Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 98 (2016) 443 – 448

Procedia Computer Science

International Workshop on Data Mining on IoT Systems (DaMIS 2016)

Real world city event extraction from Twitter data streams

Yuchao Zhou*, Suparna De, Klaus Moessner

Institute for Communication Systems (ICS), University of Surrey, Guildford GU2 7XH, United Kingdom

Abstract

The immediacy of social media messages means that it can act as a rich and timely source of real world event information. The detected events can provide a context to observations made by other city information sources such as fixed sensor installations and contribute to building 'city intelligence'. In this work, we propose a novel unsupervised method to extract real world events that may impact city services such as traffic, public transport, public safety etc., from Twitter streams. We also develop a named entity recognition model to obtain the precise location of the related events and provide a qualitative estimation of the impact of the detected events. We apply our developed approach to a real world dataset of tweets collected from the city of London.

Peer-review under responsibility of the Program Chairs

Keywords: Smart city; Twitter; city events; event extraction.

1. Introduction

Recent studies estimate 66% of the world population to be living in cities by 2050, and predict the growth of megacities, with over 10 million inhabitants each[1]. The continuous global urbanization poses several challenges to city authorities in terms of reconfiguring city services according to their needs and priorities. Being aware of city events is of significant interest to city authorities in order to plan for both known events (sports events, demonstrations etc.) as well as mitigating unforeseen circumstances (e.g. traffic incidents). In addition to information sources such as city departments, citizen sensing is being widely recognized as a complementary or corroborative information source for city events[2]. In particular, the Twitter microblogging platform has emerged as a powerful means of communication for people to share and exchange information on a wide variety of real-world events[3] and is being recognized as a near real-time city-wide event information source. Events that have an impact on human dynamics, i.e. influencing people

* Corresponding author. Tel.: +44 1483 682024; fax: +44 1483 686011.
E-mail address: yuchao.zhou@surrey.ac.uk

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.069

or traffic movements, are of particular interest to city authorities, who also want to know the impact of these events on city infrastructure[4]. Recent works on identifying events from Twitter have focused on events of a particular type, e.g. earthquakes[5], traffic[2], news[6] or general open events that may not influence city dynamics[3,7]. Our focus in this work is on identification of city related real world events. The contributions of our work are the following: 1) we develop an event classification model, reflecting the events' influence on city dynamics, 2) we propose an unsupervised method to extract city event information based on the classification model, 3) we identify the location entities in the tweets by a named entity recognition model which enables getting the precise location of the related events and 4) we provide a quantitative estimation of the impact of the event in the real world.

1.1. Background and Problem Definition

The social media platform Twitter allows users to post short text messages, or tweets, of up to 140 characters. Twitter currently has a propriety algorithm to display trending topics, consisting of terms and phrases that exhibit 'trending' behavior. While these trending topics can sometimes reflect current events (e.g. "London Olympics"), they often paraphrase conversation topics (e.g. "oneD"), with no distinctions between the different content types[3]. Due to the restricted tweet length and informal nature of social media, Twitter messages do not always follow grammar syntax and may contain mistyped words, special words, or even wrong words. Moreover, the absence of context in tweets rules out a dictionary-based method for spotting location and event terms[2].

Twitter's popularity and large user base across different cities (over 500 million users world-wide[2]) means that different types of public events get reported. This implies that a generic event extraction method that works in different cities cannot be based on simple keyword search as this may lead to only limited events being retrieved and most events being missed, especially in the absence of domain knowledge about any given city. On the other hand, techniques using event specific patterns require creating a corpus of general keywords, which will involve a huge amount of work to create a training set that will also be specific to the known and expected events of a particular city.

2. Related Work

Most of the existing approaches that detect events from tweets focus on a particular kind of event. TwitterStand[6] provides an online clustering and classification method to detect news as reported by Twitter users , however, the detected news cannot be directly linked to real world events. Sakaki et al.[5] treat a tweet as a sensor and propose a probabilistic model to detect earthquake and typhoon occurrences. The location of the event is detected through Kalman filtering and particle filtering. The work is based on the assumption that only one event will happen at a time. However, this does not apply for city events, such as traffic jams, etc., making the algorithm unsuitable for other kinds of events. Anantharam et al.[2] focus on traffic related events, providing a an automated method for creating a training set for a supervised approach. However, supervised approaches incur a heavy cost in order to be adapted to different scenarios. Balduini et al.[4] provide a city scale event detection method which links tweets with RDF data streams to support continuous query for burst detection. However, the work requires prior knowledge of the event to construct a query, making it unsuitable to detect unplanned events.

There is also some work which focuses on open domain events. Becker et al.[3] provide a combination of online clustering and classification to distinguish real world events and non-events, but the work does not provide detailed classifications nor any explanation of the detected events. Ritter et al.[7] develop an open-domain event extraction and categorization system for Twitter. The system applies an LDA-based algorithm to detect topic clusters but requires manual inspection of the clusters types.

Recent work has focused on classifying the topics of tweets into an external ontology[8]. The tweet count is used for behavior analysis and interest of visits to a cultural heritage site. In addition, Cuomo et al.[9] provide a mathematical model of visitors which shows how visitors share their experience of visiting cultural heritages on social network. Moreover, Chianese et al.[10] give a quantitative estimation of cultural heritage sensitivity by social network users. These researches in the area cultural heritage focus on users, especially visitors; analyzing on how likely a visitor is to post a message on social network and how they will interact with others, which could be adapted to our scenario of city events to establish a relationship between the number of tweets and particular types of real world events.

3. Event extraction approach

The proposed method aims to detect real-world events from a large volume of tweets. To achieve this, a classification of events is first derived, as shown in Table 1. It is based on their direct influence on city dynamics, such as on traffic flows as well as the amount of people involved. The developed classification is based on the event types discovered by Ritter et al.[7]. Twitter messages that pertain to real world events and may affect city services are extracted. However, due to some types of events using the same terms (e.g. 'perform' could be used for 'performance' as well as 'celebration'), there may be some misclassifications. Several similar event types are also subsumed into a categorization that encompasses those types, e.g. concert, festival, parade into 'culture'. Since these events will result in a similar influence on the city, it is unnecessary to classify them into separate types.

Table 1. Classification of expected real world events.

Category	Traffic influence	People involved	Examples
Traffic	High	Many	fast/slow traffic, roadwork, road incident, collision
Culture	High	Potentially many	concert, celebration, performance, exhibit, fair, festival, market, parade, firework show
Sports	Dependent on scale	Many	Sports match, race, tournament
Air quality	-	-	description of air pollution incidents
Weather	-	-	any weather description; includes wind, precipitation, temperature, cloud, sun, etc.
Disaster	-	Many	event that causes a huge damage
Non-event	-	Few	description of personal activity

To design a generic solution and avoid the need of creating a training keyword set for each city, an unsupervised method based on Twitter-LDA (Twitter Latent Dirichlet Allocation)[11] is proposed. The approach consists of the following steps:

3.1. Tweet retrieval

The tweets are retrieved from the Twitter search website API (https://twitter.com/search-home). The search statement is constrained only by place keywords and date parameters to get a complete set of tweets for a place on a certain day. An example tweet retrieved for the city of London on the 5th of February 2016 is shown below.

> Elton John performs impromptu concert in London
> rightrelevance.com/search/article pic.twitter.com/RuvZ5uScIJ

3.2. Pre-processing

Before applying TwitterLDA, the set of retrieved tweets needs to be pre-processed. This step includes tokenizing, stop words removal, and noise words removal. Tokenizing splits a sentence into tokens, which are basic elements of the sentence, such as words and punctuations. Then, stop words, which are words commonly appearing in any kind of topic in a language, are removed. This is to avoid computation effort on unimportant words. The stop words list is built from Rainball stoplist (http://www.cs.cmu.edu/~mccallum/bow/rainbow/) plus words commonly used in tweets that have little or no meaning, such as 'ha', 'hah', etc. Next, URL links and some unreadable codes are considered as noise words that will not contribute to the later steps and are removed. Thus, words are separated from the original sentence and less meaningful words are filtered out. The pre-processed tweet is shown below:

> Elton John performs impromptu concert London

3.3. Twitter LDA analysis

After pre-processing, TwitterLDA is applied on the cleaned and formatted tweets set. Twitter-LDA is a topic model designed for tweets based on LDA (Latent Dirichlet Allocation). LDA is a generative probabilistic model dealing with discrete data such as text. LDA topic model assumes a text corpus has a fixed number of topics, a document in the corpus has a mixture of topics, which form a Dirichlet distribution, and the order of words or documents does not matter. The generative process[12] of LDA model can be described as follows:

1. Randomly choose a Dirichlet distribution over topics.
2. For each word in the document
   a. Randomly choose a topic from the distribution over topics in step #1.
   b. Randomly choose a word from the corresponding distribution over the vocabulary.

In order to fit short text like tweets, Twitter-LDA model makes some modifications to normal LDA. It assumes one tweet talks about only one topic and involves a small amount of background words that do not contribute to any topic. The generative process[11] is described as follows:

1. Randomly choose a Dirichlet distribution ***b*** over background words and a Bernoulli distribution ***bt*** over decision on background words and topic words.
2. Randomly choose a Dirichlet distribution ***t*** over topical words for each topic.
3. For each user's tweet collection
   a. Randomly choose a Dirichlet distribution ***tu*** over topic.
   b. For each word
      a. Randomly choose a multinomial distribution governed by ***bt*** over decision on whether it is a background word
      b. Randomly choose a multinomial distribution governed by ***b*** over word, if the word is a background word; randomly choose a Multi distribution governed by ***t*** over word, if the word is a topical word.

Parameters of the model can be inferred through Gibbs Sampling[13], which is a Markov Chain Monte Carlo method to estimate a probability distribution. Output includes topics with keywords explanation, topic distribution for each user, and the number of tweets for each topic. During implementation, tweets collected for one day are considered as a user's tweet collection, talking about a mixture of topics. Then topics with keywords and the number of tweets for that day can be inferred through Twitter-LDA. Assuming a number of tweets talking about the same topic as in the example above (i.e. Elton John performing in London) in different ways, the output of Twitter-LDA is shown below:

john elton surprise train station piano plays performance concert watch commuters surprises crowd pancras medley hits-filled st impromptu sir station, play piano, performs pops station: leaves #eltonjohn deliver fans #music

3.4. Topic Event Labelling

Topics with top related keywords are one of outputs of Twitter-LDA. Since Twitter-LDA is an unsupervised approach, output topics are not labelled with any meaningful name. In order to link unlabelled topics to real world events, the topics are labelled based on the event type model specified in Table 1. The model is built on event types with highly related keywords for each type. A topic will be set as an event type which has the most number of keywords in the topic. The topic with no matched keywords in the event type model will be set as a non-event.

3.5. Event Scale Estimation

At the end of the Twitter-LDA process, each type of event will contain several topics, and each topic will contain a number of tweets. Thus the number of tweets talking about one type of event can be computed as well as the *Event Tweet Frequency*, which is computed by equation 1:

$$Event\ Tweet\ Frequency = \frac{No\ of\ Detected\ Event\ Tweets}{No\ of\ Sampled\ Tweets}$$

$$= \frac{Population\ Involved\ in\ Event \times Tweet\ Rate \times Event\ Influence\ Factor \times Sampling\ Rate}{Population\ at\ the\ Place \times Tweet\ Rate \times Sampling\ Rate} \quad (1)$$

where *Tweet Rate* is the percentage of people involved in an event writing a tweet, *Sampling Rate* is the rate of number of tweets sampled from the tweets collection, and *Event Influence Factor* indicates how many multiples of a tweet for an event will be posted than that in a normal situation. As *Event Influence Factor* depends on user behaviour on a city, it is set by experience. By assuming sampling is random, we can treat *Sampling Rate* is the same and divide it out. Thus we have following equation 2:

$$Event\ Tweet\ Frequency = \frac{Population\ Involved\ in\ Event \times Event\ Influence\ Factor}{Population\ at\ the\ Place} \quad (2)$$

The event scale estimation based on the population size can be calculated as given in equation 3:

$$Population\ Involved\ in\ Event = \frac{Population\ at\ the\ Place \times Event\ Tweet\ Frequency}{Event\ Influence\ Factor} \quad (3)$$

where *Population at the Place* represents the population of the city which is obtained from open data, *Event Tweet Frequency* is computed from equation 2, and *Event Influence Factor* is set by experience. The output of the algorithm for one event type (Culture) is shown below:

```
Culture || 8664.774 || john elton surprise train station
```

At the end of this step, the type of the event (*Culture* in the example above), population estimation (~8664 as shown above) and the top key terms defining the event (*john Elton surprise train station* in the example) are obtained.

3.6. Event Location Tagging

In order to find the relationships between detected anomalies and events, their location information need to be determined. Although the location information of anomalies can be obtained from sensing sites, determination of location information of the detected events is not straight-forward. This is because social media data are informal. Social media data do not always follow the grammar syntax and may contain mistyped words, special words, or even wrong words. To overcome these issues, an aggregation and rank-based location entity detection approach is proposed. For each detected event, the approach examines all the related tweets and finds the location entities in the tweets, using location named entity recognition model in OpenNLP (https://opennlp.apache.org). The detected location entities are aggregated and ranked by their occurrences. The top 2 entities are used to represent the location of the event. Also, the precise latitude and longitude can be obtained by sending a query with the top 2 entities to the Google Maps Geocoding API (https://developers.google.com/maps/documentation/geocoding/intro). The resultant output is shown below:

```
Culture || 8664.774 || john elton surprise train station ||
lat:51.5268540 || lon:-0.1245670 || London - St Pancras
```

4. **Results Visualization**

We apply our approach to identify events in the city of London. We retrieve five days (5th -9th of February 2016) of tweets that are either geo-tagged with the city of London's coordinates or mention the city in the message content. More than 30000 tweets are retrieved on a given day. Events detected on a particular day range from 44-51, with culture events being the ones most widely mentioned on Twitter. A selection of detected events for the 5th of February are visualised on a map as shown in Figure 1, depicting the event type, the population estimate, the explanatory keywords and the associated place name.

5. Conclusions

The proposed unsupervised approach is better suited to detect real world events that can inform and influence city authorities' and citizens decision making and planning, as approaches concentrating on only specific event types may not be feasible for providing a city-wide context, while open event detection approaches are not sufficient due to their lack of distinction between real world events and other non-related ordinary events. The developed LDA-based a bag

of words model can detect any topic being discussed on social media and it is supported by a keyword-based event type model to label detected topics as types of real world events. This allows non-event topics to be filtered out and enrich the explanation of the detected topics. A location detection approach has also been developed which determines the location information of related events. Moreover, our approach can also estimate the impact of the detected events according to event type, number of tweets, etc. These measurements can be aggregated based on event type and reflect the impact on the real world.

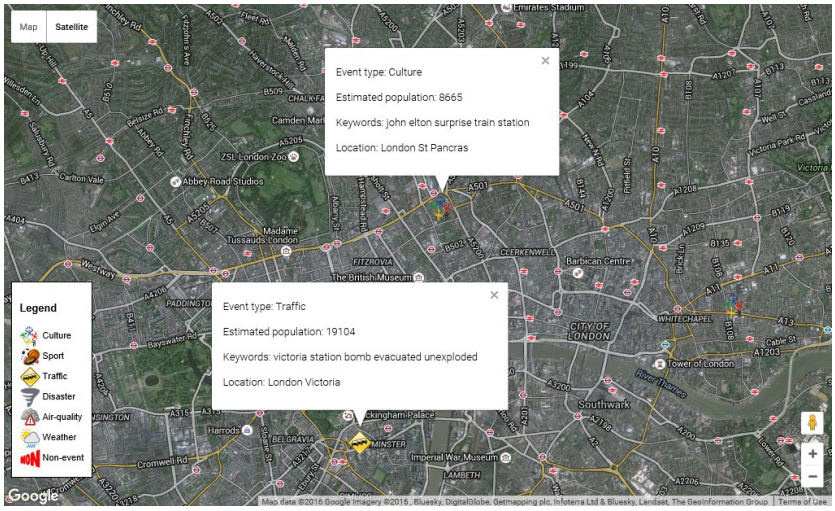

Fig. 1. Events' visualization on Google map.

Acknowledgements

This work is supported by the collaborative European Union and Ministry of Internal Affairs and Communication (MIC), Japan, Research and Innovation action, ''iKaaS" under EU Grant number 643262.

References

1. Finch, M. *Urban Migration*. Megatrends, 2015.
2. Anantharam, P., et al., *Extracting City Traffic Events from Social Streams.* ACM Trans. Intell. Syst. Technol., 2015. **6**(4): p. 1-27.
3. Becker, H., M. Naaman, and L. Gravano. *Beyond Trending Topics: Real-World Event Identification on Twitter*. in *Fifth International AAAI Conference on Weblogs and Social Media*. 2011.
4. Balduini, M., et al. *Social Listening of City Scale Events Using the Streaming Linked Data Framework*. in *Proceedings of the 12th International Semantic Web Conference - Part II*. 2013. Springer-Verlag New York, Inc.
5. Sakaki, T., M. Okazaki, and Y. Matsuo, *Earthquake shakes Twitter users: real-time event detection by social sensors*, in *Proceedings of the 19th international conference on World wide web*2010, ACM: Raleigh, North Carolina, USA. p. 851-860.
6. Sankaranarayanan, J., et al., *TwitterStand: news in tweets*, in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*2009, ACM: Seattle, Washington. p. 42-51.
7. Ritter, A., et al. *Open domain event extraction from twitter*. in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012. Beijing, China: ACM.
8. Chianese, A., et al., *An associative engines based approach supporting collaborative analytics in the Internet of cultural things.* Future Generation Computer Systems, 2016.
9. Cuomo, S., et al. *A Cultural Heritage Case Study of Visitor Experiences Shared on a Social Network*. in *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*. 2015.
10. Chianese, A., F. Marulli, and F. Piccialli. *Cultural Heritage and Social Pulse: A Semantic Approach for CH Sensitivity Discovery in Social Media Data*. in *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*. 2016.
11. Zhao, W.X., et al. *Comparing twitter and traditional media using topic models*. in *Proceedings of the 33rd European conference on Advances in information retrieval*. 2011. Dublin, Ireland: Springer-Verlag.
12. Blei, D.M., *Probabilistic topic models.* Commun. ACM, 2012. **55**(4): p. 77-84.
13. Griffiths, T., *Gibbs sampling in the generative model of Latent Dirichlet Allocation*, 2002.