On Predicting Twitter Trend: Factors and Models

Peng Zhang
Computer Science and Eng. Dept.
Arizona State University
Tempe, USA
Pzhang41@asu.edu

Xufei Wang
Computer Science and Eng. Dept.
Arizona State University
Tempe, USA
Xufei.wang@asu.edu

Baoxin Li
Computer Science and Eng. Dept.
Arizona State University
Tempe, USA
Baoxin.li@asu.edu

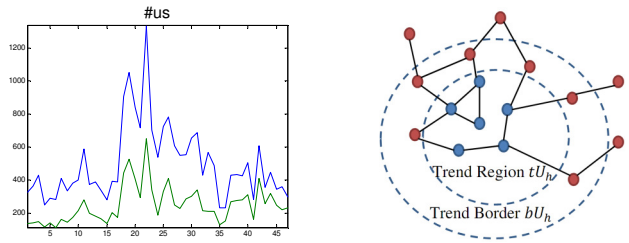
***Abstract*—In this paper, we predict hashtag trend in Twitter network with two basic issues under investigation, i.e. trend factors and prediction models. To address the first issue, we consider different content and context factors by designing features from tweet messages, network topology, user behavior, etc. To address the second issue, we adopt prediction models that have different combinations of the two basic model properties, i.e. linearity and state-space. Experiments on large Twitter dataset show that both content and context factors can help trend prediction. However, the most relevant factors are derived from user behaviors on the specific trend. Non-linear models are significantly better than their linear counterparts, which can be further slightly improved by the adoption of state-space models.**

Keywords— information diffusion, trend prediction, Twitter.

I. INTRODUCTION

Information diffusion is the process of propagation through network links. Being able to predict or simulate such a process may lead to many applications in, e.g. politics, economics [1]. In this work, we predict information diffusion on macro level as information trends of some underlying topic. We use Twitter network as a case study and focus on the prediction of ***hashtag trend*** which is a set of tweets grouped by a hashtag, i.e. a string starting with the character #, to represent a topic, event, etc. A hashtag trend is measured by the number of users and tweets involved in each time interval. Different hashtag trends may evolve with different patterns due to many relevant factors interacting in complex dynamics. Two basic issues lie in the effective prediction of hashtag trend (or information trend): relevant ***trend factors*** and appropriate ***prediction models***.

Current research of information trend prediction regarding the above two issues are reviewed below together with our efforts in this papers. On one hand, many relevant trend factors have been identified, which can generally be categorized into two categories, i.e., context and content factors. ***Content factors*** describe the content of trend through lexical, semantic, and sentimental analysis. For example, LDA topic distribution [2] is used to predict hash-tag trend. There are also simple content features, such as the fraction of tweets containing URL [2], the fraction of retweet/mention in a trend [3]. ***Context factors*** generally describe the network environment, e.g. density and centrality. User behavior is also recognized as important context factors, e.g. retweet ratio [3]. Despite these findings, most existing works use only one type of factors for prediction except only two recent work [2] [4]. In this paper, we combine all these factors for trend prediction and further discuss their relevance for better understanding of the topic. On the other hand, existing methods typically only use simple (non-)linear regression or classification models [2] [4], which are in general inadequate for handling complex trend dynamics on large-scale social networks. In this paper, we investigate several type of prediction models which are the different combination of two basic model properties, i.e. (non-)linearity and (non-)state-space modeling. The validity analysis of the features and models is evaluated on a large Twitter dataset.



(a) (b)

Fig. 1. (a) Illustration of hashtag trends. The horizon axis is time, and vertical axis is count. The blue/green line is the number of tweets/ users respectively. (b) Illustration of trend users (tU_h) and trend border users (bU_h). The blue (red) dots are users have (not) adopt the hashtag h.

II. HASHTAG TREND PREDICTION

In this paper, uppercase letters U, T and their variations by prefix and subscripts represent a set of users and tweets respectively. The lowercase u is for a user, h is for a hashtag, and t is for a time (interval). For example, $T_h(t)$ is the set of tweets with hashtag h posted in time t, and $U_h(t)$ is the users who post a tweet in $T_h(t)$. Let $tU_h(t)$ be ***trend user*** of h in t as the set of users already adopted h at that time, i.e. $tU_h(t) = \bigcup_{\tau=1}^{t} U_h(\tau)$. Then, ***trend border*** $bU_h(t)$ are the followers of $tU_h(t)$ who still have not adopted h (Fig. 1 (b)).The notation of set cardinality is $|\cdot|$. Time index t is often omitted for brevity when no confusion arises. The ***trend popularity*** of h under prediction is $[\log(|T_h(t)|), \log(|U_h(t)|)]$.

A. Trend Factors

The content and context factors for trend prediction are summarized in Table I. Details are given below.

Content factors. Since trend popularity contains both tweet $T_h(t)$ and user $U_h(t)$, our trend factors also include the two aspects. In fact, the two sets can be divided into subsets from different views as different content features.

The $U_h(t)$ can be split into three subsets based on the role of users. The first subset ($oU_h(t)$) consists of old users

rejoining trend h on t, i.e. $U_h(t) \cap tU_h(t-1)$ (Fig. 1(b)). The remaining two groups are new users with different relation to $tU_h(t-1)$. Specifically, the second subset ($fU_h(t)$) are either the followers of $tU_h(t-1)$ or users retweeting from $tU_h(t-1)$. The third subset ($sU_h(t)$) are 'self-motivated' users who either publish trend tweet by themselves or retweet from users outside our network (we only have part of the entire Twitter network). In fact, $oU_h(t)$ implies the trend's consistency by the action of old users; $fU_h(t)$ indicates the trend's propagation from trend users; $sU_h(t)$ suggests the trend's growth from the information outside the networks.

We can also split $T_h(t)$ into $oT_h(t)$, $fT_h(t)$, and $sT_h(t)$ which are the tweets produced by $oU_h(t)$ or $fU_h(t)$ or $sU_h(t)$ respectively. Another division of $T_h(t)$ is by the type of tweets: retweet ($rtT_h(t)$) as the propagation of the old trend messages; mention set ($mT_h(t)$) as the discussion among trend users; the set remains ($nT_h(t)$) as tweets of new information. The three subsets play different roles in trend diffusion. The $rtT_h(t)$ shows the propagation of h by retweet; $mT_h(t)$ indicates stickiness of h by the discussion with among users; $nT_h(t)$ implies new information injected into trend h. The subset with URL ($urlT_h(t)$) is also considered, since they are more likely to be retweeted due to the limited characters in a tweet.

Context factors describe the network environment which consist of both link topology (Structure context) and node attribute (Node context). For efficiency and relevancy, only two user sets (Fig. 1(b)) are considered: tU_h as the direct producers of h; bU_h as the direct reader of h.

Structure context factor. Three sub-graphs are considered: the sub-graph of tU_h or bU_h, and the bipartite graph between bU_h and tU_h which contains most pipelines of propagation to new users. For each of these sub-graphs, the network topology is described by three factors, i.e. centrality (in-degree), density, and reciprocity (c.f. 11~17, Table I).

Node context factors describe the network users by their actions and profiles. Besides the features [3] (c.f. 26~29, Table I) on ***interaction***, i.e. retweet and mention, as users' influence or will to propagate a trend, two other types of features are designed as users' information input/output, i.e. activeness and stimulus. ***Activeness*** is a user's frequency to generate information. The ***general activeness*** of user u in time t is:

$$act(u;t) = \alpha \cdot act(u;t-1) + |T(u;t)|\,, \qquad (8)$$

where α is the decay coefficient and $T(u;t)$ is the set of tweets posted by u in t. ***Trend activeness*** ($act_h(u;t)$) is similar to (8) by replacing $T(u;t)$ with $T_h(u;t)$, i.e. the subset of $T(u;t)$ with hashtag h. The $act_h(u;t)$ implies the participation of u in h, and $act_h(u;t)/act(u;t) \in [0,1]$ reflects u's interest on h. ***Stimulus*** is the volume of information received by a user. The ***general stimulus*** of user u in time t is:

$$stim(u;t) = \beta \cdot stim(u;t-1) + \sum_{u_i \in friend(u)} |T(u_i;t)|, \quad (9)$$

where β is the decay coefficient, $friend(u)$ are the friends of u, and $T(u;t)$ is same as (8). ***Trend stimulus*** ($stim_h(u;t)$) is similar to (9) by replacing $T(u;t)$ with $T_h(u;t)$ as the subset of $T(u;t)$ with hashtag h. The $stim_h(u;t)/stim(u;t) \in [0,1]$ reflects u's specific attention on trend h.

TABLE I. SUMMARY OF CONTENT AND CONTEXT TREND FEATURES.

Index	Description
1~3	Portion of rtT_h (retweet) or mT_h (mention) or nT_h (new) in T_h
4	Portion of $urlT_h$ (URL) tweets in T_h as $\|urlT_h\|/\|T_h\|$
5~7	Portion of oU_h (old) or fU_h (followers) or sU_h (self-motivated) in U_h
8~10	Portion of oT_h or fT_h or sT_h (tweets from oU_h or fU_h or sU_h) in T_h
11	Ratio between border and trend users: $\|bU_h\|/\|tU_h\|$ (Fig. 1(b))
12~13	Max/Average of out-degree prestige of trend user tU_h. (Fig. 1(b))
14	Density [5] of the sub-graph formed by tU_h.
15	Density [5] of the bipartite graph between $tU_h$and bU_h.
16	Reciprocity [5] of the sub-graph formed by tU_h.
17	Reciprocity [5] of the bipartite graph between bU_h and tU_h.
18~19	Average general activeness ($act(u;t)$) over tU_h or bU_h. (See (8))
20	Average trend activeness ($act_h(u;t)$) over users in tU_h. (See (8))
21	Average of $act_h(u;t)/act(u;t)$ over tU_h.
22~23	Average trend stimulus ($stim_h(u;t)$) of tU_h or bU_h. (see (9))
24~25	Average of $stim_h(u;t)/stim(u;t)$ over tU_h or bU_h.
26~27	Percentage of interaction tweets of tU_h or bU_h up to time t.
28~29	Interaction received by tU_h or bU_h divided by all their tweets up to t

Note. In this table, time index t is usually ignored for brevity.

B. Prediction Models

A prediction model can be characterized by two properties, i.e. (non-)linearity and (non-)state-space, whose combination leads to four categories. In each category, we test one typical model as the variations to the models in other categories. We start with a linear non-state-space ***ARX*** [6] model whose direct state-space extension is linear dynamic system (***LDS***) [6]. The non-linear ARX (***NARX***) can be implemented by a feed-forward neural network, whose state-space extension is recurrent nonlinear ARX (***RNARX***) [7]. Usually, non-linear models can describe more complex dynamics, and state-space models have a better memory of history. All models are summarized in Table III, together with two baseline methods. The first is ***Last Predictor*** which predicts by the last value, i.e. $\hat{\mathbf{y}}(t+1) = \mathbf{y}(t)$. The second is ***Mean Predictor*** which predicts by the mean up to t, i.e. $\hat{\mathbf{y}}(t+1) = t^{-1} \cdot \sum_1^t \mathbf{y}(t)$.

III. EXPERIMENTS AND DISCUSSION

Our Twitter dataset is collected for the purpose of *Arab Spring*. The collection involved manually defined hashtags and geographic regions related to the following countries: Egypt, Libya, Syria, Bahrain and Yemen. We collected 16.1 million Tweets by 0.67 million users from February 1, 2011 to August 31, 2011, which amounts to approximately 10% of the all Tweets hosted by Twitter following the above hashtag and geographic constraints. We select the 336 most popular hashtags with at least 5000 related tweets for trend prediction task, which sums up to 28.3 million tweets (> 16.1million because a tweet can has more than one hashtag).

We evaluate prediction by Mean Square Error (***MSE***). All prediction results in this section are estimated by 10-fold cross-validation. The features in Table I are normalized before feed into prediction models. The decay coefficients α and β for activeness (8) and stimulus (9) are 0.1.

Relevance of Trend Factors is analyzed by random forest (***RF***) which use the trend factors and popularity measures in

the past 5 days to predict trend (similar result in longer days). The mean variable importance [8] of each factor over the 5 days for user ($U_h(t)$) prediction are presented in Fig.2. Similar result can be found for $T_h(t)$, and omitted here due to limited space. Fig. 2 leads to the following points. First, node context factors on user behaviors are more important, e.g. trend stimulus/activeness. Second, the factors with specific to the information trend behavior are more important. For example, the trend stimulus/activeness is derived from user's trend related action, and factors on the topological structures between trend user and border are also important.

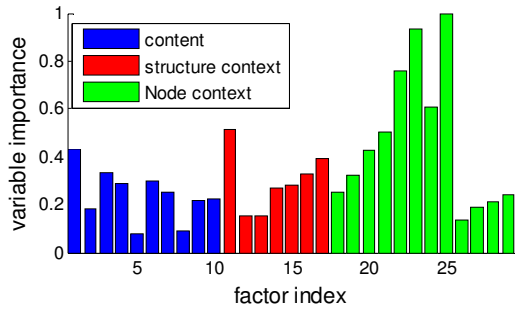


Fig. 2. The variable importance (normalized to [0,1]) of different trend factors for user ($U_h(t)$) prediction. The factor index is the same as in Table I.

TABLE II. MSE OF PREDICTION RESULT FROM DIFFERENT TREND FACTORS.

Factors	T_h(t): Mean (Std)	U_h(t): Mean (Std)
None	0.140 (0.009)	0.155 (0.008)
Content	0.136 (0.012)	0.151 (0.010)
Structure Context	0.134 (0.010)	0.147 (0.009)
Node Context	0.131 (0.008)	0.146 (0.009)
All	**0.130 (0.010)**	**0.142 (0.008)**

Note: All/None means prediction with all/no trend factors.

We also compare trend factors by the prediction result with the same RF model in Fig.2. From Table II, we have the several points. First, trend factors really help trend prediction, because the MSE with all trend factors is significantly better than that without any factors. Second, the MSE of node context, structure context, and content factors coincide with the variable importance in Fig. 2. However, the performance gap is not significant, which might because these factors are complementary and have importance on their own aspects.

Comparison of Prediction Models is summarized in Table III, where models are trained with all the features in Table I. For each model, we tried several setups and present the best result. For ARX, RF, and NARX model, we go through order 1 to 5. For LDS we check the order up to 4. For both the NARX and RNARX, the network layers vary from 1 to 2, and the hidden node number varies from 2 to 10.

The following observations come from Table III. First, the hashtag trend is predictable on some degree, for the prediction result of all models are significantly better than the baseline methods. Second, nonlinearity is critical for prediction, which is clear from the performance gap between linear models, i.e. ARX and LDS, and their non-linear counterparts, i.e. NARX and RNARX. Third, state-space is helpful but not as important as nonlinearity. In fact, the performance gap between ARX and LDS is not significant. The situation is similar for RF and RNARX. The slight improvement may due to the fact that long history might not be as important for Twitter trend. In fact, for Non-state-space models, i.e. ARX and RF, the performance will not change significantly with more than 5-day history. Another reason is that state-space models with gradient decent training might not be competent at the complicated cost function surface of high dimensional feature space. Further feature selection or regularization might help.

TABLE III. BEST PREDICTION MSE OF DIFFERENT MODELS.

Type	Model	T_h(t): Mean (Std)	U_h(t): Mean (Std)
L-NS	ARX	0.151 (0.009)	0.169 (0.006)
L-S	LDS	0.149 (0.009)	0.166 (0.007)
NL-NS	NARX	0.133 (0.007)	0.146 (0.008)
	RF	0.130 (0.010)	0.142 (0.008)
NL-S	RNARX	**0.128 (0.008)**	**0.139 (0.009)**
Baseline	Last	0.175 (0.015)	0.198 (0.018)
	Mean	0.219 (0.017)	0.235 (0.016)

Note: **L** and **NL** is short for Linear and Non-Linear; **S** and **NS** is short for State-space and Non-State-space. **Last** and **Mean** is two baseline predictors.

IV. CONCLUSION AND DISCUSSION

In this paper, we study the two basic problems in trend prediction, i.e., important factors and appropriate models, with focus on Twitter network. We investigate different factors on both tweet content and network context for trend prediction. We also compare prediction models as the combination of two basic model properties, i.e., (non)linearity and (non-)state-space modeling. We report some insightful findings from comparative experiments on large Twitter dataset. Future work includes the following two directions. First, the semantics and sentiments of a trend is relevant to its diffusion process, so further investigation on feature design is worthwhile. Second, a network may change over time and thus adaptive models may be needed to account for this issue.

ACKNOWLEDGMENT

The work was supported in part by a grant (#1135616) from the National Science Foundation. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Sheng Yu and Subhash Kak, "A Survey of Prediction Using Social Media," arXiv:1203.1647, 2012.

[2] Zongyang Ma, Aixin Sun, and Gao Cong, "Will this #hashtag be popular tomorrow?," in *SIGIR*, 2012, pp. 1173-1174.

[3] Jiang Yang and Scott Counts, "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter," in *ICWSM*, 2010

[4] O. Tsur and A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in *WSDM*, 2012, pp. 643-652.

[5] M. E. Newman, *Network: an introduction*.: Oxford Press, 2010.

[6] Lennart Ljung, *System Identification: Theory for the User (2nd edition)*.: Prentice Hall, 1998.

[7] Simon Haykin, *Neural Networks and Learning Machines, 3rd edition*.: Prentice Hall, 2008.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*.: Springer, 2009.