as DBPedia and official websites. Then tweets related to each festival can be identified and populate the Tweet class; the relationships with *Location* and *Festival* are established in the knowledge base. In addition, information from those tweets such as *Time*, *Artist* and *Show* are extracted to populate the *Performance* part of the ontology when possible. The process will be finalized by applying inference mechanism to get new information from existing data.

In the next sections, we explain in details the populating process accompanied by preliminary results. We run the main steps of our approach on 500 tweets about Cannes and Lyon extracted from the CMC CLEF 2016 collection. This collection contains 38,686,650 tweets about festivals in the world collected from May to October 2015.

4.2 Location population

The location part of the ontology is populated using Ngo *et al.* results [21]. They extract the geographic data from Wikipedia which provides the list of locations for each countries. For example, for France it includes communes (overseas departments included) with a population over 20,000. The data is structured using 3 levels: "commune", "departement", and "region". We use the country and town ("commune") of their data to populate the ontology. There are $3,885$ instances of locations for France. Concretely, we only keep a few in our first prototype since Protege is limited in the number of instances it can handle without using a database.

[2] `http://protege.stanford.edu/` Protégé is an open-source platform for building knowledge-based ontologies.

An alternative solution for geographic data could have been to use other geographic resources such as GeoName [3] or GEOnet Names Server [4], but Wikipedia provides accurate and reliable information on this topic and was enough for our Proof-of-Concept application.

4.3 Festival population

The Festival part of the ontology is populated using the list of festivals provided by DBPedia [5]. Although the information from these resources changes, the update rate is not necessarily very high to keep the ontology accurate. This structured information can be extracted using SPARQL on locally stored DBpedia or through endpoint framework [6]. In our work, for the first implementation, we query information from DBPedia through the second way.

In addition, other information related to a festival could also be retrieved from DBPedia such as the festival location and official website. In turn, it would then be possible to collect the corresponding Twitter account, hashtags (from twitter page) and keywords about the festivals and consider them as additional properties to detect festivals in tweets as presented in the section 4.4. We keep the automation of this process for later and handle this task manually for a few festivals for now for Proof-of-Concept.

4.4 Relationship between tweets, festivals and locations

We associate tweets related to specific festivals or locations. We compare the list of festivals and properties resulting from the *Festival* population (section 4.3) with the tweet contents in order to identify all tweets related to each festival. The priority is set for festival names, twitter accounts, hashtags and keywords respectively.

When considering the sub-collection of 500 tweets, we detected 137 festivals from 137 tweets including 70 festivals detected by names, 61 festivals detected by hasgtags and 6 festivals detected by Twitter account.

To recognize locations in tweets, we combine Stanford NER with other techniques such as inferring from festival location and mining the Twitter user's profile.

We use Stanford NER to recognize locations that are explicitly mentioned in tweet contents. Because numerous twitters specify locations in their text right after a hashtag (#) Stanford NER do not extract it. For this reason, we remove all hashtags in texts before using Stanford NER. In the case locations are not specified in a tweet, we infer the location from the festival that this tweet relate to. Finally, if a tweet does not contain any text about location or festival, we

[3] `http://www.geonames.org/`

[4] `http://geonames.nga.mil/gns/html/`

[5] `http://dbpedia.org/page/Lists_of_festivals`: The root page provides festivals by categories of all countries in the world

[6] `http://dbpedia.org/snorql/`

mine the Twitter user's profile to extract the home residence. We consider this hometown as the location that his tweets are about due to a conclusion from [28]: 50% users post most of their tweets in their home residence.

We set a priority for the three location extraction techniques: Stanford NER, inference mechanism and profile mining. In case a location in a tweet is recognized by more than one method, we chose the most suitable one (detected by the highest priority technique).

Using the 500 tweets, we detected 487 locations from 409 tweets including: 1) 313 locations identified by Stanford NER in 225 tweets, 2) 137 locations for 137 tweets based on the festivals 3) 245 locations recognized by Twitter users' profile. We are currently working on more sophisticated techniques to extract location from a tweet.

4.5 Performance population

From tweets that can be related to festivals or locations (see section 4.4), we use Stanford NER to extract entities such as time, artists, shows... In the 500 tweet collection, we detected 131 artists from 103 tweets, 99 time points from 99 tweets. These instances and relationships and the corresponding tweets are stored in the ontology.

4.6 Inferring new knowledge

The inference mechanism is used to infer the relationships between instances in the case they are not directly set up from previous steps. Back to an example mentioned in the introduction part, a user can extract that festival of Cannes is in May even if the time is not mentioned in the first and second tweets. It is inferred from the third tweet. In our approach, we inferred 137 locations for 137 tweets based on the festivals that these tweets related to, 30 relationships between Festival and Artist, 19 relationships between Artist and Time classes, 55 relationships between Festivals and Time.

5 Conclusion and Discussion

In this paper, we introduced an approach for building a knowledge base using Twitter and other external resources for the case of Cultural MicroBlog Contextualization (CMC) collection.

The model considers festivals organized in a specific location and related information such as time, artists or shows. By combining the festival tweet collection with DBPedia and official websites resources, we help building a more complete picture of festivals occuring in the data collection.

For this purpose, we define a festival ontology. As a background task, the population of the location and festival parts is based on resources such as DBPedia and official websites. In addition, tweets related to specific festivals or locations are retrieved and analyzed to extract related data.

We believe that by employing ontology technology, we provide an easily accessible knowledge base system. Comparing to storing data in traditional databases, our approach has several pros. Firstly, data is presented in a common language platform which can be much easily retrieved by SPARQL. A RDF data model is also easier to updated without adverse effects to the application, thus it requires less maintenance. Secondly, the inference mechanism of ontology language allows inferring new knowledge from existing data easily (in the proof-of-concept we program the inference, but ontology allows such a process). Lastly, by combining several resources such as DBPedia, websites and Twitter, our system could bring a complete and fresh knowledge about festivals by cities in the world including official information from websites and the latest stories from Twitters.

To recognize NE in the CMC collection, we combine Standford NER with inferring technique and mining user's profiles. Applying Standford NE extraction on microblogs might not be optimal; some methods have been developed on the specific case of tweets such as [2] [1] that could be tested. We will leave this task for future work.

We would also want to extract short summaries of festivals from BDpedia or official websites to propose the users a basic idea of the festivals he selects.

We suppose that the knowledge base model we built have a broad range of applications in several domains such as tourism, transportation, marketing and advertisement.

In the field of tourism, using our knowledge base to build a graphical recommender system with highly informative summaries about events, famous people, related activities aggregated from tweets would be valuable. Tourists do not have to spend time to search and process information for their need. Moreover, latest news, opinions and feedback are more likely to appear in tweets rather than in official websites.

In the transportation domain, a system based on our knowledge base that would suggest a suitable route or transportation mean to avoid crowds, traffic jams or other problems could be welcomed by travels.

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

References

1. Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., Aswani, N. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In RANLP, (pp.83-90) (2013).
2. Ritter, A., Clark, S., Etzioni, O. Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2011).
3. Nebhi, K. Ontology-based information extraction from twitter (2012)

4. Iwanaga, I. S. M., Nguyen, T. M., Kawamura, T., Nakagawa, H., Tahara, Y., Ohsuga, A. Building an earthquake evacuation ontology from twitter. In Granular Computing (GrC), 2011 IEEE International Conference on (pp. 306-311), IEEE (2011)
5. Narayan, S., Prodanovic, S., Elahi, M. F., Bogart, Z. Population and Enrichment of Event Ontology using Twitter. Information Management SPIM (2010)
6. Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N. Ontology-based sentiment analysis of twitter posts. Expert systems with applications, (pp. 4065-4074) (2013)
7. Cheng, Z., Caverlee, J., Lee, K. You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 759-768) (2010)
8. Chandra, S., Khan, L., Muhaya, F. B. Estimating twitter user location using social interactions–a content based approach. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 838-843) (2011)
9. Fink, C., Piatko, C. D., Mayfield, J., Finin, T., Martineau, J. Geolocating Blogs from Their Textual Content. In AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0 (pp. 25-26) (2009)
10. Abel, F., Celik, I., Houben, G. J., Siehndel, P. Leveraging the semantics of tweets for adaptive faceted search on twitter. In The Semantic Web–ISWC 2011 (pp. 1-17). Springer Berlin Heidelberg (2011)
11. Wang, Z., Cui, P., Xie, L., Chen, H., Zhu, W., Yang, S. Analyzing social media via event facets. In Proceedings of the 20th ACM international conference on Multimedia (2012)
12. Sakaki, T., Okazaki, M., Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In: The 19th international conference on World wide web, ACM (2010)
13. Quack, T., Leibe, B., Van Gool, L. World-scale mining of objects and events from community photo collections. In: The 2008 international conference on Content-based image and video retrieval ,ACM (2008).
14. Watanabe, K., Ochi, M., Okabe, M., Onai, R. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In: The 20th ACM international conference on Information and knowledge management (pp. 2541-2544) (2011)
15. Lee, R., Sumiya, K. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks (pp. 1-10) (2010)
16. Zhao, Q., Mitra, P., Chen, B. Temporal and information flow based event detection from social text streams. In AAAI (Vol. 7, pp. 1501-1506) (2007)
17. Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., ... Jaimes, A. Sensing trending topics in Twitter. Multimedia, IEEE Transactions on, (pp.1268-1282) (2013)
18. Li, H., Srihari, R. K., Niu, C., Li, W. Location normalization for information extraction. In Proceedings of the 19th international conference on Computational linguistics-Volume 1 (pp. 1-7) (2002)
19. Huang, Y., Liu, Z., Nguyen, P. Location-based event search in social texts. In Computing, Networking and Communications (ICNC), 2015 International Conference on (pp. 668-672) (2015)

20. Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., Lee, B. S. Twiner: named entity recognition in targeted twitter stream. In: The 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 721-730) (2012)
21. Ngo, Q. H., Doan, S., Winiwarter, W. Using Wikipedia for extracting hierarchy and building geo-ontology. International Journal of Web Information Systems, (pp. 401-412) (2012).
22. Wimalasuriya, D. C., Dou, D. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science.(2010)
23. Nagarajan, M., Purohit, H., Sheth, A. P.A Qualitative Examination of Topical Tweet and Retweet Practices. ICWSM, pp. (295-298) (2010).
24. Benson, E., Haghighi, A., Barzilay, R. Event discovery in social media feeds. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 389-398) (2011).
25. Hobbs, J. R., Pan, F. An ontology of time for the semantic web. ACM Transactions on Asian Language Information Processing (TALIP), (pp.66-85) (2004).
26. Krishnan, V., Manning, C. D. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In: The 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 1121-1128) (2006).
27. Weng, J., Lee, B. S. Event Detection in Twitter. ICWSM, 11, 401-408 (2011).
28. Lee, B., Hwang, B. Y. A Study of the Correlation between the Spatial Attributes on Twitter. In Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on (pp. 337-340) (2012).
29. Liu, X., Zhang, S., Wei, F., Zhou, M. Recognizing named entities in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 359-367) (2011).
30. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J. Overview of the INEX 2012 tweet contextualization track. Initiative for XML Retrieval INEX, 148 (2012).
31. Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., Sanjuan, E.. Overview of the CLEF 2016 Cultural Microblog Contextualization Workshop, Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016), Lecture Notes in Computer Science (LNCS) 9822, Springer, Heidelberg, Germany, 2016
32. Ermakova, L., Mothe, J. IRIT at INEX 2012: Tweet Contextualization. In CLEF (Online Working Notes/Labs/Workshop) (2012)