On Predicting Twitter Trend: Important Factors and Models

Peng Zhang, Xufei Wang, Baoxin Li
Arizona State University
Baoxin.Li@asu.edu

Abstract

Trend prediction for social media has become an important problem that can find wide applications in economics, social studies, etc. In this paper, we investigate two basic issues in trend prediction, i.e., what are the important factors and what may be the appropriate models. To address the first issue, we consider different content and context factors by designing features from tweet messages, network topology, and user behavior, etc. To address the second issue, we investigate several prediction models that have different combination of the two basic model properties, i.e. (non-)linearity and (non-)state-space modeling. Our study is based on the hashtag trend of a large Twitter dataset with more than 16M tweets and 660k users. We report some insightful findings from comparative experiments. In particular, it is found that the most relevant factors are derived from user behavior on information trend and that non-linear state-space models are more effective for trend prediction.

Introduction

Information diffusion is a network process in which information propagates through network links. Being able to predict or simulate the outcomes of such a process may lead to many applications in social studies, e.g. economics, politics. (Yu & Kak, 2012). In this work, using Twitter network as a case study (Kwak, et al., 2010), we investigate the problem of predicting Twitter trends, which measure, at macro level, information diffusion regarding some underlying topic or event. Twitter is an interactive social media platform where users share ideas and communicate by tweets, messages of less than 140 characters. The term *trend* in Twitter refers to the dynamics of a set of tweets grouped by a hashtag, which is a string of characters starting with the character #, to represent a topic, event, etc. For example, #icwsm13 is the hashtag for the topic on the ICWSM 2013 conference. The *popularity* of a twitter trend is measured by the number of users and tweets involved in the trend or hashtag. So our prediction object is: given the history of a hashtag trend, how many users and tweets will appear in the next time interval?

Different twitter trends may evolve with different patterns (Figure 1) due to many factors. Therefore, effective trend prediction should consider both the design of *trend factors* and the selection of *prediction models*. Although trend prediction has been studied in other related fields, e.g., time series analysis and system identification (Ljung, 1998), many important questions still need to be answered in the context of social networks. For example, trend diffusion is complex due to the interaction among a large set of network nodes, which requires more sophisticated prediction model. Also, how to exploit additional information channels in social networks such as user activities should be investigated.

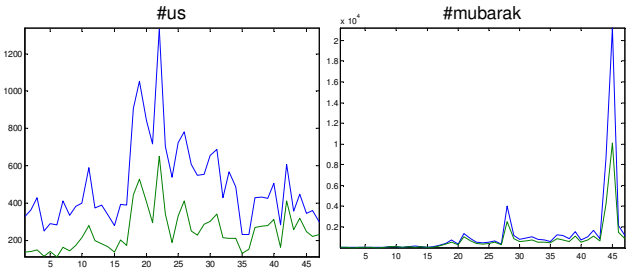

Figure 1: An illustration of Twitter trends. The horizon axis is time, and the vertical axis is count. In each plot, the blue/green line is the number of tweets/users respectively.

Current tweet trend research mainly focuses on the investigation of relevant trend factors. On the one hand, existing methods typically only use simple (non-)linear regression or classification models, which are in general inadequate for handling sophisticated trend dynamics on large-scale social networks. On the other hand, many relevant trend factors have been identified, which can generally be categorized into two categories, i.e., context and content factors. Content factor is derived lexically from tweets for describing the trend topic, quality, etc. Context factor is derived from network structure and user behavior to describe the environment for trend diffusion. Most prior works only predict trend with one type of factors, except two recent papers (Ma, et al., 2012) (Tsur & Rappoport,

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2012), which, however, covered only limited context and content factors.

In this paper, we investigate two basic issues in trend prediction, i.e. what are the important factors and what may be the appropriate models. To address the first issue, we consider different content and context factors by designing features from the body of the tweets, the network topology, and user behavior, etc. To address the second issue, we investigate several prediction models by considering different combination of the two basic model properties, i.e. (non-)linearity and (non-)state-space modeling. The analysis of the features and models is done through experiments with a large Twitter dataset,

The main contributions and findings of this paper are as below. First, a comprehensive study on different aspects of trend factors on a large, real dataset was conducted. Second, different categories of prediction models, covering (non-)linear and (non-)state-space models, were investigated and comparatively evaluated using the same dataset. The analysis of the experimental results suggests that, for trend prediction on the Twitter network, context factors based on user behavior is most useful, and nonlinear state-space models appear to be more effective.

Related Work

Since this paper is about the two basic issues of trend prediction, the related works also lies in two aspects.

Relevant Trend Factors

Current research has identified many factors relevant to information trend diffusion. These factors can be generally divided into 2 categories, i.e. content and context. Content factor describes the information trend content by lexical analysis. For example, it is recognized in (Lehmann, et al. 2012) and (Romero, Meeder and Kleinberg 2011) that trend may follow different temporal pattern across its content topic, e.g. politics, sports, and the LDA topic distribution is used to predict trend (Ma, Sun and Cong 2012). The TF-IDF similarity between the information trend and user's interest is also relevant (Yang and Leskovec 2011). There are also simple content features, such as the fraction of tweets containing URL (Ma, Sun and Cong 2012), the fraction of retweet/mention in a trend (Yang and Counts 2010), and hashtag itself (Tsur and Rappoport 2012).

Context factor describes the diffusion environment of the information trend. On the one hand, network topology is shown to be related to the scale and speed of trend diffusion, e.g. the network density (Lerman and Ghosh 2010) and the border of sub-graph formed by users already adopt the trend (Romero, Meeder and Kleinberg 2011). On the other hand, the importance of user's behavior is also recognized, e.g. the retweet ratio and mention ratio of information trend contributors (Yang and Counts 2010).

There are also works of information diffusion on micro level as trend adoption behavior of individual users, e.g. retweet (Lerman, et al. 2012) or hashtag adoption (Yang, Sun and Mei 2012). The findings are very similar to that of macro information trend and they also provide inspirations on macro trend factors. For example, we generalize the attention limit of user (Lerman, et al. 2012) to design user stimulus as a context factor for trend prediction (Table 1).

Despite these relevant trend factors, most current works only predict trend with one type of factors, except the two recent papers (Ma, Sun and Cong 2012) (Tsur and Rappoport 2012). However, they only covered several limited context and content factors. In this paper, different factors are combined together for trend prediction, and their importance is comprehensive discussed.

Trend Prediction Models

The most popular type of trend prediction method is regression/classification (Ma, Sun and Cong 2012) (Tsur and Rappoport 2012). Our work belongs to such category. However, most current works only use simple linear regression model. As a result, it lost the temporal history which is proved to be important for trend prediction (Yang and Leskovec 2011). In this paper, we also consider dynamic state-space model where the influence of temporal history is accumulated into latent state variables.

Another popular method is to model the information trend by differential stochastic equations (Szabo and Huberman 2010) (Matsubara, et al. 2012). The advantage of these methods is that the dynamic mechanism of trend is explicit. However, there are too many relevant trend factors, so the stochastic equation is always based on some assumption or simplification of major factors.

There are also studies on global temporal pattern of information trend by clustering (Yang and Leskovec 2011) (Lehmann, et al. 2012). Then a classifier is assigned to each cluster based on content or context features. These methods can predict the trend in a whole range rather than the next time interval. However, it is not easy to update the prediction with the upcoming observations. Also, it will not provide any dynamic mechanism on information trend.

Notations and Problem Statement

Twitter is a graph $G = \langle U, E \rangle$ where users U are connected to each other by links E. A link $E_{i,j}$ from user u_i to u_j means u_i is a *follower* of u_j, and u_j is a *friend* of u_i. The total friends of u_i is its out-degree; the total followers of u_j is its in-degree. Each user's tweet will be broadcasted to all

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

his followers, so links are the basic information diffusion pipeline. Twitter users can also communicate with each other via retweet and mention as the main cause of information diffusion. *Retweet* is identified by the use of RT @username or via @username in a tweet; *Mention* is the use of @username if it is not a retweet.

In this paper, uppercase letters U, T, E and their variations by prefix and subscripts represent a set of users, tweets, and links respectively. While lowercase letter 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 set of users contribute to $T_h(t)$. Let $tU_h(t)$ be *trend user* as set of users already adopted trend h before t, and *trend border* $bU_h(t)$ is the followers of $tU_h(t)$ who still have not adopt h before t (Figure 2).The notation of set cardinality is $|\cdot|$. Time index t is often omitted for brevity when no confusion arises.

In this paper, our prediction object $\boldsymbol{O}_h(t)$ is the *popularity measure* of trend h as the number of tweets and users:

$$\boldsymbol{O}_h(t) = [\log(|T_h(t)|), \log(|U_h(t)|)], \quad (1)$$

where the logarithm is to compress the large dynamic range of trend popularity. Let $\boldsymbol{I}_h(t)$ be the vector of relevant trend factors. Then $\boldsymbol{O}_h(t)$ is estimated by a prediction model $\mathcal{F}(\cdot\,|\boldsymbol{\theta})$as:

$$\widehat{\boldsymbol{O}}_h(t+1) = \mathcal{F}(\boldsymbol{O}_h(1:t), \boldsymbol{I}_h(1:t)|\boldsymbol{\theta}), \quad (2)$$

where $\boldsymbol{\theta}$ is model parameter vector, and $\boldsymbol{O}_h(1:t)$ and $\boldsymbol{I}_h(1:t)$ are the popularity measures and trend factors from the beginning time 1 to current t.

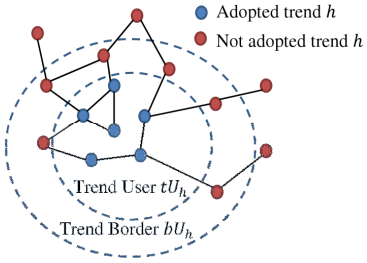

Figure 2. Illustration of trend user tU_h and trend border bU_h.

Relevant Trend Factors

The relevant trend factors can be generally grouped into 2 categories, i.e. content and context factors. In this section, we will and discuss the specific feature design for these factors which are summarized in Table 1.

Content factors

Content factors are extracted from the information trend itself to describe its topic, quality, character, etc. In this paper, we only focus on simple content features, e.g. number of retweet and mention. Despite of its simplicity, such content factors are proved to be useful for trend prediction (Asur, et al. 2011). Although deep content factors, e.g. trend topic, are also relevant to trend diffusion (Romero, Meeder and Kleinberg 2011), it is hard to find appropriate features. Traditional semantic analysis, e.g. LDA topic distribution, only leads to less effective features than the simple content factors (Ma, Sun and Cong 2012) . So we left the design of deep content factors as an open problem for future investigation.

Since trend popularity (1) is defined on both the tweet set $T_h(t)$ and user set $U_h(t)$ of current time interval t, our relevant trend factors include the two aspects. In fact, the two set can be categorized from different view which leads to the different content features below.

One simple division of $T_h(t)$ is by twitter types:

$$T_h(t) = rtT_h(t) + mT_h(t) + nT_h(t) \quad (3)$$

where $rtT_h(t)$ is the retweet set as the propagation of old trend messages, $mT_h(t)$ is the mention set as the discussion among trend uses, and $nT_h(t)$ is the remaining 'new' tweets set as new information injected into trend h. So the three subsets may play different row in trend diffusion.

Another division of $U_h(t)$ is based on the role of user's contribution to current trend h:

$$U_h(t) = oU_h(t) + fU_h(t) + sU_h(t), \quad (4)$$

The meaning of (4) is explained below. Then, $oU_h(t)$ is the set of active 'old' users as the intersection between $tU_h(t)$ and $U_h(t)$; $fU_h(t)$ is either the followers of $tU_h(t)$ or users retweeting trend twitter from $tU_h(t)$; $sU_h(t)$ is 'self-motivated' users who either publish trend tweet by themselves or retweet from users outside our network (since we can only crawl subset of the entire Twitter network). In fact, $oU_h(t)$ is the trend reproduce in $tU_h(t)$; $fU_h(t)$ is the information diffusion from $tU_h(t)$; $sU_h(t)$ new injection to the trend from outside source.

The $T_h(t)$ can also categorized following (4) as:

$$T_h(t) = oT_h(t) + fT_h(t) + sT_h(t), \quad (5)$$

where $oT_h(t)$, $fT_h(t)$, and $sT_h(t)$ are the set of tweets posted by $oU_h(t)$, $fU_h(t)$, and $sU_h(t)$ respectively.

Due to the limited characters of Tweets, the existence of URL is also a very important content measure. So the subset of $T_h(t)$ with URL, i.e. $urlT_h(t)$, is also relevant.

Context factors

Context factors describe the network environment in which information trends are diffusing. A network consists of nodes (user) and linkages (structure), which leads to the two aspects of context factors below.

Structure context factors are all kind of metrics describing the topological structure of network, e.g. network density, centrality, transitivity, similarity (Newman 2010). Since trend users tU_h are information trend producer, it is nature that the structure of sub-graph formed by tU_h is very relevant to trend diffusion. Also, trend border bU_h as followers of tU_h is very important since they are the users

directly exposed to the information trend (Romero, Meeder and Kleinberg 2011). Therefore, the topological structure of sub-graph bU_h and its relationship to tU_h are also used for trend prediction. The structure context feature to be extracted can be divided into 3 categories. First, the prestige (centrality) of tU_h is the user's influence for information diffusion based on network structure. Second, the network property of both the sub-graph of tU_h and bU_h, such as density. The *density* of directed graph $G = \langle U, E \rangle$ is:

$$density(G) = |E|/\big(|U| \cdot (|U| - 1)\big), \quad (6)$$

which is the number of links divided by all possible links. Third, the tie-strength between tU_h and bU_h, and our choice is *reciprocity* defined as the portion of co-links:

$$reciprocity(G) = |coLink(E)|/|E|, \quad (7)$$

where $coLink(E) \subset E$ is the subset of co-links. Following the above ideas, the density and reciprocity of bipartite graph can be defined similarly.

Table 1. Summary of Content and Context trend factors.

Ind.	Symbol	Description
Content Factor		
1	$prop.\text{rt}T_h$ [1]	Retweet proportion: $\|rtT_h\|/\|T_h\|$; see (3)
2	$prop.\text{m}T_h$	Mention proportion: $\|mT_h\|/\|T_h\|$; see (3)
3	$prop.\text{n}T_h$	New tweet proportion: $\|nT_h\|/\|T_h\|$; (3)
4	$prop.\text{url}T_h$	URL tweet proportion: $\|urlT_h\|/\|T_h\|$
5	$prop.\text{o}T_h$	Tweet proportion by oU_h: $\|oT_h\|/\|T_h\|$; (5)
6	$prop.fT_h$	Tweet proportion by fU_h: $\|fT_h\|/\|T_h\|$; (5)
7	$prop.\text{s}T_h$	Tweet proportion by sU_h: $\|sT_h\|/\|T_h\|$; (5)
8	$prop.oU_h$	Trend user proportion: $\|oU_h\|/\|U_h\|$; (4)
9	$prop.fU_h$	Follower user proportion: $\|fU_h\|/\|U_h\|$;(4)
10	$prop.sU_h$	Self-motivated proportion:$\|sU_h\|/\|U_h\|$;(4)
Structure Context		
11	$rat.bU_htU_h$	Ratio of border to trend $\|bU_h\|/\|tU_h\|$
12~13	$prest.tU_h$ [2]	Max/Average prestige of trend user tU_h
14	$dens.tU_h$	Sub-graph density of tU_h; see (6)
15	$dens.tU_hbU_h$	Bipartite graph density of $tU_h$and bU_h
16	$recp.tU_h$	Sub-graph reciprocity of tU_h; See(7)
17	$recp.bU_htU_h$	Bipartite graph reciprocity of bU_h and tU_h
Node Context		
18~19	$act.\&U_h$ [7]	Average general activeness of $\&U_h$; (8)
20	$act_h.tU_h$	Average trend activeness of tU_h; see (8)
21	$rat.act_h.tU_h$	Activeness ratio: $act_h.tU_h/act.tU_h$
22~23	$stim_h.\&U_h$	Average trend stimulus of $\&U_h$; see (9)
24~25	$rat.stim_h.\&U_h$	Stimulus ratio: $stim_h.\&U_h/stim.\&U_h$
26~27	$freq.int.\&U_h$	Interaction frequency of $\&U_h$
28~29	$ratio.int.\&U_h$	Interaction ratio of $\&U_h$;

1. In this table, time index t is usually ignored for brevity.
5. The prestige here is the average out-degree.
3. '$\&U_h$' means trend users tU_h or border users bU_h (Figure 2).

Node context factors describe the social network users based on their action and other profile. Both trend users tU_h and trend border bU_h is important for information diffusion, so the node context features will derived from the two user sets. Our node context consists of the following aspects as the input, output, and style of a user set.

Activeness is the user's action frequency in Twitter, i.e. posting tweets. It can be taken as the 'temperature' of a user to generate information. The *general activeness* of user u on time t is $act(u;t)$ defined as:

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

where α is the decay coefficient and $T(u;t)$ is the tweets posted by u on time t. *Trend activeness* ($act_h(u;t)$) is defined similar to (8) by replacing $T(u;t)$ with $T_h(u;t)$ which the tweets of trend h posted by u on t. In fact, high trend activeness implies a high probability of a user's participation in h. Furthermore, the ratio between $act(u;t)$ and $act_h(u;t)$ reflects user's interest on a specific trend.

Stimulus is the volume of information, i.e. tweets, received by a user. The g*eneral stimulus* of user u on time t is derived from all tweets posted by friends:

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

where β is the decay coefficient, $friend(u)$ is the friend set of u, and $T(u;t)$ is the same as (8). Similarly, *trend stimulus* $stim_h(u;t)$ is derived from tweets of trend h posted by friends, which is to replace the $T(u_i;t)$ with $T_h(u_i;t)$.The ratio between trend and general stimulus will reflect user's 'attention' on a specific trend. In fact, attention-limited similarity measure will improve the prediction of information adoption behavior (Lerman, et al. 2012).

Action style is how a user acts in Twitter. In fact, user interaction, i.e. retweet and mention, is the main course of information diffusion in Twitter. The *interaction frequency* of a user u is percentage of tweets posted by u that are interactions up to current time. *Interaction rate* is the number of interaction tweets received by a user divided by the number of all tweets he posted (Suh, et al. 2010). In fact, interaction frequency reflects the will of a user to spread information, while interaction rate reflects his influence.

Trend Prediction Models

A model is a mapping function form input variables to output prediction objects. Since a model can be characterized by two properties, i.e. linearity and state-space, the full combination leads to 4 model types to be introduced here.

Preliminaries

Following the convention of system identification (Ljung 1998), we use $\mathbf{y}(t)$ for the o_y dimensional *output vector* under prediction, and $\mathbf{u}(t)$ for the o_u dimensional *input vector*. For state-space model, the *state vector* is $\mathbf{x}(t)$ with a dimension of o_x.

Non-state-space model predicts the next output $\mathbf{y}(t+1)$ from several past input and output vectors:

$$\mathbf{y}(t+1) = \boldsymbol{f}(\boldsymbol{\varphi}_{na,nb,nk}), \quad (10)$$

Where $\boldsymbol{f}(\cdot)$ is prediction function, $\boldsymbol{\varphi}_{na,nb,nk} = [\mathbf{y}(t) \dots \mathbf{y}(t+1-n_a), \mathbf{u}(t-n_k) \dots \mathbf{u}(t+1-n_k-n_b)]$ is *predictor vector*, n_k is input delay, and the model order is n_a and n_b . Non-state-space model is widely used due to its simplicity and effectiveness. However, the prediction is limited to previous n_a output and n_b input vectors cascaded in $\boldsymbol{\varphi}_{na,nb,nk}$. The history beyond the model order will be discarded.

State-space model solves the limited-memory problem of non-state-space model by introducing state vector (t) :

$$\begin{aligned} \mathbf{x}(t+1) &= \boldsymbol{h}(\boldsymbol{\varphi}_{na,nb,nk}, \mathbf{x}(t), \dots, \mathbf{x}(t-n_d)) \\ \mathbf{y}(t+1) &= \boldsymbol{g}(\boldsymbol{\varphi}_{na,nb,nk}, \mathbf{x}(t+1), \dots, \mathbf{x}(t+1-n_d)) \end{aligned}, \quad (11)$$

where $\boldsymbol{\varphi}_{na,nb,nk}$ holds the same meaning as (10), $\boldsymbol{h}(\cdot)$is the state-transition function, $\boldsymbol{g}(\cdot)$ is the output emission function, and n_d is the state order. State-space model can predict future with full history, because the effect of previous observations is accumulated in its state variables. In fact, (11) is just a general form of state-space model, and the $\boldsymbol{h}(\cdot)$ and $\boldsymbol{g}(\cdot)$ can be either linear or non-linear with many different implementations as introduced below.

Prediction Models

ARX model (Auto Regressive model with eXternal input) (Ljung 1998) is a linear non-state-space model defined as:

$$\mathbf{y}(t) = \sum_{j=1}^{n_b} \mathbf{B}_j \cdot \mathbf{u}(t-n_k-j) - \sum_{i=1}^{n_a} \mathbf{A}_i \cdot \mathbf{y}(t-i), \quad (12)$$

where the model parameter $\mathbf{A}_i$ and $\mathbf{B}_j$ is a matrix of size $o_y \times o_y$ and $o_y \times o_u$ respectively. The input vector can be ignored by setting $n_b = 0$, which is also called **AR** model. ARX model can be estimated with simple least square method which makes it the most popular baseline model.

Nonlinear ARX (*NARX*) *model* (Figure 3) is a feed-forward neural network implementation of non-linear non-state-space model (1). The nonlinearity come from the neuron sigmoid transfer function operating on linear combination of layer's output. NARX can be trained by back-propagation. With enough layers and hidden nodes, NARX can approximate any function well (Haykin 2008).

Random forest (*RF*) (Hastie, Tibshirani and Friedman 2009) is a popular ensemble non-linear non-state-space method (10) as a set of regression trees. The robustness of RF lies in the independence among different trees which is achieved in two steps. First, each tree is constructed independently by bootstrap samples. Second, each tree node split is decided on a random subset of predictor variables, i.e. $\boldsymbol{\varphi}_{na,nb,nk}$ in (10). The final prediction result of RF is the average of each regression tree.

Linear dynamic system model (LDS) is a state-space model under linear assumption of state transition ($\boldsymbol{h}$) and output emission ($\boldsymbol{g}$) (See (11)). With $n_a = 0, n_b = 1, n_k = 1$ and $n_d = 1$, LDS can be formulated as:

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{A} \cdot \mathbf{x}(t) + \mathbf{B} \cdot \mathbf{u}(t) + \mathbf{w}(t) \\ \mathbf{y}(t+1) &= \mathbf{C} \cdot \mathbf{x}(t) + \mathbf{D} \cdot \mathbf{u}(t) + \mathbf{v}(t) \end{aligned}, \quad (13)$$

where $\mathbf{v}(t) \sim N(\mathbf{0}, \mathbf{R})$ and $\mathbf{w}(t) \sim N(\mathbf{0}, \mathbf{Q})$ are zero mean Gaussian noise. The *LDS model order* is the dimension of state vector $\mathbf{u}(t)$. Another popular state-space model is HMM (Bishop 2006), however it is not appropriate for regression due to its discrete sate space in nature.

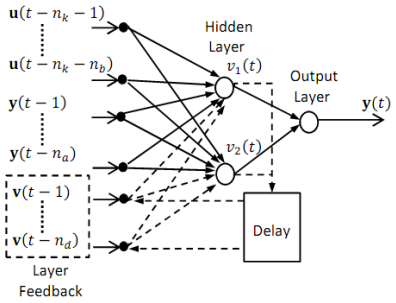

Figure 3. An illustration of NARX and RNARX network with only 1 hidden layer of 2 nodes. RNARX is the extension of NARX with the dashed feedback loop.

Recurrent Nonlinear ARX (RNARX) network is state-space (11) extension of NARX with the introduction of delayed feedback loops around each layer except for the last output layer (Figure 3). The layer feedback can improve the network capability of modeling complicated dynamics, since the influence of history will accumulate on hidden nodes (or state) througth the feedback loops.

Table 2. Statistics of Arab Spring Twitter Dataset

Name	Value	Name	Value
Tweets	16,043,422	Retweet	5,336,868
URL ratio	40.33%	Hashtag Ratio	97.48%
Users	666,168	Links	86,710,704
Mean friend	130.20	Reciprocity	19.9%

Experiments and Discussion

Dataset and Preprocessing

Our Twitter dataset is collected for the purpose of Arab Sprint. The collection involved the manually defined hashtags and geographic regions related to the following countries: Egypt, Libya, Syria, Bahrain and Yemen. Meanwhile, the Twitter network among the related users is crawled by the Twitter API as well. We collected 16.8 million Tweets form 0.67 million users from February 1, 2011 to August 31, 2011 using the streaming API. The crawled Tweets during this course account for approximately 10% of the all Tweets hosted by Twitter. The statistics of our dataset is summarized in Table 2. The collection process leads to a high hash-tag ratio, which means most tweets contain at least one hash-tag. Therefore, the dataset is appropriate for our hash-tag trend prediction purpose.

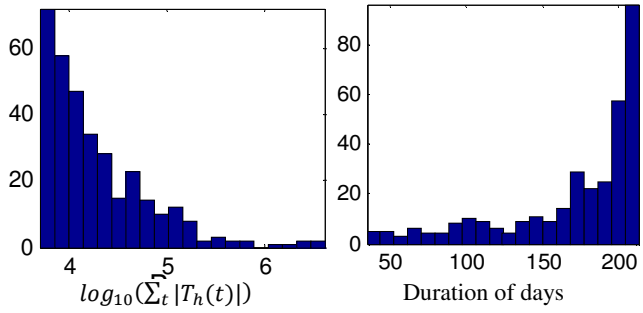

Figure 4. Left: the histogram of total tweet number over hashtag trends. Right: the histogram of trend duration.

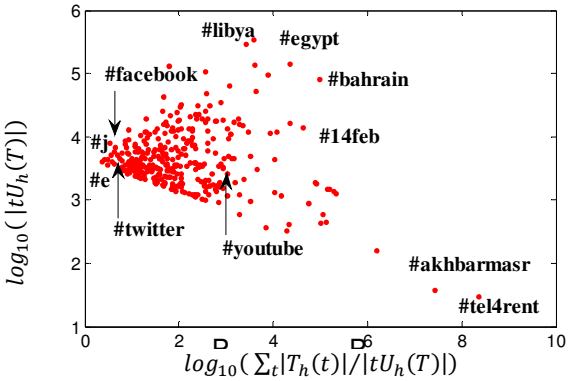

Figure 5. The scatter plot of total users ($|tU_h(T)|$) involved and their average tweets ($\sum_t |T_h(t)|/|tU_h(T)|$) in the trend.

We select the 336 most popular hash-tags with at least 5000 related tweets for trend prediction task. With the high hashtag ratio in Table 2, the 336 hashtag sum up to 28,333,656 tweets which is above the total tweets number 16,043,422 (Table 2) due to replicate count. Basic statistics of these trends are illustrated in Figure 4, while Figure 5 gives more insight on the trend characters. For example, in some hashtag trends, e.g. #tel4rent and #akharmasr, only few users posted a lot of tweets as advertisement for some products. The most popular trends are the Arab Spring hashtags, e.g. #egypt and #libya, have a large population of trend users and a high average tweets.

We evaluate prediction by Mean Square Error (MSE). All prediction results in this section are estimated by 10-fold cross-validation, where 90% trends are used for training and the rest 10% are left for testing. The features in Table 1 are normalized by Z-score transform before feed into prediction model.

Comparison of Trend Factors

With the so many content and context factors listed in Table 1, several questions may arise: will these factors really help? Which type of factor is most relevant for trend prediction? The above questions are answered by both feature importance and prediction accuracy analysis.

Factor Importance Analysis

We use random forest (RF) (Hastie, Tibshirani and Friedman 2009) for feature importance analysis, where the *variable importance* measure is the decay in test accuracy by permuting out-of-bag samples. The importance is normalized to [0,1] by dividing the highest value. We set the prediction vector $\boldsymbol{\varphi}_{na,nb,nk}$ (10) to be $n_a = n_b = 5$ and $n_k = 1$ which is the input and output value in the previous 5 days. The reason is that the RF prediction result will not change significantly after 5 days. In this paper, the decay coefficient α and β for activeness (8) and stimulus (9) are set to 0.1. In fact, the result may not change too much for small decay coefficients, while large value may lead to performance decrease. Following the factor index of Table 1, we illustrate the mean variable importance of each factor over 5 days. We only present the importance for the popularity measure of U_h(t) because the result of T_h(t) is similar. The detailed mean and std of 15 most important trend factors are listed in Table 3.

Based on the statistics in Table 3 and Figure 6, we can see that the importance of node context factors, i.e. stimulus and activeness, are significantly better than other trend factors. Stimulus is a measure of trend information input to users. Large trend stimulus (4) means a high exposure to trend tweets, so users are likely to participate the trend. On the other hand, activeness is a measure of trend information output by users. Large trend activeness (8) means a high frequency to trend participation, so users are likely to post a trend tweet again.

The second important type of trend factors is the structure context factors about the relation between trend users and trend border (Figure 2). This is also intuitive. For example, a high $rat.bU_htU_h$ (size ratio) means the trend are broadcasted to a lot of users (bU_h), and a $recp.bU_htU_h$ (reciprocity) means these listeners has a good relationship with the trend propagators (tU_h).

Form the above discussions, we have two observations. First, trend factors based on user behaviors are more likely to be important, e.g. trend stimulus/activeness. This is coherent to the founding of (Cha, et al. 2010) that user interaction graph is more appropriate for influence measure. Second, the factors should be designed with specific to the information trend. 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 (Figure 2) are also important. On the contrary, the action style node context factors, i.e. interaction frequency/ratio, are less important for trend prediction, because they are just some general information about user's behavior.

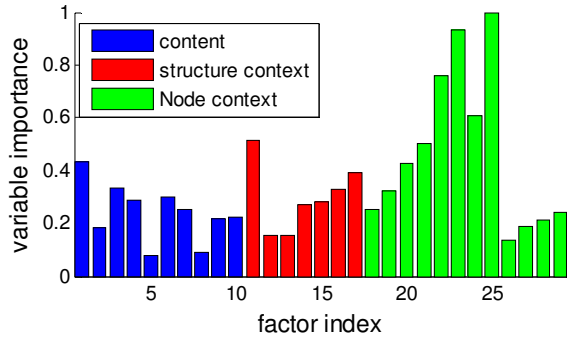

Figure 6. The variable importance of different trend factors for user (U_h(t)) prediction. The factor index is in Table 1.

Table 3. Trend factors sorted by RF variable importance.

Ind.	Symbol	T_h(t): Mean (Std)	U_h(t): Mean (Std)
22	$stim_h.tU_h$	1.000 (0.376)	0.917 (0.197)
23	$stim_h.bU_h$	0.788 (0.261)	0.964 (0.350)
24	$rat.stim_h.tU_h$	0.744 (0.195)	1.000 (0.381)
25	$rat.stim_h.bU_h$	0.883 (0.384)	0.722 (0.151)
21	$rat.act_h.tU_h$	0.560 (0.179)	0.559 (0.157)
11	$rat.bU_htU_h$	0.383 (0.029)	0.716 (0.076)
20	$act_h.tU_h$	0.412 (0.139)	0.529 (0.098)
17	$recp.bU_htU_h$	0.336 (0.023)	0.560 (0.039)
1	$prop.\mathrm{rt}T_h$	0.327 (0.184)	0.441 (0.166)
7	$prop.\mathrm{s}T_h$	0.332 (0.077)	0.346 (0.056)
16	$recp.tU_h$	0.232 (0.029)	0.435 (0.068)
19	$act.bU_h$	0.236 (0.043)	0.415 (0.061)
15	$dens.tU_hbU_h$	0.236 (0.014)	0.412 (0.023)
6	$prop.fT_h$	0.288 (0.075)	0.341 (0.097)
3	$prop.\mathrm{n}T_h$	0.271 (0.135)	0.354 (0.114)

Note: the index and symbol of trend factor is the same as Table 1. The factors are sorted by the mean importance of T_h(t) and U_h(t).

Table 4. 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' means prediction with all trend factors; 'None' means prediction without trend factors, i.e. $\boldsymbol{n_b} = \mathbf{0}$ for $\boldsymbol{\varphi}_{na,nb,nk}$in (10).

Prediction Performance Analysis

We further investigate the importance of different type of factors by comparing their prediction performance. For the sake of consistency, we use the same RF prediction model as the above factor importance analysis.

From the statistics of prediction MSE in Table 4, we can see the following points. First, the trend factors (Table 1) really helps to improve trend prediction, because the MSE with all trend factors is significantly better than that without any factors. Second, the MSE order of node context, structure context, and content factors follows the variable importance in Figure 6. However, the difference is not significant. This might be explained by Table 3 that these factors are complementary and have importance on their own aspects.

Comparison of Prediction Models

In this subsection, we try to investigate different types of models for trend prediction. A model is usually characterized by two aspects, i.e. linear/non-linear, state-space/non-state-space, which leads to a combination of 4 model categories. For each category, at least 1 typical model is chosen so that they have some similarity to models of other category. To be specific, LDS (13) is the direct state-space extension of ARX (12), and NARX (Figure 3) is the direct non-linear extension of ARX. Moreover, RNARX is the state-space extension of NARX. So comparison among these models can reflect the effect of different model categories.

The comparative prediction result is summarized in Table 5 under evaluation of MSE. The prediction models are trained with all factors in Table 1. For each model, we tried a series of model setups, and present the best result. For ARX, RF, and NARX model, we go through order 1 to 5 for the predictor vector $\boldsymbol{\varphi}_{na,nb,nk}$ in (10). For LDS (13) we check the model order up to 4. Both the NARX and RNARX models are 1 layer network, and the hidden node number varies from 2 to 10. The predictor order of RNARX (Figure 3), i.e. n_a and n_b, is fixed to 1, and the delay order n_d is 1.

To show the effectiveness of above models, we also designed two baseline prediction methods. The first is called *Last* Predictor which simply predicts future as last value, i.e. $\hat{\mathbf{y}}(t+1) = \mathbf{y}(t)$. The second is called *Mean* Predictor which simple predicts future as mean value up to latest time, i.e. $\hat{\mathbf{y}}(t+1) = t^{-1} \cdot \sum_1^t \mathbf{y}(t)$. Their prediction results are also included in Table 5.

Table 5. 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/NL is the short for Linear/Non-Linear; S/NS is the short for State-space/Non-State-space.

We have the following observations from Table 5. First, the twitter hashtag tend can be predictable on some degree, because the MSE of all prediction models are significantly better than the two baseline methods, i.e. Last predictor

and Mean predictor. Second, nonlinearity is very important for trend prediction, which is clear from the performance gap between linear models, i.e. ARX and LDS, and their non-linear peers, i.e. NARX and RNARX. The non-linearity may mainly due to the complex mechanism of network information trend diffusion. Third, state-space is helpful, but the benefit is smaller than 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 history information might not be as important for Twitter trend as it is for other trend, e.g. economics. 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 possible explanation is that state-space models usually use local gradient training methods, which might not be competent at the complicated cost function surface due to high feature dimension. Further feature selection or regularization might help the situation.

Conclusion

In this paper, we study two basic problems in information trend prediction, i.e. important factors and appropriate models. We designed features of different trend factors from both tweet content and network context. We also investigate model categories as the combination of two basic properties, i.e. (non)-linearity and (non)-state-space. Experiments on large Twitter dataset lead to the following observations. Both content and context factors will help trend prediction. However, node context factors of user's behavior on trend, e.g. trend stimulus and activeness, are most relevant. As for the prediction model, non-linear models are significantly better than their linear peers, which may mainly due to the complex information diffusion process in large social network. State-space can help to improve prediction but only on a slight degree.

Future work can lies in two aspects. First, semantic content of information trend is very relevant to their diffusion process (Romero, Meeder and Kleinberg 2011) (Lehmann, et al. 2012). However, how to model the effect appropriately is still an open and interesting problem. Second, network environment may change over time. Also, there will be some unique factors for each individual trend. Therefore, adaptive model may lead to further improvement.

Reference

Asur, S.; Huberman, B. A.; Szabo, G.; and Wang, C. 2011. Trends in social media: Persistence and decay. *In 5th International AAAI Conference on Weblogs and Social Media*.

Bishop, C., 2006. *Pattern recognition and machine learning*, Springer New York.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. *In 4th international AAAI conference on weblogs and social media (ICWSM), AAAI*.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2001. *The elements of statistical learning*. Springer.

Haykin, S. 2008. *Neural networks and learning machines*. Prentice Hall.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? *In Proceedings of the 19th international conference on World Wide Web (WWW),* 591–600. ACM.

Lehmann, J.; Goncalves, B.; Ramasco, J. J.; and Cattuto, C. 2012. Dynamical classes of collective attention in twitter. *In Proceedings of the 21st international conference on World Wide Web (WWW),* 251–260. ACM.

Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *In Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, AAAI.

Lerman, K.; Intagorn, S.; Kang, J.-H.; and Ghosh, R. 2012. Using proximity to predict activity in social networks. *In Proceedings of the 21st international conference companion on World Wide Web (WWW)*, 555–556. ACM.

Ljung, L. 1998. *System Identification: Theory for the User*. Prentice Hall.

Ma, Z.; Sun, A.; and Cong, G. 2012. Will this# hashtag be popular tomorrow? In Proceedings of the 35th international ACM SIGIR conference, 1173–1174. ACM.

Matsubara, Y.; Sakurai, Y.; Prakash, B. A.; Li, L.; and Faloutsos, C. 2012. Rise and fall patterns of information diffusion: Model and implications. *In Proceedings of the 18th ACM SIGKDD* , 6–14. ACM.

Newman, M. 2010. *Networks: an introduction*. Oxford University Press, Inc.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. *In Proceedings of the 20th international conference on World Wide Web (WWW)*, 695–704. ACM.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010.Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. *In IEEE Second International Conference on Social Computing*, 177–184. IEEE.

Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Communications of the ACM* 53(8):80–88.

Tsur, O., and Rappoport, A. 2012. What's in a hashtag?:content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM)*, 643–652. ACM.

Yang, J., and Counts, S. 2010. Predicting the speed, scale, and range of information diffusion in twitter. In *4th international aaai conference on weblogs and social media (ICWSM)*, AAAI.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)*, 177–186.

Yang, L.; Sun, T.; Zhang, M.; and Mei, Q. 2012. We know what@ you# tag: does the dual role affect hashtag adoption? In

Proceedings of the 21st international conference on World Wide Web (*WWW*), 261–270. ACM.

Yu, S., and Kak, S. 2012. A survey of prediction using social media. *arXiv preprint arXiv*:1203.1647.