Modeling Tweets Responding to Rapid Events

黄　俊† 　岩井原　瑞穂†

† Graduate School of Information, Production and Systems, Waseda University

2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135 Japan

E-mail: junhuang@akane.waseda.jp, iwaihara@waseda.jp

Abstract Rapid events refer to remarkable events which people would have high probability to mention immediately on Twitter when they occur or when they are going to occur, for instance, an astonishing goal in a soccer game. Different from normal ones, responses to these rapid events last only for a very short time period. Modeling tweets responding to rapid events introduces two challenges not encountered in previous work: background tweets which do not reflect any events and overlap situations, sometimes it is a mixture of series of events. In this paper, we discuss a numerical model which can explain tweet volume responding to a series of rapid events. We predefine three typical responses to different situations: 1) posting a huge number of related tweets right after the event, 2) tweeting during an on-going event, 3) making comments in advance of foreseeable results. Our target is to estimate strength of each event from a mixture of consecutive responses adaptively.

Keyword User reaction, Tweet, Modeling

1. Introduction

The explosion of social media is allowing unprecedented access to the thoughts and reactions of a large audience in response to breaking news, from entertainment and politics to technologies and products. In this paper, we call the news above as rapid events, and we choose to focus on Twitter, the most popular microblogging service provider, which offers great opportunities for analyzing the reactions of a wide audience with respect to rapid events.

Different from normal events, discussions for a rapid event only last for very short time period (e.g. several minutes, hours at most). We have to admit that, from the view of Twitter, people will continuously response to a certain topic for a long period of time. However, a burst of tweets may be expected right after an exciting event happens. With further observation, we discover two other bursty situations: making comments when an event is on-going and commenting in advance of some predictable results. For example, a huge number of tweets were posted during the new iPhone 4s releasing. In the case of last several minutes of football games, audience are likely to comment numerously for the coming winning or losing moment of their favorite teams. Hence, three typical burst of Twitter user reactions to rapid events are defined and we call them “BOA” for short (Before, On-going and After).

Our target is to study the user behaviors of Twitter in order to obtain a better understanding of both Twitter content and events and, estimate strength of each event to provide possibility for ranking these events in future. To achieve this goal, we first build our Tweet corpus from a Twitter firehose, a streaming search API of Twitter, spanning from June 11, 2010 to July 11, 2010 and focus on tweets containing target country names from the list of 32 World Cup participants. Then we provide a detailed study of a set of rapid events (goals, red card and yellow card) with respect to the game state (which team is taking lead, etc.) and a set of user engagement measures. Since the user behavior itself is hard to observe, we can only monitor game related features like team inclination, delay information, shallow linguistic features, etc. By proving correlation between features above and user-related features, we can finally capture user behavior.

The rest of the paper is organized as follows. We first review related work in Section 2. Then we will show features observed from real data in Section 3 and propose a probabilistic model with EM algorithm [1] for our problem in Section 4. Conclusion and future work will be given in Section 5.

2. Related Work

Our work is related to several lines of work in topic detection, intrusion detection, abnormality detection and multilingual natural language processing.

First, the work on Topic Detection and Tracking (TDT) [6, 7, 8] all aims to detect and track events from a steam of news stories, thus is related to our work. However, this body of work is focusing on topics which are often modeled as probabilistic distribution of terms. However,

tweets responding to a rapid event often lack reference to a topical term: users often mention emotions without referring to an event. This implies that just monitoring changes of term distributions fail to capture user responses. Another main difference between our work and TDT works is that we consider rapid events whose life circles are much shorter than the ones of events that TDT works focus on. This will cause big challenges that the boundaries of a series of rapid events may not be clear and there will be overlaps between events.

Second, the work of intrusion and abnormality detection has been studied extensively in recent years for network security issue. We borrow several concepts of change point detection in time series mining and adopt them to our work to distinguish abnormal changes of features, which is caused by a rapid event, from normal fluctuation over time. In the works of Kenji Yamanishi [3], a dynamic syslog mining method was proposed to monitor network failure which is treated as an abnormal situation. In the probabilistic related intrusion detection works[4], EM algorithm is often used for selecting and updating parameters. In our probabilistic estimation of significance of game events, we also use EM algorithm for estimating the model parameters, of the mixture Gaussian model.

Considering Twitter as a multilingual microblogging platform, most of our features were language-independent. However, in order to collect the sentiment information of tweets, we build a multilingual sentiment vocabulary by extracting sentiment phrases with high frequency.

Further, our approach is distinguished from previous work in the following regards:

1. Although, in most previous work, user behavior and tweets on rapid events have not been related explicitly, our paper gives a clear connection between them and presents empirical evidence of correlations between changes of features extracted from events related tweets and user reaction.
2. Previous work of Twitter users' response to events main focus on events with large time span, while Twitter has its realtime nature and user reaction can be captured immediately, hence we propose work of modeling tweets responding to rapid events, on which discussions will only last minutes, at most to hours.
3. A useful set of user engagement measures is proposed and types of reaction to rapid events are also properly defined to cover almost all the common situations.

3. Observation

In order to illustrate our work clearly, first we will use an example form results of real data observation to explain our problem.

3.1. Data Set

We choose the domain of football games to start data collection. Soccer is a good topic candidate because there exists multiple types of rapid events through the whole game, e.g. goals, free kick, cards, game states, etc. We collected tweets related to all the 64 matches in World Cup 2010 were collected

Utilizing Twitter search API [12], we can receive 15 tweets for each hash tag of country name in an atom file. Due to the limitation of connection to Twitter, we can get about 2250 latest tweet every hour. We assume tweets in an atom file are randomly sampled from all the tweets containing the hash tag, and statistically study the entire tweets from the received atom files.

Each entry of tweet contains 12 attributes among which 4 will be examined: *published time*, *title*, *author* and *language*. Note that the atom file has its own timestamp, which can be used for delay calculation and we will discuss it later in the next subsection. Utilizing these attributes, we create three content-based *user engagement measures*: *delay*, *linguistic features*, and *sentiment index (Team Inclination)*.

Figure.1 describes how these measures change during the game between Brazil and Netherlands on July 2. Due to limitation of the space, we have split Figure. 1 into several parts.

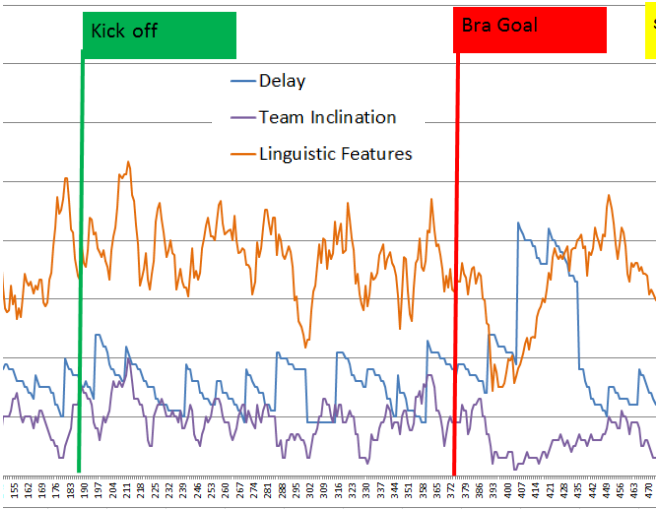

Figure.1 (a) Features at the beginning of the game

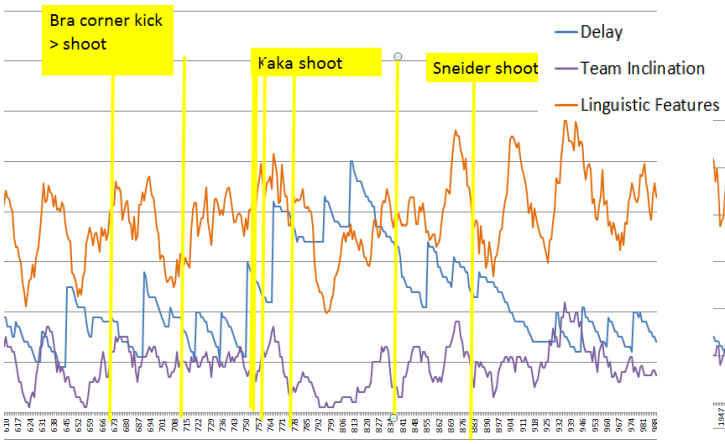

Figure.1 (b) Features at numerous events period

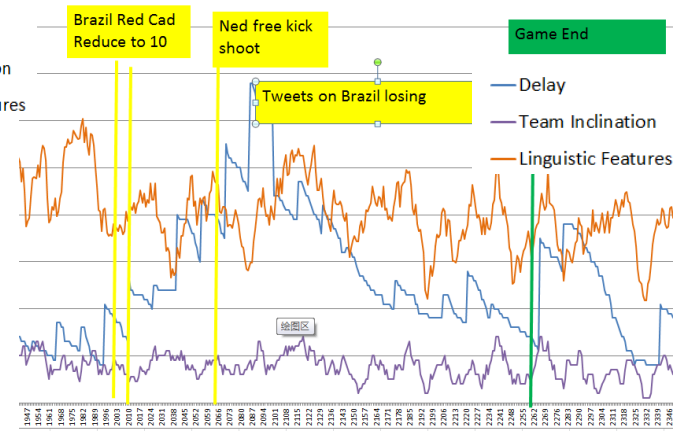

Figure.1 (e) Features approaching to the end

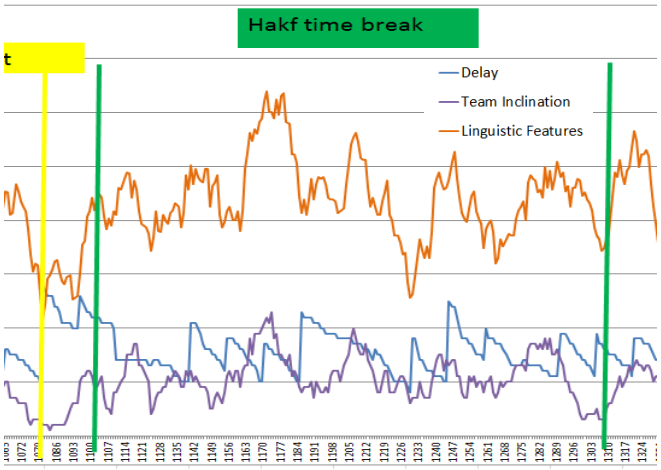

Figure.1 (c) Features during half time break

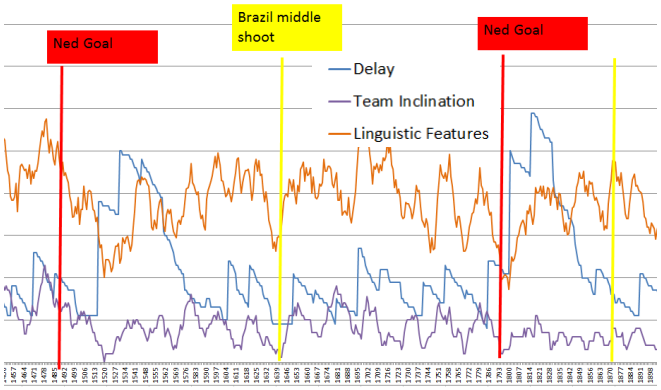

Figure.1 (d) Features after two goals

All these charts share same x-axis, which is the sequence of tweets we get over the game, the y-axis is the value of the features, however the magnitudes are different, and we just focus on the changes and shape of the feature curves.

Netherlands (NED)	Statistics	Brazil (BRA)
11	Shots	15
5	Shots on goal	4
2	Goals Scored	1
19	Fouls Committed	20
20	Fouls Suffered	17
4	Corner kicks	8
23	Free kicks Shots (scored)	17
0 / 0	Penalty Kicks (Goals/Shots)	0 / 0
3	Offsides	2
0	Own Goals	1
4	Yellow cards	1
0	Red Cards	1

Table 1 Overall statistics of the game provided by FIFA [11]

3.2. Rapid Events and Game States

Table 1 reports overall statistics of the game events in the match mentioned above.

- **Rapid Events**: Events like goals are the most discussed rapid events (more than 50% of all, frequency of tweets explicitly mentioning goals among all tweets related with rapid events), while suffered fouls are the least discussed (only about 8%). We had expected audience to talk more about events with high arguments like committed fouls or offsides. Instead, users seem highly engaged with events which may have big chance to create a goal, like corner kicks and free kicks.

- **Game State**: A game state is a vector of game-related features, like score, # of red cards, #of yellow cards, start/end of $1^{st}/2^{nd}$ half, etc. The game state will update after a game event included in the features listed above occurs.

Each game event, that causes game state changes, has the following properties:

- **Event Advantage** : this property has a value from {-1, 0,1}, indicating the side a game event is advantages for, team 1, team 2 or neutral, e.g. if we mark the value of goal scored by Brazil as 1, thus the value of goal scored by Netherlands is -1.
- **Event Predictability** : this property gives a binary value to indicate whether occurrence of a game event is predictable, like half end, penalty kick, or unpredictable, like goals. This is used to evaluate whether people are reacting to an event going to happen.

3.3. Delay

Theoretically, the responding time of a user should contain two parts: 1.Gap between the occurrence of the event and timestamp of starting writing the tweet; 2. Cost of time in editing the tweet content. However, these two parts are not directly observable.

Given the mechanism of Twitter search API, we find an alternative solution: once Twitter server receives a search request, it starts to collect tweets containing target hash tag from the pipeline of tweets. As long as 15 related tweets are collected it will send back the results in an atom file. Meanwhile, heavy server workload also leads to some delay. As showed in Figure.1, after a goal, average delay value increases because the explosion of tweet volume caused heavy workload.

A high delay value indicates it cost long time to collect 15 tweets, and we can infer that few people engaged in discussion of target hash tag in this time period, or totally contrary side, lots of people joined discussion.

To distinguish these two different situations, we designed a delay value $d(t)$ for a tweet t and standard deviation $\sigma(a)$ of tweets in on atom file a of as following:

$$d(t) = T(a) - T(t)$$

$$\sigma(a) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T(t_i) - \mu(a))^2}$$

$$\mu(a) = \frac{1}{N}\sum_{i=1}^{N} T(t_i)$$

Where, $T(a)$ represents timestamp of the creation of atom file a, and $T(t)$ represents the publish time of the tweet t. N is the number of tweets in an atom file. t_i is the i-th tweet in this file.

3.4. Linguistic Features

In this subsection, some shallow linguistic features are delivered in order to carry out analysis from the aspect of natural language process. Table 2 shows some tweet examples we collected.

Tweets
It's 1-1. A Brazilian mistake ... Sneijder's free kick comes off of Melo and it's an own goal. http://bit.ly/b52p0F
Goaaaaallll!! Hup Holland Hup! Suicide goal! Brazil dong dong!
Damn_ 1:1 Brazil vs. Netherlands.! That Goal was more than unexpected lol.
This match is on fire brazil VS Holland!!!!!!!!!!
Its ok holland don't be too happy! Brazil is gna get to u ;)
oh no brazil :(

Table 2 Tweet example

Based on our observation, people prefer use simple sentences when they are getting excited, that is after some rapid events. Another significant character is that repeated words and letters and usage of exclamation mark(!), question mark(?) are associated with emotional involvement. Thus, we defined following the features to measure excitement of users:

- **Length($L(t)$)** : number of tokens separated by the character of blank.
- **Fog Index($FI(t)$)** : Gunning fog index [9] estimates the years of formal education needed to understand the text on a first reading. The complete formula is:

$$0.4 * \left(\left(\frac{words}{sentences}\right) + 100 * \left(\frac{complex\ words}{words}\right)\right)$$

- **Run_Length($RL(t)$)**: this is an index for the repeated letters

$$L = log(1 + sum\ of\ (run\_length - k))$$

Here, *run_length* is the number of consecutive letters, where *run_length* > k. k is a parameter, works as a threshold. For example, if k=2 *run_length* of "Goooaaaall!!!"=1+3+4+2+3=13.

With these features, we can define a tweet complexity

c(t):

$$c(t) = w_1 * L(t) + w_2 * FI(t) - w_3 * RL(t) - w_4 * f(t)$$

Where, w_1, w_2, w_3, w_4 are 4 weight factors for adjusting magnitude of these features. *f(t)* stands for the number of occurrences of keywords, including team and player names, and occurrences of special marks.

3.5. Team Inclination

As we know, tweets on a football game come from three different sources: Supporters of team 1, supporters of team 2 and neutral audience. It is necessary to figure out who is reacting when something occurred. Thus, we propose following features:

- **Team Inclination** : We assume if a tweet contains more entities from team 1, the author may prefer team 1, otherwise he may be a fan of team 2, if equal, then he is neutral audience.

$$TI(t) = O1(t) - O2(t)$$

$$O1(t) = log\ (1 + freq(T1) + w\ freq(P1))$$

$$O2(t) = log\ (1 + freq(T2) + w\ freq(P2))$$

  Here, *freq(T1)* and *freq(P1)* refer to the occurrences of team name and player names of team 1, and w is a weight, here w equals to 1. All these entities are collected form an ontology extracted from the official site of FIFA World Cup [10].
- **Sentiment Lexicon** : sentiment lexicon help to determine whether this tweet is a cheer or complaint. Due to the problems of informal language and mixture of different languages, we sorted all the tweets to collect words with high frequency and manually picked up items with sentiment tendency. Our previous work on tweet sentiment classification also made contribution to collect the lexicon [5].

3.6. BOA

This subsection discusses how the features mentioned above changes in different type of reactions to rapid events.

- **B-type** : The typical predictable event in this game is the end of the game. In Figure.1 (e), when the game situation is approaching to the end, Brazil got a red card with 10 men to against the Dutch and trail for on goal, numerous tweets were tweets on their losing. This caused a significant raise of the delay value.
- **O-type** : In Figuare.1 (b), during this period, the delay value increased again for high-frequency rapid events. During the exciting moment, people are likely to post tweet.
- **A-type** : In Figure.1 (a) and (d), there are three goals, after goals, people cheer and shout with simple sentences associated with repeated words and letters, thus the index of linguistic feature have some drops. Meanwhile, there are also manifest raise of delay.

Besides, we also find in the period of half time break, people have more time to post long tweet, and the shape of index is steady.

4. Probabilistic Estimation of Significance of Game Events

Let $w_n\ (n = 1,..,N)$ be collected tweets. For $n = 1,..,N$, let t_n be the creation time of w_n, and let $\boldsymbol{x}_t$ be the *d*-dimension vector holding *d* tweet feature values of w_n, where the features are those discussed in Section 3. These features vectors are observed data from tweets. Now we consider probabilistic estimation of how each game event affects tweets.

We introduce the following assumptions:

1. Each *game event* (or *event* for short) is categorized into *predictable* (like half ends, corner kicks) and *unpredictable* (like goals, red cards). A change of the game state is also regarded as a game event.
2. Predictable events cause *before tweets*, while unpredictable events do not cause them. Both predictable and unpredictable events cause *after tweets*. Before tweets contribute to before bursts, while after tweets contribute to after bursts. On-going bursts are regarded as a mixture of before and after tweets on a series of game events. *Background tweets* are those containing game-related hashtags but occurring independently from game events, such as simple cheering tweets occurring in idle situations.
3. We adopt a probabilistic assumption such that the influence of a predictable game event on before tweets grows exponentially towards the event, and the influence of a predictable or unpredictable event on after tweets decays exponentially after the event. Formally, let t_i be the time event e_i occurred. The influence of e_i on before tweets at time t grows by growth function $g_i(t) = \exp((t - t_i)/r_g)$ for $t_i - t_e \leq t \leq t_i$ and $g_i(t) = 0$ for $t < t_i - t_e$ and

$t_i - t_e < t$, where $r_g > 0$ is a growth parameter and t_e>0 is a parameter called *influence window*. Likewise, the influence of e_i on after tweets at time t decays by decay function $d_i(t) = \exp(-(t-t_i)/r_d)$ for $t_i \le t \le t_i + t_e$ and $d_i(t) = 0$ for $t < t_i$ and $t_i + t_e < t$, where $r_d > 0$ is a decay parameter and t_e>0 is influence window.

4. Although the parameters r_d, r_g, t_e above can be dependent on events, for simplicity we assume fixed values for all the events. These values can be sampled from tweet data.
5. Although users can be influenced by other users' tweets, we assume that such influence is negligible compared with the influence from watching the game. Thus each tweet is occurring independently from other tweets, only responding to game events.

Following the assumption that tweets occur independently from each other, we try to model tweet probability as a finite mixture of Gaussian distributions. We categorize tweets into k *tweet groups* $G_1, \ldots, G_k$ such that G_1 is a set of background tweets, $G_i\ (i = 2, \ldots, k)$ is a set of before or after tweets on an identical game event. We model each tweet group by one d-dimensional Gaussian distribution with unknown mean value vector μ_i and variance co-variance matrix $\Sigma_i \in R^d \times R^d,\ i = 1, \ldots, k$. Here, R^d is the d-dimensional real vector space. We try to statistically estimate from which tweet group an observed tweet w_n occurred, using k-mixture Gaussian model. By $p(\boldsymbol{x}_n|\theta), n = 1, .., N$, we represent the conditional probability that tweet w_n presents feature vector $\boldsymbol{x}_n$ under unknown parameters θ. Using unknown mixture proportions π_i, $i = 1, \ldots, k\ \ 0 < \pi_i < 1$ such that $0 < \pi_i < 1$ and $\sum_{i=1}^{k} \pi_i = 1$, $\mathrm{p}(\boldsymbol{x}_n|\theta)$ can be written as:

$$\mathrm{p}(\boldsymbol{x}_n|\theta) = \sum_{i=1}^{k} \pi_i \delta_i(t_n)\, p(\boldsymbol{x}_n|\mu_i, \Sigma_i)$$

Here, $\delta_i(t)$ is the *lifetime function* of group i, such that $\delta_i(t)$ is either a growth function $g_i(t)$, decay function $d_i(t)$, or constant $\delta_i(t) = 1$, depending on whether i-th tweet group is before tweets, after tweets, or background tweets, respectively. Here, background tweets are constantly occurring throughout the game following the parameters μ_1 and Σ_1. Now the unknown parameters θ consist of $(\pi_i, \mu_i, \Sigma_i), i = 1, \ldots, k$. We discuss estimation of θ by the EM algorithm approach. From the estimated mean μ_i and variance co-variance matrix Σ_i, we can characterize tweet groups. Also, the mixture proportions π_i can tell us ranking of significance between tweet groups.

The literature[1] presents EM algorithms for multi-dimensional k-mixture Gaussian model. However, since our model involves lifetime function $\delta_i(t)$, the algorithm needs to be extended. We introduce hidden index variable z for representing from which model the observed vector $\boldsymbol{x}_n$ is occurring.

$$\mathrm{p}(z = i, \boldsymbol{x}_n|\theta) = \pi_i \delta_i(t_n)\, \mathrm{p}(\boldsymbol{x}_n|\mu_i, \Sigma_i)$$

The E-step of the EM algorithm, at iteration j, is to compute the following conditional log-likelihood expectation function Q under given parameters $\theta^{(j)}$:

$$Q(\theta|\theta^{(j)}) = \sum_{n=1}^{N} \sum_{i=1}^{k} \gamma_i^{(j)}(n) \log(\, \pi_i \delta_i(n)\, \mathrm{p}(\boldsymbol{x}_n|\mu_i, \Sigma_i))$$

Here, $\gamma_i^{(j)}(n) = \frac{\pi_i^{(j)} \delta_i(n) \mathrm{p}(\boldsymbol{x}_n|\mu_i, \Sigma_i)}{\sum_i \pi_i^{(j)} \delta_i(n) \mathrm{p}(\boldsymbol{x}_n|\mu_i, \Sigma_i)}$. Then the M-Step of the EM algorithm is to find $\theta^{(j+1)} = \arg max\ Q(\theta|\theta^{(j)})$, by solving $\frac{\partial Q}{\partial \theta} = 0$. The solution is as follows:

$$\pi_i^{(j+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_i^{(j)}(n),\ \ \mu_i^{(j+1)} = \frac{\sum_{n=1}^{N} \gamma_i^{(j)}(n) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_i^{(j)}(n)},$$

$$\Sigma_i^{(j+1)} = \frac{\sum_{n=1}^{N} \gamma_i^{(j)}(n) (\boldsymbol{x}_n - \mu_i^{(j+1)})(\boldsymbol{x}_n - \mu_i^{(j+1)})^T}{\sum_{n=1}^{N} \gamma_i^{(j)}(n)}.$$

Starting from given initial parameters $\theta^{(0)}$, the EM algorithm iterates the above E-step and M-step, computing $(j+1)$-th parameters using j-th parameters, until the parameters converge. The resulting parameters $\pi_i\ (i = 1, \ldots, k)$ represents the relative significance of tweet group i.

5. Conclusion and Future Work

In this paper, we defined and studied a novel problem of user behavior to rapid events. Based on our observation to the tweet data collected, we delivered several features to estimate user responses to the events, including delay information, linguistic features and team inclination. Moreover, a model on probabilistic estimation of significance of events with EM algorithm to select and update parameters is proposed.

Verifying this model on whole data set we collected will be the next mission of our work. Carry out extension to other domain will be necessary to prove our generality.

References

[1] Geoffrey J. McLanchlan, Thrutanbakam krishnan, “The EM Algorithm and Extensions”

[2] Ana-Maria Popescu. Marco Pennacchiott, “Dancing with the Stars, NBA Games, Politics:An Exploration of Twitter Users’ Response to Events”, AAAI 2011

[3] Kenji Yamanishi, Yuko Maruyama, “Dynamic Syslog Mining for Network Failure Monitoring”, KDD 2005

[4] Jun-ichi Takeuchi, Kenji Yamanishi, “A Unifying Framework for Detecting Outliers and Change Points from Time Series”, IEEE Transactions on Knowledge And Data Engineering Vol. 18, No4. April 2006

[5] Jun Huang, Mizuho Iwaihara, “Realtime Social Sensing of Support Rate for Microblogging”, Proc. 2nd SNSMW, LNCS 6637, pp. 357-368, April 2011.

[6] J. Allan, R.Papka and V.Lavrenko. “On-line New event Detection and Tracking”, SIGIR 1998

[7] Y. Yang, T. Pierce and J. Carbonell, “A Study of Retrospective and On-line Event Detection”, SIGIR 1998

[8] G. P. C. Fung, J. X. Yu, P. S. Yu and H. Lu, “Parameter Free Bursty Events Detection in Text Streams”, VLDB 2005

[9] http://gunning-fog-index.com/

[10] http://www.fifa.com/worldcup/archive/southafrica2010/index.html

[11] http://www.fifa.com/worldcup/archive/southafrica2010/matches/round=249718/match=300061507/index.html

[12] https://dev.twitter.com/docs/using-search