Predicting Bursts and Popularity of Hashtags in Real-Time

Shoubin Kong[1], Qiaozhu Mei[2], Ling Feng[1], Fei Ye[3], Zhe Zhao[2]
[1]Dept. of CS&T, Tsinghua University
{kongsb09@mails., fengling@}tsinghua.edu.cn
[2]School of Information, University of Michigan
{qmei, zhezhao}@umich.edu
[3]School of Statistics, Capital University of Economics and Business
feiye08@gmail.com

ABSTRACT

Hashtags have been widely used to annotate topics in tweets (short posts on Twitter.com). In this paper, we study the problems of real-time prediction of bursting hashtags. Will a hashtag burst in the near future? If it will, how early can we predict it, and how popular will it become? Based on empirical analysis of data collected from Twitter, we propose solutions to these challenging problems. The performance of different features and possible solutions are evaluated.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous

Keywords

Hashtag, burst, real-time prediction

1. INTRODUCTION

As one of the leading platforms of social communications and information dissemination, Twitter has become a major source of information for common Web users. Conversations on Twitter are featured with their "burstiness", the phenomenon that a topic of discussion suddenly gains a considerable popularity, and then quickly fades away. Such bursting topics are usually triggered by breaking news, real world events, malicious rumors, or various types of behavior cascades such as campaigns of persuasion.

These bursting topics, usually referred to as trending topics, provide users with fresh discoveries and timely updates of events. Much research work has investigated the value of the bursting topics in a broader context. Bursts of topics have been demonstrated to have a predictive power of product sales, stock market, search engine queries, elections, and even outbursts of diseases. Therefore, an earlier detection of such bursting topics implies an increased revenue, a reduced damage, a timely treatment, and better decision-making in general. To help people access bursting topics in time, Twitter deploys a list of trending topics as they are detected.

However, it may be already too late to react even if a "burst" can be *detected* in no time. On April 23, 2013, a false claim about explosions at the White House and the injury of the president sent by the hacked account of the Associated Press quickly became an explosive burst on Twitter [1]. Although the rumor was debunked as soon as it raised the concern of observers, damage had been made - the bursting topic had shaken the stock market so badly that the Dow Jones Indices experienced a sudden drop of more than 100 points. If only we can *predict* the outbreak of a topic *before* it bursts! But can we?

Hashtags, user-specified strings prefixed with a # symbol, have been widely used as identities of topics on Twitter. From a 10% random sample of the tweet stream, we can identify hundreds of thousands of new hashtags each day. However, only few hashtags can become trending topics. Among the few, it is likely that a smaller proportion can be predicted early on. In this study, we present the real-time prediction of these hashtags before they actually burst. There has been some closely related work. Nikolov *et al.* [6, 1] selected the slice of time series centered at the trend onset of trending topics as reference signals, and new hashtags were classified by comparing observed time series to these signals. Lin *et al.* [4] proposed a framework to capture dynamics of hashtags based on their topicality, interactivity, diversity, and prominence. Ma *et al.* [5] predicted the popularity of hashtags on daily basis. Our work, conversely, concentrates on predicting the bursts of hashtags as early as possible, and the popularity of those hashtags if they burst. Built on top of our preliminary work [3], we present in this paper formal definitions of four different states in the life cycle of a bursting hashtag and two corresponding prediction tasks. We provide a detailed examination of the contribution of different types of features. The major contribution of this paper includes:

1. We provide formal definitions of **four states** in the life cycle of a bursting hashtag, based on which we present **two** real-time prediction **tasks** of bursting hashtags.

2. We present solutions to the two prediction tasks (**burst** and **popularity**), and evaluate the performance of different types of features and different methods with real data.

SIGIR'14, July 06 - 11 2014, Gold Coast, QLD, Australia.
Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609476.

[1] http://www.foxnews.com/us/2013/04/23/hackers-break-into-associated-press-twitter-account/

2. PROBLEM STATEMENT

2.1 Definitions

Given a hashtag, the tweets containing this hashtag are counted at equal time intervals to generate a time series $\langle C_1, C_2, ..., C_t, ...\rangle$. In this paper the granularity of the time interval is set to 1 minute. For example, C_1 denotes the count of tweets with this hashtag within the very first minute that the hashtag first appears.

A very small proportion of hashtags will actually burst. We monitor the time series of all hashtags, but only those satisfying certain criteria will trigger the prediction(s). The function that evaluates the criteria is called a "prediction-trigger." As described in [3], a prediction-trigger checks the count of a hashtag within a five-minute sliding window ever after its first presence. If the count is greater than a threshold, the hashtag is marked *active*, indicating that it has the potential to burst in the near future. That hashtag thus enters a bursting life cycle consisting of four key states:

State 1. **Active**. If there exists a time interval t_a in the life cycle of a hashtag, where $C_{t_a} + C_{t_a+1} + C_{t_a+2} + C_{t_a+3} + C_{t_a+4} > \phi$, we say that the hashtag becomes active since time t_a. Predictions for that hashtag will be triggered as soon as it becomes active.

State 2. **Bursting**. As defined in [3], within 24 hours after a hashtag becomes active, if there exists a time interval t where the count of a hashtag during that time interval, C_t, is greater than $max(C_1 + \delta, 1.5C_1)$, we say that the hashtag bursts and t is the *start* of the burst. Note the threshold is defined according to the definition of spikes in [2]. When C_1 is sufficiently large (i.e., $C_1 > 2\delta$), $C_1 + \delta < 1.5C_1$, and $1.5 \cdot C_1$ should be used as the criterion of a burst. Otherwise $C_1 + \delta$ is used as the criterion.

State 3. **Off-Burst**. For a hashtag that is busting, if starting from a time interval t', all the counts $C_{t'}$ are smaller than $max(C_1 + \delta, 1.5C_1)$ in the following 24 hours, we say that the hashtag is off-burst and t' is the *end* of the burst.

State 4. **Inactive**. As the "off-burst" state is defined corresponding to the "bursting" state, the "inactive" state is defined according to the "active" state. If a hashtag can no longer meet the condition for triggering another prediction in 24 hours, the hashtag is considered inactive. A complete life cycle of the bursting hashtag comes to an end.

We adopt the same setting in the previous work [3] and set both parameters δ and ϕ to 50. When this bursting hashtag becomes inactive, it is treated as a new hashtag and will be monitored again. In other words, when a hashtag comes to the end of the current life cycle, a next life starts.

Table 1 shows the distribution of bursting hashtags in difference states of their life cycles, estimated from the 10% sample of tweet stream. The three columns in the table, PAB, POB, and PAI, are proportions defined as follows:

$$PAB = \frac{\#hashtags\ already\ bursting\ at\ time\ t}{\#hashtags\ that\ burst\ at\ all} \quad (1)$$

$$POB = \frac{\#hashtags\ already\ off\ burst\ at\ time\ t}{\#hashtags\ that\ burst\ at\ all} \quad (2)$$

$$PAI = \frac{\#hashtags\ already\ inactive\ at\ time\ t}{\#hashtags\ that\ burst\ at\ all} \quad (3)$$

Table 1: Bursting hashtags in different states.

Time since active	PAB(%)	POB(%)	PAI(%)
5 minutes	30.32	4.35	0.43
15 minutes	51.02	39.53	14.60
30 minutes	68.03	47.29	19.38
1 hour	80.88	59.90	32.84
3 hours	92.68	81.15	61.95
6 hours	**95.12**	88.50	73.46
24 hours	100	**96.11**	88.60
48 hours	100	99.60	**98.15**

Among all hashtags that have ever bursted in their life cycles, about 95% started *bursting* within 6 hours since they became *active*; about 96% were *off-burst* within 24 hours since active; and about 98% were already *inactive* within 48 hours since active.

2.2 Prediction Tasks and Solutions

Based on the above definitions, multiple prediction tasks can be done in real-time. When a new hashtag appears, it is monitored and checked whether it becomes active. When it satisfies the condition, predictions are triggered. Given a hashtag that has already reached, or is predicted to reach one of the four states, one can attempt to predict whether it will reach the next state and how long it takes to reach the next state. One can also make predictions for the actual statistics of a hashtag when it reaches a particular state. For example, one can predict whether a hashtag is going to burst, and if yes how long it takes to burst, and how long the burst will last. Among many possibilities, we selectively discuss two particular prediction tasks in this paper.

Task 1. Will an active hashtag burst in the near future? This problem is framed as a binary classification task. For a hashtag that is marked as active but is not currently bursting, a binary label will be assigned to indicate whether it will burst within 24 hours since active.

Solution: We have reported that weighted SVM achieves the best performance in [3]. In this paper, we provide details on the effectiveness of different types of features.

Task 2. If a hashtag is detected to be bursting, how popular will it be? Since about 96% of bursting hashtags are off-burst within 24 hours since active, our goal is to predict the popularity of a hashtag in 24 hours since active. The popularity (Pop) of each bursting hashtag is indicated by the total count of tweets containing the hashtag. Note that predicting the exact value of Pop is extremely difficult and generally not necessary. Therefore we relax the problem and predict the natural logarithm of Pop, $log(Pop)$. This problem can be framed as a regression task.

Solution: We explore five different regression models, including Linear Regression (LR), Classification And Regression Tree (CART), Gaussian Process Regression (GPR), Support Vector Regression (SVR) and Neural Network (NN). The evaluation results are summarized in Section 4.

The first prediction task is much more challenging. Firstly, our goal is to predict the bursting hashtags before they actually burst, not to detect them when they are already bursting. This is an inherently difficult problem. Secondly, the observations are largely unbalanced. Fig. 1 shows that among the active hashtags that have triggered the predic-

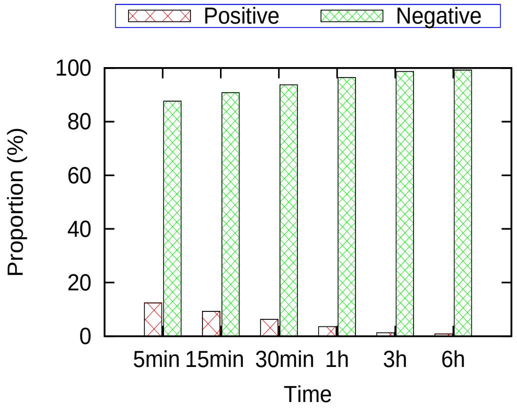


Figure 1: Proportions of bursting hashtags in active hashtags, over time. Positive: hashtags that are going to burst after t; Negative: hashtags that are active but never burst.

Table 2: Importance of features in Task 1. Removing time series features results in the largest drop of performance. This decrease is less significant when predictions occur early.

Features	F_1-score (prediction at time t since active.)					
	5m	15m	30m	1h	3h	6h
All	0.326	0.372	0.368	0.360	0.290	0.250
-Meme	0.319	0.379	0.358	0.344	0.238	0.229
-User	0.332	0.382	0.352	0.338	0.171	0.225
-Content	0.332	0.368	0.352	0.349	0.228	0.220
-Network	0.317	0.381	0.361	0.355	0.179	0.206
-Hashtag	0.322	0.382	0.354	0.313	0.272	0.197
-Prototype	0.313	0.375	0.364	0.360	0.288	0.235
-Time*	**0.308**	**0.334**	**0.313**	**0.242**	**0.147**	**0.069**

* Time series features.

tion, only a small proportion of them are going to burst. It is quite challenging to precisely predict so few bursting hashtags from the the rest of hashtags that are active but never burst. The proportion of hashtags that are going to be bursting even goes down to **0.8%** at the 6th hour after they becomes active. This is not surprising as most bursting hashtags have already bursted at that time.

3. FEATURE SPACE

In this section, we present different types of features which may be effective in predicting whether a hashtag is going to burst and how popular it will become.

Meme Features. This category consists of "count" features and "ratio" features. The former include number of tweets, authors, retweets, and mentions associated with the hashtag; the later include the ratio of authors, retweets, mentions, and urls among tweets with the hashtag.

User Features. Three features about the users who adopted the hashtag are considered: activity of the users, sum of their followers, and maximum number of followers, which may reflect the potential scale of future diffusion.

Content Features. We pay special attention to sentiment indicators from the content of tweets, including counts of emoticons, strength of word-level sentiments, and number of special signals. Special signals include words with repeated letters (e.g. goooooood), repeated exclamation marks (!!!) and repeated question marks (???).

Network Features. A social network can be constructed through retweets and mentions associated with a hashtag. Several graph features are extracted from this network of users, including graph order, graph density, average degree and entropy of degree distribution.

Hashtag Features. The length of the hashtag, the count of the hashtag with different cases and the co-occurrence frequency of hashtags are considered in this category.

Time Series Features. Coefficients of polynomial curve fitting and symbolic sequences are defined to represent the shape of the time series of the hashtag. Besides, some derivative features are defined to describe the characteristics of the time series, such as dormant period, mean value and standard deviation of the time series, mean value and standard deviation of the absolute first-order derivative, etc.

Prototype Features. Prototypes here refer to historical hashtags that are similar to the one to be predicted. Top-k prototypes for the hashtag to be predicted can be found from the historical data. For Task 1, the number of bursting hashtags among the top-k prototypes is used as a feature. For Task 2, top-k **bursting** prototypes are extracted, and their weighted average popularity is used as a feature. Considering k from 1 to 10, we can get 10 prototype features.

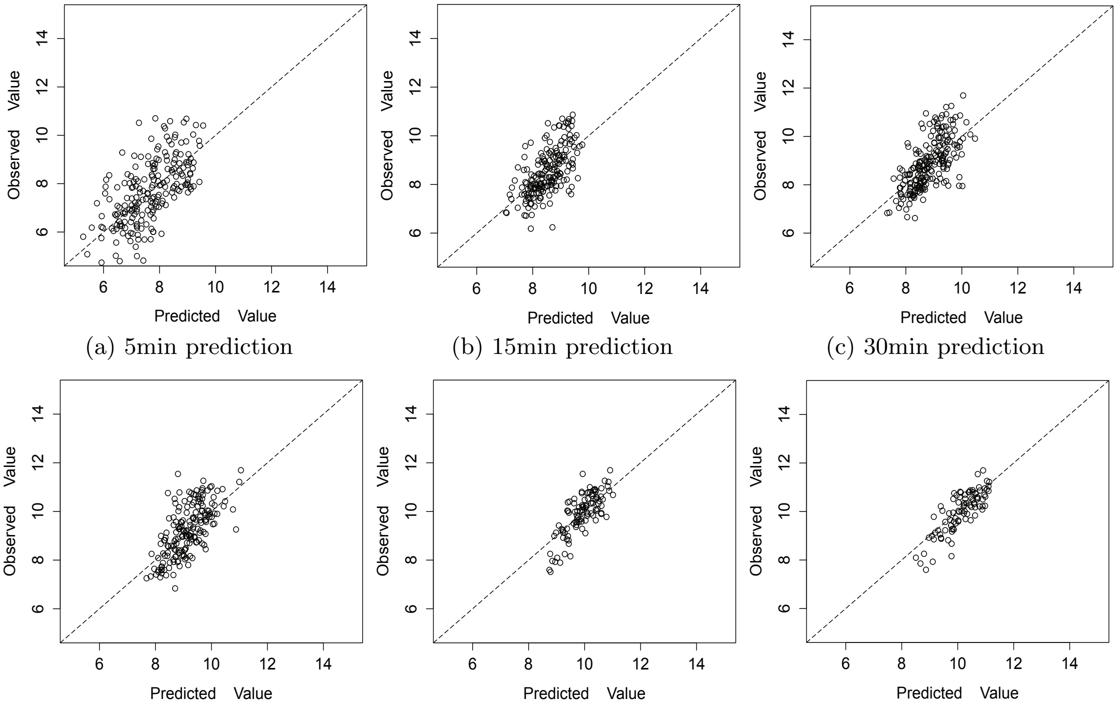
4. EXPERIMENTS

Predictions were made at six representative moments after a hashtag is marked as active, which can be divided into three stages, early stage (5min, 15min), middle stage (30min, 1h) and late stage (3h, 6h). Predictions after 6 hours since *active* were not considered because at that time only **5%** of bursting hashtags have not started bursting. Our data set consists of a three-month training set (2012.11-2013.1) and a one-month test set (2013.3). Accuracy is not a reasonable metric to evaluate the performance of Task 1. The classifier can easily yield high accuracy by predicting all items into the negative class. F_1-score, incorporating precision and recall, is used as the performance metric for Task 1. $RMSE$ (Root Mean Square Error), a common metric for regressions, is used as the metric to evaluate Task 2.

Our previous work [3] has already compared different algorithms for Task 1. Here we focus on evaluating the contribution of different types of features. We apply weighted SVM for a number of times, each time removing only one type of features. From Table 2, it can be observed that:

- Time series features are the most effective type of features. When they are removed, the drop of F_1-score is the most significant. The earlier in the life cycle of a hashtag, the less important the time series features. When they are removed, the F_1-score decreases by 72.41% if the prediction happens at 6 hours after the hashtag becomes *active*; if the prediction happens 5 minutes after *active*, however, removing time series features only results in a 5.64% drop.

- When the predictions are made early on (i.e., 5 minutes after the hashtag becomes active), the prototype features are the next most useful categories of features following time series features, and followed by network features and meme features. The inclusion of user features and content features actually hurts the perfor-



(a) 5min prediction (b) 15min prediction (c) 30min prediction

(d) 1h prediction (e) 3h prediction (f) 6h prediction

Figure 2: Predicted vs. Observed *log(Pop)*.

mance of early detection, while they become useful for predictions at a later stage.

- No matter which type of features are removed, the drop of F_1-score at the late stage is more sensitive than earlier stages. This is because the class distribution becomes so skewed at the late stage and different types of features have to be used together to train a good model.

For those bursting hashtags predicted correctly, how early are they predicted in advance of their bursts? According to the statistics of the results, when predicted at 5 minutes since *active*, the correct predictions happen on average **55 minutes** earlier than the actual bursts. And this value is larger for predictions at later stages of the life cycle. Therefore correct prediction of bursting hashtags is at least an average of 55 minutes earlier than the start of their bursts.

Table 3 shows the comparison of methods for Task 2. GPR achieves the best performance in most cases. Only when predicted 6 hours after becoming *active*, is SVR slightly better than GPR. Fig. 2 shows the comparison between predicted values and observed values of $log(Pop)$. It can be seen that the predicted values correlate the observed values very well.

Table 3: Comparison of algorithms for Task 2.

Method	*RMSE* (prediction at time t since active.)					
	5m	15m	30m	1h	3h	6h
LR	1.093	0.835	0.820	0.914	0.665	0.625
CART	1.138	0.846	0.960	0.941	0.662	0.667
GPR	**1.050**	**0.764**	**0.775**	**0.783**	**0.552**	0.506
SVR	1.091	0.776	0.789	0.792	0.555	**0.492**
NN	1.379	1.068	0.950	1.018	0.841	0.631

Acknowledgement

The work is supported by National Natural Science Foundation of China (61373022, 61073004, 60773156), and Chinese Major State Basic Research Development 973 Program (2011CB302203-2). It is also partially supported by the National Science Foundation under grant numbers IIS-0968489 and IIS-1054199, and partially supported by the DARPA under award number W911NF-12-1-0037.

5. REFERENCES

[1] G. H. Chen, S. Nikolov, and D. Shah. A latent source model for nonparametric time series classification. In *Advances in Neural Information Processing Systems*, pages 1088–1096, 2013.

[2] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.

[3] S. Kong, Q. Mei, L. Feng, and Z. Zhao. Real-time predicting bursting hashtags on twitter. In *Web-Age Information Management*. Springer, 2014.

[4] Y. R. Lin, D. Margolin, B. Keegan, A. Baronchelli, and D. Lazer. # bigbirds never die: Understanding social dynamics of emergent hashtag. *arXiv preprint arXiv:1303.7144*, 2013.

[5] Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 2013.

[6] S. Nikolov. *Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series*. PhD thesis, Massachusetts Institute of Technology, 2012.