Modeling and Characterizing Social Media Topics Using the Gamma Distribution

Connie Yee, Nathan Keane, Liang Zhou
Text Analytics and Machine Learning
Thomson Reuters
New York, NY 10036, USA
{connie.yee,nathan.keane,l.zhou}@thomsonreuters.com

Abstract

We present a novel technique to identify emerging or important topics mentioned on social media. A sudden increase in related posts can indicate an occurrence of an external event. Assuming that the sequence of posts is a homogeneous Poisson process, this sudden change can be modeled using the Gamma distribution. Our Gamma curve fitter is used to return a set of emerging topics. We demonstrate our algorithm on Twitter data and evaluate empirically using the Reuters News Archive and manual inspection. Our experimental results show that our algorithm provides a good picture of the emerging topics discussed on Twitter.

1 Introduction

Over the past decade, microblogging sites, such as Twitter, have emerged as an important source of real-time news updates, with each microblogger acting as an information source. In contrast with news writing and reporting, microbloggers post content that is brief and uses colloquial language.

Some posts are reactions to events that have already broken out to the public. For content that originated in standard media outlets, such as newswire, the social medium can act as a filter and amplifier (Asur et al., 2011). Other posts serve as originators of events. For example, Twitter has been observed to lead newswire in reporting on sporting events and natural disasters (Petrovic et al., 2013). For sporting events, such as the FIFA World Cup, millions of users turn to microblogs to comment on what they just witnessed at a stadium or watched on television.

We are interested in discovering events related to both content from news outlets and content that originates on social media. An event occurrence can be detected by the volume and sudden change in volume of posts. After examining the distributions of the volumes of topics in Twitter, we observe two main categories of topics:

- Long-lasting topics that Twitter users frequently discuss in their daily lives, such as the foods they ate and the activities they are currently doing
- Emerging topics[1], or topics of importance to the general public, such as sporting events and natural disasters

Long-lasting topics tend to have a uniform distribution of volume over time, while emerging topics usually contain spikes in volume.

In this paper, we aim to detect the emerging topics by modeling a topic's frequency distribution with the Gamma distribution. It is a suitable function for modeling if we assume that the posts responding to an event arrive as a homogeneous Poisson process.

We begin with an initial set of event candidates by taking a topic modeling approach and assume that the words in a topic cluster represent one event. The event candidates are the inputs to the curve fitting algorithm, which returns the events that have valid model parameter values. We consider the outputs of our algorithm to be the emerging topics.

This paper is organized as follows: Section 2 introduces related work in event detection in Twitter.

[1] We view an emerging topic as an event so we use the words "topic" and "event" interchangeably in this paper.

Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015, pages 117–122,
Denver, Colorado, June 4, 2015.

Section 3 explains our modeling algorithm and the theory behind it. Section 4 reports our experimental results, which are evaluated in Section 5. We conclude and discuss future work in Section 6.

2 Related Work

Event detection in Twitter has been well-researched in recent years. Some focus on a keyword-based approach, such as through hashtags or term n-grams, to track trends. Shamma et al. (2011) investigated using a normalized term frequency to identify peaky and persistent topics. A challenge with a bursty term analysis is the difficulty in capturing an event with just a single string of words. Furthermore, the ability to identify an event requires that at least one term has a burst of relative frequency.

Other research has leveraged topic models as a means of learning clusters of events that are associated with an event. Topic models express a distribution over terms and thus are more descriptive than single keywords. Of the research that is based on topic modeling, much has been in the form of retrospective event detection models (Ramage et al., 2010). Recently, more work has been performed in the area of on-line processing of documents as they arrive (Lau et al., 2012), temporal topic models (Hong et al., 2011), and user-temporal mixture models (Yin et al., 2013).

There has been some prior work to incorporate the above-mentioned types of event detection methods with the properties of the topics or events. Zubiaga et al. (2014) aimed to classify trending topics by running a classifier using 15 features that consider the way a topic spreads.

Much of the focus of unsupervised methods has been on particular types of tweets or terms. Yang and Leskovec (2011) examined patterns of temporal behavior for hashtags. They presented the K-spectral centroid clustering algorithm to determine six classes of common temporal patterns that tweets containing hashtags follow. Further research by (Matsubara et al., 2012) proposed a general model for the rise and fall patterns of influence propagation. Zhao et al. (2012) studied a global bursty pattern derived from multiple types of tweets (posts, retweets, URL-embedded tweets) and modeled the smoothness of the state context. Their model was solely tested on keywords.

Shapes are a concise way of describing temporal variable behaviors. Each shape can be assessed by attributes, such as the rate a spike increases (Gregory and Shneiderman, 2012). There is evidence in data from the digital web site `digg.com` that the novelty of a topic determines how it decays over time (Wu and Huberman, 2007). Asur et al. (2011) observed that the number of tweets across trending topics can be characterized by a log-normal distribution and a linear decay. The trending topics were provided by the Twitter Search API and mostly consisted of two to three word expressions.

3 Modeling Topic Frequency Distributions

3.1 Topic Modeling and Segment Selection

This section describes how we form our initial set of event candidates and then select the segment of the frequency distribution for the next step of our algorithm.

Topics can be extracted from textual corpora through probabilistic topic models. Latent Dirichlet Allocation (LDA) is a widely adopted generative model for topic modeling (Blei et al., 2003). For each document, there is a multinomial distribution over topics. For each topic, there is another multinomial distribution over words. A popular algorithm for LDA model parameter estimation and inference is Gibbs sampling (Griffiths and Steyvers, 2004).

We used an LDA algorithm, similar to the MALLET topic model package (McCallum, 2002), with an efficient Gibbs sampling to identify 50 topics per day as event candidates. Each tweet was treated as one document. The resulting topics were then analyzed as follows:

1. Count the number of tweets that contain at least 30% of the topic in 15-minute intervals.

2. Determine the most relevant portion of the time series to model. Identify the highest peak and the points immediately preceding and following it, whose volumes are at least $x\%$ of the peak volume. We experimented with x ranging from 10–90% in increments of 10% and selected $x = 30$ based on manual inspection.

3.2 Modeling Tweet Frequency

In this section, we explain how we model the number of tweets regarding a particular event.

We envisage the arrival of tweets as a Poisson process. A Poisson process is a widely-used stochastic process for modeling the times at which arrivals enter a system. The sequence of interarrival times $X_1, X_2, ...$ in the Poisson process is a sequence of independent and identically distributed (IID) random variables, each having a probability density of an exponential, $f_X(x) = \lambda e^{-\lambda x}$, for some rate $\lambda > 0$ and $x > 0$. A unique property of the Poisson process is the memoryless quality. This means that the distribution of the remaining arrivals is the same as the original arrival time distribution, i.e. the remaining arrival time has no "memory" of previous arrivals.

Using the Poisson distribution, we model a poster tweeting after an event as an IID random variable with an exponential density function $f_X(x)$. Assuming a homogeneous Poisson process, where the posting rate λ for this event is constant, a second poster independently tweeting after the same event also has an exponential density function $f_X(x)$.

The interarrival times of tweets after an event then become the sum of n IID random variables, each with the density function $f_X(x) = \lambda e^{-\lambda x}$. Given that the density of the sum of two independent random variables can be found by convolving their densities, the convolution of multiple exponential distributions is called the Gamma density (Akkouchi, 2005). Thus, the time of the n^{th} post, T_n, follows a Gamma distribution.

If we let N_t be the number of posts in time interval $[0, t]$, it can be shown that $\{N_t \geq n\}$ and $\{T_n \leq t\}$ represent the same event. Using this duality, we can fix the time interval and model the frequencies of the tweets.

3.3 Curve Fitting and Parameter Estimation

The Gamma distribution has three different types, one of which is the two-parameter gamma distribution, given by (1).

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty; \alpha, \beta > 0 \tag{1}$$

The parameter α is known as the shape parameter, since it most influences the peakedness of the distribution, while the parameter β is called the scale parameter, which mostly influences the spread of the distribution.

Since there is no closed-form solution for the Gamma distribution, we used a heuristic search method to estimate the parameters of the distribution. A commonly used nonlinear optimization technique called the Nelder-Mead simplex algorithm (Lagarias et al., 1998) was employed for this purpose.

To avoid the need to normalize the time series, we fit the time series segments to the three-parameter probability density function. It can be obtained from (1) by adding a scaling factor A_0 and replacing x by $x - \mu$, where μ is the location parameter, as in (2).

$$f(x; \alpha, \beta, \mu) = \frac{A_0}{\Gamma(\alpha)\beta^{\alpha}} (x - \mu)^{\alpha-1} e^{-(x-\mu)/\beta}, \quad x \geq \mu; \alpha, \beta > 0 \tag{2}$$

The estimated values for α and β, as well as the sum of squared errors, or χ^2, were further analyzed. A threshold on A_0 can be optionally set so that only tweets that meet a minimum volume level are considered.

4 Experimental Results

Our experiments were conducted in a retrospective fashion, whereby we assumed the full document collection was given as input.

4.1 Data Cleaning and Topic Modeling

First, we gathered approximately 127 million tweets spanning 2014-06-14 0:00 GMT to 2014-06-27 11:59 GMT from Twitter Decahose, which is a feed of 10% of all tweets. We then conducted preprocessing by removing stopwords, URLs, and non-ASCII characters.

Following the data cleaning, we ran LDA on each of the 14 days of tweets to obtain 700 topics. Out of the 700 raw topics, we achieved convergence with defined χ^2 for 36 topics. Table 1 lists four topics that were randomly selected for further examination.

Date	Topic	Top Words
2014-06-14	Stanley Cup	game kings cup win hockey
2014-06-15	Wonder Goal	goal messi argentina france #worldcup
2014-06-19	Biting	england rooney suarez goal uruguay
2014-06-27	Player Contract	money pay million shaw united

Table 1: Selected topics.

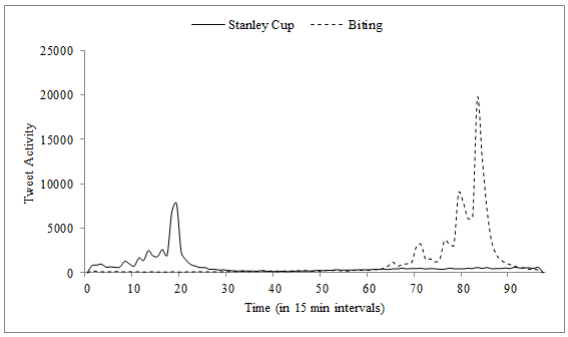

Figure 2: Distributions of two topics with $\beta = 0.08$ and different αs.

4.2 Curve Fitting

The frequency distribution of the "Stanley Cup" topic over a 24-hour window is shown in Fig. 1. The curve segment between the two labeled points served as the input to the curve fitter, which estimated α and β to be 48.55 and 0.08, respectively.

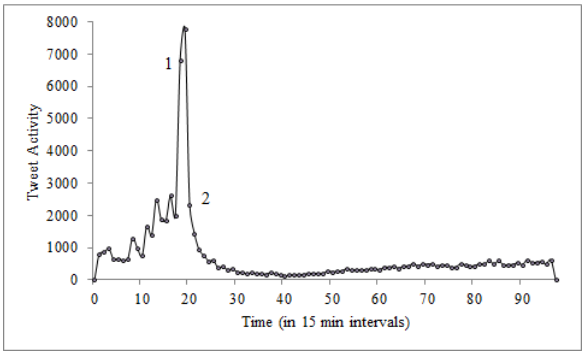

Figure 1: Frequency distribution of the "Stanley Cup" topic.

To better understand these estimated parameter values, we can compare it to another topic with the same β value. Fig. 2 shows the distributions of the "Stanley Cup" (solid line) and the "Biting" (dotted line) topics. The "Biting" topic, which refers to a shocking biting incident during the World Cup, has a sharper peak, thereby translating to a higher α value of 129.25. On the other hand, the "Stanley Cup" topic denotes an expected or planned event whose outcome happened to be predictable.

We can analyze the effect of the β parameter by keeping α constant. Fig. 3 shows two topics with the same α value. The solid line represents the "Player Contract" topic, while the dotted line is the "Wonder Goal" topic. The latter topic refers to one of the greatest soccer goals made by a player. While this event is impressive enough to make it on social media, it appears to dissipate quickly and is likely soon replaced by the next great play in the World Cup. Its α value is a mere 1.69. In contrast, the "Player Contract" topic with α of 8.03 is discussed over the course of ten hours, as the signing of a well-known player to a new team can have great implications for the coming season.

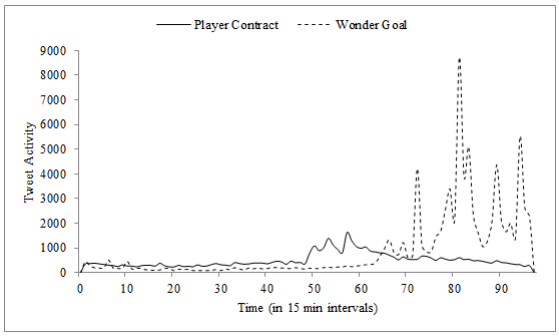

Figure 3: Distributions of two topics with $\alpha = 1.74$ and different βs.

5 Evaluation

We evaluated empirically all the events discovered from the curve fitting algorithm. For purposes of evaluation, we considered an event to be an actual event if it falls in one of two categories:

- *news*, if it reached the standard media outlets
- *social*, if it was solely discussed on social media

By determining the number of news and social events, and dividing it by the total number of events discovered, we calculated precision, as defined in (3). Our algorithm achieved 77.8% precision.

$$Precision = \frac{|news| + |social|}{total} \quad (3)$$

5.1 News Events

We leveraged the news domain to identify news events. Traditional news media, such as Reuters, typically span a wide range of categories, from fashion to finance. Although its distribution over the categories differs from that in Twitter, it is safe to assume that if an event is mentioned in newswire, it carries some importance.

We performed a query-based search in the Reuters News Archive to collect documents written within one day of the event date. By querying stories both before and after the event, we analyzed events that originated either in newswire or on social media. A news story was counted if it contained at least five of the top ten words. 15 of the 36 topics had at least one corresponding story in Reuters News, and concentrated on major sporting events.

5.2 Social Events

There are events that fail to reach the standard media outlets but are significant in the social media context. We inspected the remaining 21 topics which lacked a corresponding news story and categorized them into three main areas, as shown in Table 2.

Entertainment	#shawntotop shawn buy follow
	follow sos love luke
Daily Life	happy birthday day love hope
	day happy fathers dad
Twitter Related	tweet cool funny haha post
	follow ya follback yo click

Table 2: Examples of events not mentioned in newswire.

After examining some representative tweets, we concluded that the "Entertainment" events were largely based on the Twitter users' interests, such as a new music album release. They were labeled as social events. The "Daily Life" and "Twitter Related" topics are examples of long-lasting topics that do not carry much news nor social significance.

6 Conclusion and Future Work

Our novel technique based on the Gamma distribution offers a useful starting point for using the shapes of the frequencies to determine whether a topic is an emerging topic. Although some long-lasting topics were also detected, the algorithm is able to provide a good picture of the news and social events discussed on social media. Some advantages of our method are that it is unsupervised and independent of how the initial set of event candidates are formed, which means that LDA can be replaced with a different topic model.

While we made simplifications and assumptions in our algorithm, there are several directions for future research. One area is to relax the assumption of modeling the sequence of posts as a homogeneous Poisson process. Since the posting rate λ for an event likely changes over time, we can divide the entire sequence into smaller segments and model each separately. In addition, removing cyclical or seasonal topics before curve fitting may help eliminate false positives.

References

Mohamed Akkouchi. 2005. On the convolution of exponential distributions. *Soochow Journal of Mathematics*, 31(2):205-211.

Sitaram Asur, Bernardo A. Huberman, Gabor Szabo, and Chunyan Wang. 2011. Trends in social media: persistence and decay. In *Proceedings of the 5th International AAAI conference on Weblogs and Social Media*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Machon B. Gregory and Ben Shneiderman. 2012. Shape identification in temporal data sets. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 305–321. Springer London.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Science*, 101: 5228-5235.

Liangjie Hong, Dawei Yin, Jian Guo, and Brian D. Davison. 2011. Tracking trends: incorporating term volume into temporal topic models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 484–492.

Jeffrey C. Lagarias, James A. Reeds, Margaret H. Wright, and Paul E. Wright. 1998. Convergence properties of

the Nelder–Mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147.

JeyHan Lau, Nigel Collier, and Timothy Baldwin. 2012. On–line trend analysis with topic models: #twitter trends detection topic model online. In *Proceedings of COLING 2012*, pages 1519–1534.

Yasuko Matsubara, Yasushi Sakurai, B. Aditya Prakash, Lei Li, and Christos Faloutsos. 2012. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 6–14. ACM.

Andrew K. McCallum. 2002. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Sasa Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton. 2013. Can Twitter replace Newswire for breaking news? In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*.

Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 130–137.

Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2011. Influence and passivity in social media. In *Machine Learning and Knowledge Discovery in Databases*, pages 18–33. Springer Berlin Heidelberg.

David A. Shamma, Lyndon Kennedy, Elizabeth F. Churchill. 2011. Peaks and persistence: modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pages 355–358. ACM.

Fang Wu and Bernardo A. Huberman. 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 177–186. ACM.

Hongzhi Yin, Bin Cui, Hua Lu, Yuxin Huang, and Junjie Yao. 2013. A unified model for stable and temporal topic detection from social media data. In *Data Engineering (ICDE)*, pages 661–672. IEEE.

Wayne X. Zhao, Baihan Shu, Jing Jiang, Yang Song, Hongfei Yan, and Xiaoming Li. 2012. Identifying event–related bursts via social media activities. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1466–1477. Association for Computational Linguistics.

Arkaitz Zubiaga, Damiano Spina, Raquel Martinez, and Victor Fresno. 2014. Real–time classification of Twitter trends. *Journal of the Association for Information Science and Technology*.