5D-1

Using entropy and observed tweeting behavior to identify large events

Muhammad Asif Hossain Khan[1], Tomohiro Sakamaki[2], Masayuki Iwai[3], Yoshito Tobe[4], Kaoru Sezaki[5]

1 Graduate School of Information Science and Technology, The University of Tokyo 2 Graduate School Frontier Science, The University of Tokyo

3 Institute of Industrial Science, The University of Tokyo 4 School of Science and technology , Tokyo Denki University 5 Center for Spatial Information Science, The University of Tokyo

Abstract- In this paper, we present a method of identifying the occurrence of major events involving city dwellers. Instead of using data collected in controlled laboratory environment as is done in most contemporary research, we have used Tweeter's open platform to monitor the tweeting behavior of over 500 users for four months. Analyzing this vast data we have extracted a set of features to identify each user's normal threshold of tweets which varies dynamically depending on the time of the day. We have also used entropy, which gives indication about the density of twitter users at any hour of a day. Using these thresholds, and empowered with the knowledge of expected tweeter density at any time, we have devised a model that can predict occurrence of major events experienced by the Tweeter users in real time.

I. Introduction

Use of Online Social Networks (OSN) like Facebook and micro-blogging sites like twitter are gaining popularity in an unbelievable rate. Only the US traffic to Twitter grew 1,382% between February 2008 and February 2009, from 475,000 unique visitors to 7 million [1]. From January 2010 to August 2010 per day tweets have increased from 30 million to 90 million [2]. This clearly indicates that these social networking sites are going to be a very valuable source of information for researchers trying to model human behavior.

Research efforts have been made to use the information in the Tweets to find the center and trajectory of earthquake [3]. Some other researchers have investigated the intentions of twitter users at community level [4]. Researchers are considering that the large social interactions that take place over these sites can be used to study human behavior, e.g. people's marketing trend, propagation of ideas etc. But, we believe that we are getting carried away by the growth rate of these sites. It is very important first to verify that the users of these sites really represent the community. For example, for long term study of human behavior, like modeling the mobility pattern of a community based on the geo-tag information posted along with tweets, it is very important to first identify whether the twitter users are representative of that community. One way to do that is to investigate the change in behavior of the OSN users in response to some events, which has impact over the whole community – not just a small interest group (e.g. in case of events like upcoming rock concert). Such universal events could be either large natural catastrophe or common social events like observation of New Year eve or religious events like Christmas.

We examined the reprehensibility of Twitter users around Tokyo and found that their twitting behavior during New Year or Christmas certainly differs from that of any normal day. Based on the twitting behavior of over 500 twitter users, we also calculated the entropy for every hour of each day of the week, which reveals the expected density of twitter users at any hour of the day.

II. Method of Experiment

We monitored the twitting behavior of over 500 users for more than four months and developed profile for each of them. We used the Twitter API to crawl all tweets made within 30-Kilometer radius from the Imperial Palace in Tokyo. Twitter REST APIs does user and IP based rate limiting. Hence, it is not possible to crawl all tweets generated from an area. Therefore, we gradually selected over 500 most frequent tweeters and have followed them more intensely.

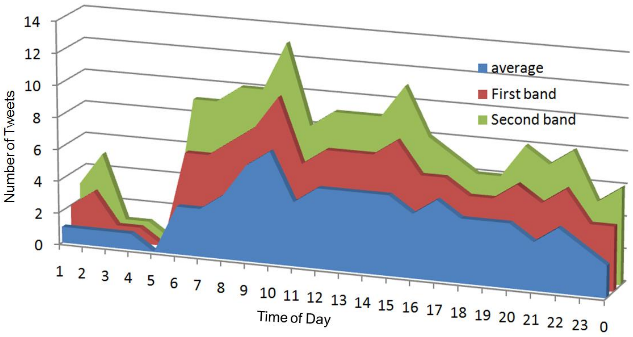

Figure 1. Profile for one particular user

$$\overline{X}_h(u_k) = \left\lceil \frac{\sum_d \overline{x}_{h,d}(u_k)}{7} \right\rceil \tag{1}$$

$$\overline{x}_{h,d}(u_k) = \begin{cases} \left\lceil \dfrac{N_{h,d}(u_k)}{f_{h,d}(u_k)} \right\rceil \\ 0 \quad if, f_{h,d}(u_k) = 0 \end{cases}$$

$$FirstBand_h(u_k) = \overline{X}_h(u_k) + \frac{1}{\sqrt{7}} \sqrt{\sum_d (\overline{x}_{h,d}(u_k) - \overline{X}_h(u_k))^2} \tag{2}$$

$$SecondBand_h(u_k) = \overline{X}_h(u_k) + \frac{2}{\sqrt{7}} \sqrt{\sum_d (\overline{x}_{h,d}(u_k) - \overline{X}_h(u_k))^2} \tag{3}$$

We profiled each of these users individually. A user profile (Fig. 1) includes the average number of tweets (Eq. 1) made by this user at any hour of a day. The other two profile features are the First band (Eq. 2) and the Second band (Eq. 3). The terms used in the equations have the following meanings:

$\overline{X}_h(u_k)$: Average expected tweets from user u_k at hour h

$\overline{x}_{h,d}(u_k)$: Average expected tweets from user u_k at hour h of day d of any week

$N_{h,d}(u_k)$: Total number of tweets done by user u_k at hour h of day d of any week

$f_{h,d}(u_k)$: Number of different weekdays, *d,* during which user u_k made any tweet at hour *h*

As the twitting behavior of a user is normal phenomenon, we claim that the probability that a user's number of tweet will not exceed the First band is 0.68 and that it will not exceed the Second band is 0.95. So, a 'significant' percentage of users crossing the second band certainly will indicate occurrence of some event. Now, the value for this 'significant' number will not remain same for every hour of the day. To determine this number we calculated the entropy (Eq. 4) (Fig 2) for every hour of a day based on the crawled data for each user. A high value of entropy means many users participate in twitting during that time and low entropy indicates that very few users are active at that period. The 'significant' number should be inversely proportional to the entropy value for any time of a day.

$$H(h,d) = -\sum_{k} P(N_{h,d}(u_k)\log P(N_{h,d}(u_k)) \qquad \textbf{(4)}$$

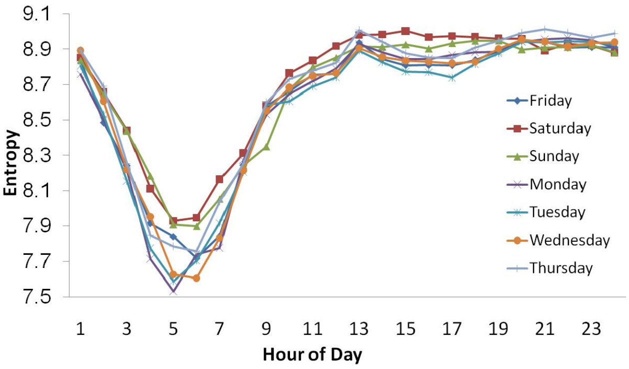

Figure 2. Entropy of different hour of day in different day of a week

III. Results and Discussion

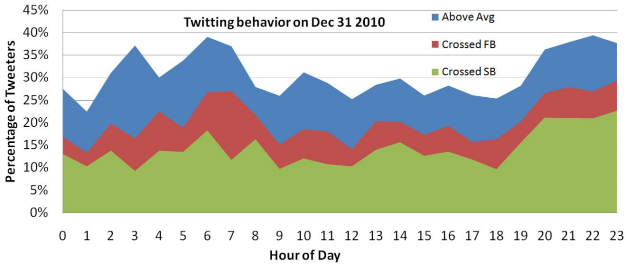

Figure 3. Twitting behavior on Dec 31 2010

We observed the twitting behavior of the users on December 25 and 31 of 2010 and January 1 of 2011. We then compared it to that of a normal weekday. The results for December 31 is shown in Figure 3. It clearly depicts that a significant percentage of tweeters are crossing not only their expected average tweet for Friday, but also a lot of them are crossing their second band. Similar phenomenon was observed for December 25 and January 1 also. Figure 4 shows same result for December 7 2010, which was a normal weekday. The difference is evident. In Figure 5 we presented a comparison of how many percentage of users cross their second band in an eventful and normal day. Our claim is substantiated here too.

In figure 4 at hours 5 and 6 the average tweet is quite high though nothing significant took place at that time according to our knowledge. This substantiates our claim that the value of 'significant' number as mentioned in the previous section should be made higher for those hours of the day when the entropy is low.

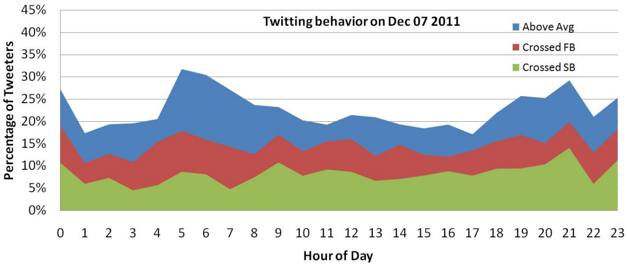

Figure 4. Twitting behavior on Dec 07 2010

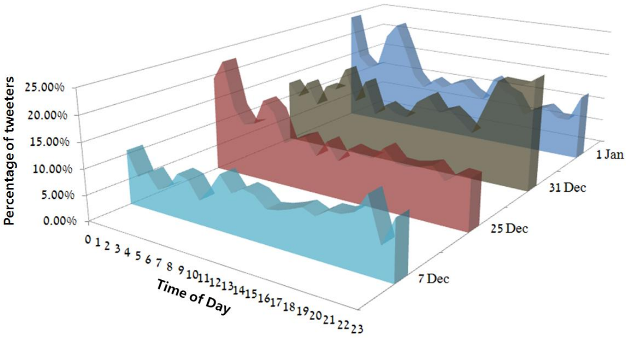

Figure 5. Comparison of Second band crossing among different days

IV. Conclusion and Future Work

Our model is able to detect major events in real time based on the twitting behavior of users. This indicates that twitter users are representative of a community. We are currently also gathering the geo-tag information posted with the tweets. We hope to build a human mobility model for a community using this geo-tag information. However, the problem that we have observed from our experiment is that the inconsistency of the twitter users in posting their geo-tag information. In our next project, we hope to overcome this problem by using some interpolation algorithms and also using some other data sources which might be used as a complement of twitter data source.

V. References

[1]. http://www.businessinsider.com/twitter-traffic-grows-1382-in-a-year-2009-3

[2]. http://techcrunch.com/2010/09/14/twitter-seeing-90-million-tweets-per-day/

[3]. T. Sakaki, M. Okazaki and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proc. WWW 2010*, 2010.

[4]. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007.*