From Smart Cities to Smart Neighborhoods: Detecting Local Events from Social Media

Yang Li
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Glasnevin, Dublin 9, Ireland

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

ABSTRACT

There are several examples of work which uses data from social media to detect events which occur in our real, physical world. Our target for event detection is to partition a large geographic region, a whole city in our case, into smaller districts based on geotagged Tweets and to detect smaller local events. We generate a language model for Tweets from each district and measure the KL divergence on incoming Tweets to detect outliers. When these reach a sizable volume or intensity and are consistent, this indicates an event within that district. We used Tweets drawn from Dublin city and we describe experiments on partitioning the city into districts and detecting local events within districts.

1. BACKGROUND AND RELATED WORK

Much research work is reported in the literature utilizing the characteristics revealed by Twitter features, including the realtime detection of live events. Event detection has long been a research topic across many application areas and using many sources of data or information [7]. Early work leveraged natural language processing tools, such as named-entity extraction for online news event identification. Such tools work well on well-structured text like newspaper articles and TV transcripts, but do not perform well over some forms of social media such as Twitter. To address this, other methods have been proposed. Twitterstand [5] gathers and disseminates breaking news from Twitter using an online clustering method to cluster similar Twitter messages. Sakaki et al. [4] classify Twitter contents using a Support Vector Machine. Twitcident [1] enables filtering, searching, and analyzing Twitter information streams during incidents as they are happening as well as providing a faceted search interface to dive deeper into these Tweets. Other works [6] also reports real-time event detection from Twitter based on temporal and textual features of Tweets.

These previous works successfully detect breaking news or live events in Twitter streams globally, and their methods are sensitive to large-scale events which attract a large number of possibly global Tweets, such as the Presidential inauguration in the USA. This is because their target events generate significant boosts to the mainstream of Twitter and a significant volume of event Tweets which can be detected. Yet Twitter users often post information about local, community-specific events such as a local flood, a fire, or a road closure because of a tree falling, where traditional news coverage at a regional or national level is non-existent and indeed it is quite difficult to confirm if such events have actually happened. We illustrate some of these later in Table 2. The motivation for our work is examine whether Tweets, localised to a small geographical region, can be used to detect unusual events happing at a *local* level within a city. Our contribution is to use Tweets from Dublin city to partition the city into smaller regions, model the typical Twitter content for each region and then use a sufficiency of outlier Tweets to indicate the likelihood of local-level events in areas of Dublin city.

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

2. EVENT DETECTION IN SOCIAL MEDIA

We work on a relatively small data-set, Tweets from Dublin city. For the purpose of the detection of unusual socio-geographic events, we first determine normal crowd behavior in a geographical region of the city in terms of Twitter activity. After mapping geo-tagged Tweets onto defined partitions on a map, we focus on sudden increases or decreases in the number of Tweets happening in a geographical partition or the topics of discussion, which can be clues to an unusual event happening. Our assumption is that local events can be reported on Twitter and the content of such Tweets is a semantic irregularity to the typical Twitter behaviour of a region, i.e. people do not normally Tweet about floods, fires or road closures unless there are such events happening.

To detect unusual local events for a given large area we first partition the city area into sub-areas by establishing socio-geographic boundaries. We adopt a clustering-based space partition method that reflects geographical distribution of a dataset and better deals with heterogeneous regions. Some research works divided their target area into equally sized grids with different granularities. We chose not to use this approach because an appropriate cell size is difficult to determine and does not consider the geographical distribution of Tweets.

We adopt the K-means clustering method based on the geographical occurrences of our Tweets. The K-partitioned regions are demonstrated in different colors on a unit graph, as shown in Figure 1. As a result, we achieve an appropriate socio-geographic boundary setting for the target region by distributing the actual occurrences of Tweets. We partition Dublin into 25 regions, a number which is a guesstimate as to what would be best. When we compare the partition results to the actual population distribution of Dublin city area according to the Central Statistical Office data, as in Figure 1, we see the partition results are acceptable, so 25

seems to have been a reasonable choice. Hotspots can easily be identified, such as the city center where there is high population density and a high volume of Tweets, as well as some low population areas with a high Tweet volume such as Dublin Airport, and the Phoenix Park.

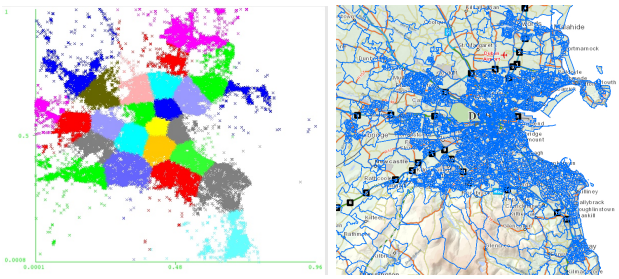


Figure 1: Geo-social partitioning of Dublin into 25 clusters and population distribution of Dublin

We make a major assumption that for each location there is consistency and periodicity in Twitter activity, such as appearances of regular users in regular locations and perhaps Tweeting about regular topics of interest. While some deviations outside usual or regular activities are caused by, for example, holidays, or visits from friends, these are mostly restricted to individuals. However when the same deviations are picked up by multiple users at the same time, same location, same topic, then this leads us to believe that we can recognize local events from inconsistencies in Twitter users' behavior at a regional level, including a change in the topic of Tweets, a so-called semantic irregularity.

We now explain how we set up the measurements of regularity. Within each partition of the city, there are Tweets generated over time, and in our work we analyze weekday and weekend days differently. This is because partitions have different activities for weekday vs. weekends such as offices which will be relatively quiet during weekends whereas shopping areas will be more active. The regularity of the total amount of Tweets are calculated using the average of each day during a rolling one month period, and with $\pm$ 1.0 standard deviation, assigned into hourly bins and any number outside the 1.0 standard deviation are considered as unusual activity, as shown in Figure 2.

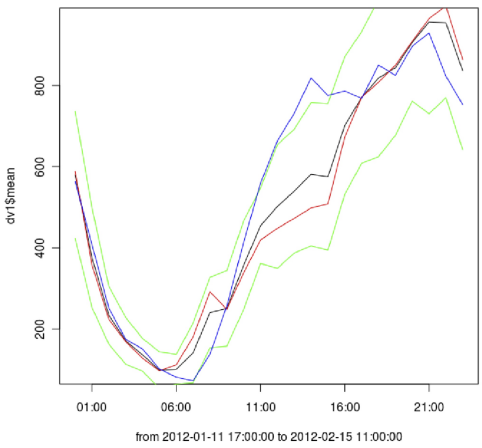


Figure 2: Twitter occurrences in hourly bins

For every partition we store a set of regular active Twitter users. If there are many visiting Twitter users sending Tweets from the partition, we consider this as another clue of irregular Twitter activity.

Measuring semantic regularity of Tweets in partitions is more complex. For each geo-tagged Tweet in our collection, we use all of the texts in each partition to build a language model that represents the semantic consistency of the partition. In order to preserve the semantics of Tweet contents we do not apply any stop-word filtering, and special characters such as "#" and "@" are not removed.

We use a language modeling approach to build individual models for each of the 25 partitions in the city allowing us to estimate the probability that a new Tweet issued from a given partition can be ranked by the probability that it was "generated" by the model. More concretely, given a set of locations L, and a Tweet T, our goal is to rank the locations by $P(L|T)$. Rather than estimate this directly, we use Bayesian inversion:

$$P(L|T) = \frac{P(T|\theta_L)P(L)}{P(T)} \quad (1)$$

where L is the model of the location. Assuming independence between terms:

$$P(T|\theta_L) = \prod_i P(t_i|\theta_L) \quad (2)$$

The probability of a term, given a location, $P(T_i|\theta_L)$, is estimated with Dirichlet smoothing [8]:

$$P(t|\theta_L) = \frac{c(t,L) + \mu P(t|\theta_C)}{|L| + \mu} \quad (3)$$

where μ is a parameter, set empirically, *c(t,L)* is the term frequency of a term t for partition L, $|L|$ is the number of terms in partition L. In this work we assume the prior probability of the partitions, *P(L)*, is distributed uniformly. We ignore $P(T)$, since it is the same for all partitions, and thus does not affect the ranking. partitions can be ranked directly by the probability of having generated the Tweet, or they can be ranked by comparing the model yielded by the Tweet, to the model of the partition, using Kullback-Leibler (KL) divergence. When ranking by KL divergence, we let θ_T be the language model for the Tweet T and L be the language model for the partition L. We use the Lemur Toolkit [2] for building our language models and carrying out our experiments.

Our aim is to detect geo-social events that result in unusual Twitter user behavior. For this, we define a socio-geographic boundary as under unusual conditions when its indicators, Number of Tweets (NT), Number of Users (NU) and Semantic Regularity (SR) satisfy the following function:

$$F = \alpha NT + \beta NU + \gamma SR \quad (4)$$

In function (4), F is a measure for the scale of an unusual event, α, β, and γ are coefficients for normalizing the measurements of each regularity. If the F is over a threshold, we predict that it is an indication that an unusual event is happening.

3. EXPERIMENTS

We crawled geo-tagged Twitter messages through the Twitter Streaming API. We setup a bounding box which covers

the Dublin area and from 24/Jan/2013 to 19/Mar/2013 we crawled English-only Tweets with exact geo-locations attached. This yielded 387,800 Tweets in total from 14,533 unique users, each of which we mapped to one of our 25 city regions. To test how well our language model represents the consistency of partitioning, we compared our predicted locations for Tweets to actual locations. We used location accuracy (Acc), which calculates the percentage of correct predictions over all test examples and we obtained an Acc value of 0.3347. We also used Mean Reciprocal Rank (MRR), obtaining a figure of 0.4290. Based on our experimental results we find that with our identified city partitions, the language models generated from the contents created inside each of the partitions provide good consistency for defining the regularity of each partition.

4. USER TWEETING BEHAVIOUR ANALYSIS

We now focus on two aspects of users' Tweeting behaviour: geographic (where we Tweet) and temporal (when we Tweet).

4.1 Analysis of Geographical Behaviour

One would expect that people typically exhibit strong periodic behaviour in their movement as they move back and forth between their homes and workplace [3]. We observed this pattern in our users' Tweeting locations using the 25 partitions into which the Dublin city area was partitioned. We identified 5,875 unique users from our dataset who generated 95% of the overall Tweets, which reduced our total number of Tweets to 368,476 and we eliminated users who only generate 1 or 2 Tweets within a month as these are possibly visitors to the city.

We observed strong periodic behaviour in the distribution of locations from where Tweets were sent. In Table 1, we see that almost 44 % of users sent Tweets from only 1 or 2 of 25 different zones across the city during this one-month period. It is reasonable to assume that these locations are the users'

Table 1: User Tweeting in different zones

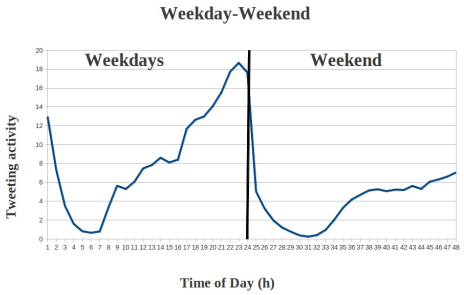
Number of zones	% of overall users
1	21.8%
2	22.7%
3	18.8%
4	13.7%
5-25	23%

homes, workplaces or leisure places. We also found that 23% of users generated Tweets across at least 5 seemingly random zones during this period and Tweets sent from these non-regular locations are of particular interest to our event detection task. If people only Tweet from their regular locations, their contents can be expected to be similar. Thus if we want to find irregular, unexpected event-related content, Tweets sent from non-regular locations should be of use.

4.2 Analysis of Temporal Dynamics

The volume of Tweets generated over time exhibits characteristics which potentially represent, in some way, each user's daily living patterns. Through studying temporal Tweeting behaviour, we can group users with similar daily life patterns. We aggregate the number of Tweets into hourly bins for each 24 hours, for weekdays and for weekends. Figure 3 shows trends from users' Tweeting patterns for weekdays and weekends in terms of the average number of Tweets generated per hour. Users are much less active during the

Figure 3: Overall Tweeting behaviour

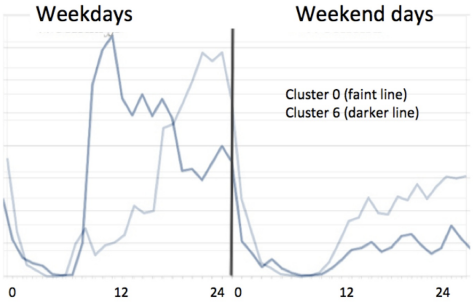


weekend than weekdays, and the boost in volume starts much later in the weekend, 2pm as compared to 8am during weekdays.

We focused on 805 users who sent more than 100 Tweets in a month, and clustered these users by their temporal Tweeting features. For each user, there are 48 features, each representing the average number of Tweets per one-hour window for weekdays and for for weekend days. We used the EM algorithm clustering from WEKA to assign these 805 users to 10 clusters. Within each cluster we detect instances where users have noticeably unique characteristics in their temporal Tweeting patterns, as shown below.

Figure 4 shows the aggregated activities of two groups. Cluster 0 consists of very active users, 10 times more active than average in terms of hourly Tweeting volume and we consider these people as general Twitter users, who are just more active than others. By contrast, users in cluster 6

Figure 4: Tweet distributions for Clusters 0 and 6



show completely different Tweeting patterns and we infer that these people are typical office workers, their Tweeting times peaking mostly during their lunch breaks, and after dinner, and they don't stay out late at night socialising.

5. DISCUSSIONS AND FUTURE WORK

Unlike other areas of multimedia information retrieval, there are no standardised test collections of content, and limited standard tasks to execute on harvested Twitter content.

Event and Date	Time	GPS Coordinates	Related Twitter Content
Local flooding in Glencree Valley Jan 25, 2013	16:45:10	53.1809595,-6.1887448	The flooding around #Glencreevalley #Enniskerry is crazy! Watch out drivers! #Aaroadwatch
	16:50:08	53.182842,-6.191808	my car is like a floating boat #Enniskerry #flooding
Car crash on O'Connell Street caused by heavy rain, Jan 25, 2013	17:28:32	53.1809595,-6.1887448	@aaroadwatch bus and car collision on o'Connell street sb
	17:30:32	53.348604,-6.2597	@RobbieH46 slowly....it's a fecking car crash!!!!
	17:30:50	53.347887,-6.259207	Poor man or women in car crash.. #sayapray dangerous driving in this weather #5wordweather @spin1038
Heavy traffic jam Blanchardstown, Mar 09, 2013	17:17:11	53.3948484,-6.3912147	massive traffic jam in blanch won't be home till Christmas
	17:21:49	53.394718,-6.389326	traffic freaks me out!!!
	17:05:01	53.393323,-6.393317	Caught in a traffic jam
Pipe burst, cut off water supply Clongriffin Jan 07, 2013	14:22:16	8. 53.404341,-6.158719	@DonnieWahlberg its raining we have no water because of a burst pipe I am bogged down in housework but I am happy and having fun anyway :-)
	22:32:06	53.2853,-6.22825	@seanm91 apparently while attempting to fix the water pipe they damaged the gas line #incompetence

Table 2: Examples of Detected Real-time Events

For event detection on a city-wide or national scale, like Presidential elections, international sports matches, major concerts or other major social occasions, there is a groundtruth against which event detections can be compared. But who knows if there really was slow traffic on the M50 near the Blanchardstown exit on the morning of 5th March 2013. Instead we point to anecdotal examples of four local events which occurred and were detected by our method and which are shown in Table 2.

6. CONCLUSIONS

In this paper, we examined a way to comprehend the dynamics of small, local areas within a city through social media based on consistencies across Twitter users' behaviour, covering location, time and content which does not form part of a language model for each of our 25 regions. We ran a series of experiments which showed consistency across these and we demonstrated detecting events at a local level.

An algorithm for detecting local events in real time based on location, time, and content, of Tweets has not been presented before and our method provides good classification performance at a local, almost parochial level. Although event detection from social media, especially Twitter, has been studied for some time there are still many challenges, especially for processing information at a fine-grained local level and we believe that such information, when relayed or forwarded (re-Tweeted) automatically to the right person, will be of use. Our next challenge is detecting the Twitter users to notify about such locally detected events.

7. REFERENCES

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 305–308. ACM, 2012.

[2] J. Allan, J. Callan, K. Collins-Thompson, B. Croft, et al. The LEMUR toolkit for language modeling and information retrieval. *The Lemur Project. http://lemurproject. org (accessed 25 January 2012)*, 2003.

[3] N. Eagle and A. S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[4] T. Sakaki, 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*, pages 851–860. ACM, 2010.

[5] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. ACM, 2009.

[6] H. Sayyadi, M. Hurst, and A. Maykov. Event Detection and Tracking in Social Streams. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.

[7] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proc. 21st annual international ACM SIGIR conference on Research and Development in information retrieval*, pages 28–36. ACM, 1998.

[8] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. 24th annual international ACM SIGIR conference on Research and Development in information retrieval*, pages 334–342. ACM, 2001.