Will This #Hashtag Be Popular Tomorrow?*

Zongyang Ma
zma4@e.ntu.edu.sg

Aixin Sun
axsun@ntu.edu.sg

Gao Cong
gaocong@ntu.edu.sg

School of Computer Engineering
Nanyang Technological University, Singapore, 639798

ABSTRACT

Hashtags are widely used in Twitter to define a shared context for events or topics. In this paper, we aim to predict hashtag popularity in near future (*i.e.,* next day). Given a hashtag that has the potential to be popular in the next day, we construct a hashtag profile using the tweets containing the hashtag, and extract both *content* and *context* features for hashtag popularity prediction. We model this prediction problem as a classification problem and evaluate the effectiveness of the extracted features and classification models.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

Keywords

Twitter, Hashtag, Popularity prediction, Hashtag clarity

1. INTRODUCTION

Micro-blogging services represented by Twitter are changing our way of accessing and disseminating information online. Because of the fast information disseminating rate, various topics emerge within a very short time period, particularly some public relations crises. This calls for effective prediction on what topics would emerge into significant ones in near future. As an important feature in Twitter, hastags (keywords prefixed with # symbols) are widely used to define a shared context for specific events, topics, or meme [1]. We therefore propose to predict hashtag popularity in near future, which could identify topics/events that might be critical for relevant organizations to response earlier.

Hashtag has attracted significant research interest. In [2], the authors analyze hashtags in eight categories (*e.g.,* politics, celebrity) for the differences in the mechanics of information diffusion. The evolution of hashtag popularity over time (*e.g.,* usage patterns before and after bursty peaks) is analyzed in [1]. Different from both studies which perform retrospective analysis of hashtags using historical data, we aim to predict hashtag popularity for newly appearing hashtags in a Twitter stream in the next day.

Most germane to this work is the recent paper on predicting hashtag popularity in weekly basis [3]. Our work is significantly different. First, we predict the *number of users* who adopt a hashtag, but not the *number of tweets* annotated by the hashtag in [3]. We therefore derive 8 features from the virtual community formed by users who adopt a hashtag; these context features are not used in [3]. Second, our content features are derived from *the content of the tweets* annotated by a hashtag whereas almost all content features in [3] are derived from *the hashtag itself* (*e.g.,* number of characters in a hashtag). Third, we target on time critical applications to predict hashtag popularity on daily basis rather than weekly basis, because most bursty hashtags are popular for a few days only [1].

To the best of our knowledge, we are the first to predict hashtag popularity in daily basis, using both content and context features. Additionally, we are the first to utilize clarity and topic modeling as content features for hashtag analysis.

*This work was partially supported by Singapore's National Research Foundation's research grant NRF2008IDM-IDM004-036, and Singapore MOE AcRF Tier-1 Grant RG13/10.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. HASHTAG POPULARITY PREDICTION

Problem Definition. In our problem setting, all tweets received from a Twitter stream are partitioned into consecutive fixed time-intervals by their time-stamps. The time-interval could be an hour, a few hours, or a day, depending on the number of tweets received as well as the time criticalness of the prediction. We define the *popularity degree* of a hashtag h in time-interval t, denoted by Φ_t^h, to be the number of users who post at least one tweet annotated by h within time-interval t. Given the collection of tweets published in t and a hashtag h of interest, our task is to predict Φ_{t+1}^h.

Note that predicting the *exact value* of Φ_{t+1}^h is extremely hard and is often not necessary. Therefore we relax the problem and predict the *range* of popularity degree. We define 5 ranges: $[0, \phi)$, $[\phi, 2\phi)$, $[2\phi, 4\phi)$, $[4\phi, 8\phi)$, and $[8\phi, +\infty)$, representing *not popular*, *marginally popular*, *popular*, *very popular*, and *extremely popular*, respectively. Note that the ranges are defined following an exponential order and ϕ is application-dependent.

Let T_t^h be the collection of tweets annotated by hashtag h in time-interval t. T_t^h is also known as *hashtag profile*. Next, we detail the content and context features extracted from a hashtag profile.

Content Features. The 6 content features extracted from T_t^h are listed in Table 1. We elaborate *hashtag clarity* and *hashtag topic vector* features in more detail. Hashtag clarity quantifies topical cohesiveness of all tweets in T_t^h. It is computed as the Kullback-Leibler (KL) divergence between the uni-gram language model inferred from T_t^h and the background language model from the entire tweet collection $\mathcal{T}$. If a hashtag refers to a specific topic, then the high probabilities of a few topic-relevant words distinguish its tweets from the background.

$$Clarity(T_t^h) = \sum_{w \in T_t^h} P(w|T_t^h) log_2 \frac{P(w|T_t^h)}{P(w|\mathcal{T})}$$

Hashtags in a similar topic (*e.g.,* political, music) may follow similar popularity trends [1]. We use Latent Dirichlet Allocation (LDA) to identify the topic distribution of a hashtag. Each hashtag profile T_t^h is considered as a document and 20 topics are inferred

Table 1: The 6 content features and 8 context features

Type	Description
Content Feature	(1) number of tweets in T_t^h, (2) fraction of tweets containing URL, (3) fraction of re-tweet, (4) fraction of tweets with mention "@", (5) hashtag clarity, and (6) 20-dimension hashtag topic vector.
Context Feature	(1) number of users $\|U_t^h\|$, (2) average authority of users, (3) density of G_t^h, (4) fraction of users forming triangles in G_t^h, (5) ratio between the number of connected components and the number of nodes in G_t^h, (6) average edge weights in G_t^h, (7) number of border users, and (8) 15-dimension exposure probability vector.

from all such documents in our data. A 20-dimension topic vector is then assigned to each hashtag profile with the entries quantifying the likelihood of the hashtag belonging to the corresponding topic.

Context Features. We first construct a directed weighted graph $\mathcal{G} = \langle U, E \rangle$ to model user relations in the data collection. In $\mathcal{G}$, a user $u \in U$ is a node and a directed edge $e(u_p, u_q) \in E$ from u_p to u_q is weighted by the number of times u_p mentions u_q in her tweets, similar to that in [2]. The authority scores of users are computed in this global user graph using PageRank algorithm.

We consider all users who post at least one tweet in hashtag profile T_t^h form a virtual community. By extracting their relations from the global user graph, we form a community graph $G_t^h = \langle U_t^h, E_t^h \rangle$. From G_t^h, 8 context features (see Table 1) are derived to capture the current popularity of h, the influential power of its users, the connectivity among these users, and the distribution of the users exposed to h. We elaborate the last two features in Table 1.

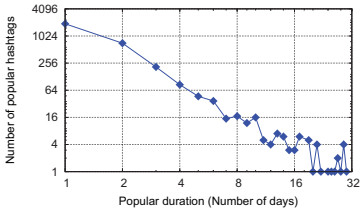
Based on the global graph, border users are those each has at least one edge from users in G_t^h but has not adopted hashtag h, *i.e.,* $\{u_q | \exists e(u_p, u_q), u_p \in U_t^h, u_q \notin U_t^h\}$. Exposure probability vector depicts border user distribution in a more detailed manner: $P(k)$ $(1 \leq k \leq 15)$ is the ratio of border users who have k edges from users in G_t^h. A border user with more exposure to a hashtag is expected to be more likely to adopt the hashtag.

3. EXPERIMENTS

Experimental Setting. We used tweets published by Singapore-based users from Jan to Aug 2011. The dataset consists of more than 31M tweets from over 2M users. The global user graph constructed based on mention relation consists of 214K users and 680K edges. Note that, only users participated in mentions are included in this graph. In our experiments, we set the time-interval to be *a day* and set $\phi = 25$. That is, a hashtag used by fewer than 25 users in a day is considered not popular.

A key issue in the experiments is that, among all hashtags appeared at least once in a day, which hastags should be selected for popularity prediction. Figure 1(a) plots the number of popular tags $(\Phi_t^h \geq \phi)$ against their popularity duration in number of days. Observe that a large number of tags are popular for only a day. In fact, a much larger number of tags has never been popular. In our experiments, we therefore choose to predict the popularity degree of *newly appearing hashtags* that are *at least marginally popular*, $\Phi_t^h \geq \phi$. A hashtag is considered *new* if it has not gained marginal popularity in the past 7 days. Note that, the same hashtag may gain popularity in different time periods for different reasons. For example, #apple may gain popularity at different time periods for releasing different products.

We conducted experiments using five commonly used classification methods: Naïve Bayes (NB), Decision Tree (C4.5), k-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Maximum Entropy (MaxEnt). Their prediction accuracies are compared with three baseline methods, which do not use any extracted features.



(a) Popular tag vs popular duration

Popularity	No. hashtags
[0, 25)	1609
[25, 50)	614
[50, 100)	353
[100, 200)	211
≥200	111
Total	2898

(b) Hashtag distribution

Figure 1: Popular tag vs duration; and distribution of instances

Table 2: Hashtag popularity prediction accuracy by 8 methods

Method	Micro-F_1	Macro-*Pr*	Macro-*Re*	Macro-F_1
Random	0.197	0.198	0.205	0.163
Lazy	0.254	0.251	**0.450**	0.317
PriorDist	0.385	0.209	0.210	0.209
NB	0.397	0.341	0.390	0.338
KNN	0.500	0.426	0.419	**0.422**
C4.5	0.527	0.413	0.408	0.409
SVM	0.582	0.401	0.329	0.296
MaxEnt	**0.591**	**0.437**	0.387	0.382

The baseline methods are Random (predicts the popularity range of Φ_{t+1}^h randomly), Lazy (predicts the range of Φ_{t+1}^h the same as Φ_t^h), and PriorDist (predicts the range of Φ_{t+1}^h randomly following a prior probability distribution in the 5 ranges).

Results. We conducted 10-fold cross validation and evaluated prediction accuracy using Micro-F_1, Macro-*Pr*ecision, Macro-*Re*call and Macro-F_1. Note that, because each instance has exactly one correct label, Micro-*Pr*ecision/*Re*call is the same as Micro-F_1.

Table 2 reports the prediction accuracies by the 8 methods, in increasing order of Micro-F_1. All baseline methods perform the worst. Observe that MaxEnt achieves the best Micro-F_1 of 0.591 which is triple of Random and 54% of increment over the best baseline PriorDist. SVM is the second best performing method by Micro-F_1; MaxEnt also achieves the best Macro-*Pr* followed by KNN. Surprisingly, Lazy prediction yields the best Macro-*Re*. The main reason is the skewed distribution of our categories (see Figure 1(b)). Because of the small number of instances in *very popular* and *extremely popular* ranges, most classifiers fail to learn effective patterns for accurate prediction. Lazy prediction enjoys high accuracy mainly in these two ranges. By Macro-F_1 KNN is the best performing method.

Our analysis on features reveals that context features are relatively more effective than content features (detailed results not reported due to page limit). The top-5 most effective features are: (1) number of users, (2) number of border users, (3) number of tweets, (4) hashtag clarity, and (5) fraction of users forming triangles.

4. SUMMARY

In this paper, we propose to predict hashtag popularity to identify fast emerging topics attracting collective attention. Our preliminary results demonstrate the effectiveness of both content and context features. In our future work, we will continue to investigate other possible features, and more importantly to investigate effective prediction models specifically for hashtag popularity prediction.

5. REFERENCES

[1] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *WWW*, pages 251–260, 2012.

[2] D. M. Romero, B. Meeder, and J. M. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW*, 2011.

[3] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *WSDM*, pages 643–652, 2012.