6-2012

Prediction of Topic Volume on Twitter

Yiye Ruan

Hemant Purohit
Wright State University - Main Campus

David Fuhry

Srinivasan Parthasarathy

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

Repository Citation

Ruan, Y., Purohit, H., Fuhry, D., Parthasarathy, S., & Sheth, A. P. (2012). Prediction of Topic Volume on Twitter. *Proceedings of the 4th International ACM Conference on Web Science*, 397-402.
https://corescholar.libraries.wright.edu/knoesis/596

Prediction of Topic Volume on Twitter

Yiye Ruan
Department of Computer Science and Engineering
Ohio State University
ruan@cse.ohio-state.edu

Hemant Purohit
Kno.e.sis, Department of Computer Science and Engineering
Wright State University
hemant@knoesis.org

David Fuhry
Department of Computer Science and Engineering
Ohio State University
fuhry@cse.ohio-state.edu

Srinivasan Parthasarathy
Department of Computer Science and Engineering
Ohio State University
srini@cse.ohio-state.edu

Amit Sheth
Kno.e.sis, Department of Computer Science and Engineering
Wright State University
amit@knoesis.org

WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.
ACM 978-1-4503-1228-8

Abstract

We discuss an approach for predicting microscopic (individual) and macroscopic (collective) user behavioral patterns with respect to specific trending topics on Twitter[1]. Going beyond previous efforts that have analyzed driving factors in **whether** and **when** a user will publish topic-relevant tweets, here we seek to predict the **strength** of content generation which allows more accurate understanding of Twitter users' behavior and more effective utilization of the online social network for diffusing information.

Unlike traditional approaches, we consider multiple dimensions into one regression-based prediction framework covering network structure, user interaction, content characteristics and past activity. Experimental results on three large Twitter datasets demonstrate the efficacy of our proposed method. We find in particular that combining features from multiple aspects (especially past activity information and network features) yields the best performance. Furthermore, we observe that leveraging more past information leads to better prediction performance, although the marginal benefit is diminishing.

Keywords

Social Networks, User Engagement, Volume Prediction, People-Content-Network Analysis (PCNA)

ACM Classification Keywords

H.2.8 Database Management: Database Applications - Data Mining

[1] http://www.twitter.com

Introduction

The ubiquitousness of Twitter as a micro-blogging service has revolutionized and reshaped how information spreads in cyber-space. By sending short *tweets* from computers and mobile devices, Twitter users can easily publish content, interact with others and engage in online discussion. Apart from being a social network, Twitter has also proved valuable in many situations including epidemic surveillance [3], emergency response [10], political campaigns [13], etc.

Researchers have been studying Twitter and other social networks from multiple facets in recent years, and an affluence of works have been presented [2, 9, 5, 14, 7]. Here we are interested in predicting Twitter users' behavior of generating topic-relevant tweets, and especially estimating the amount of relevant tweets they will write in the future. Accurate prediction of this value can benefit understanding online community sustainability (i.e. to estimate the amount of discussion within a community in the future), viral marketing (i.e. to design strategy which maximizes the reach of viral messages) and many more applications. Unfortunately, previous studies modeled the prediction problem as binary classification (for instance, "whether a user writes tweet or not in given time frame" [8] , "whether the group size exceeds a threshold or not" [5]), which has a coarse granularity and cannot provide a precise estimate of topic discussion volume.

In this paper, we focus on two tasks: predicting the *microscopic* (individual) and *macroscopic* (collective) volume of topical tweets that Twitter users will generate within a time frame. Beyond prior tweeting activity and content analysis, we posit that the underlying follower-followee network has a critical role to play in predicting the strength of this signal. We extract a series of features from tweet content, user network structure, neighboring friends' influence and user past activity, and build a linear regression model extending our earlier effort [8] on building a unified framework for effective study using multiple dimensions, namely People-Content-Network analysis (PCNA). Experiment results show that our model achieves decent accuracy on both tasks.

Methodology

Problem Statement

The first (microscopic) task of our study is to predict how many tweets relevant to a topic each user will write on each day in a time interval. Specifically, given a topic t we want to estimate the relevant tweet volume $vol_i^t u$ by each user u on each day $d_i \in [d_s, d_e]$, using only information from the past h days before d_i (i.e. $[d_{i-h}, d_{i-1}]$). The second (macroscopic) task is to predict the total amount of relevant tweets that a group of users will generate on each d_i, i.e. $\sum vol_i^t u$. Please refer to Table 1 for notations.

Let us start with explaining the notion of a tweet "being relevant" to a topic. For each topic, we initially collect tweets based on manually-selected seed keywords using Twitter Streaming API. We build a *topic-context set* by fetching concepts and entities from the top 3 Wikipedia pages retrieved from Google search for the topic, inspired from approach in [12]. This set is complemented by extracting top 10% frequent terms (uni-grams, bi-grams and hashtags) from the tweet corpus. We keep human in the loop to maintain this set's relevancy. A tweet is considered *relevant* to the topic if its text contains at least one element in this set.

We also observe the need for limiting the user search space as it is impractical and unnecessary to consider all tens of millions of registered Twitter users. Therefore, on the day d_i we only consider the prediction of users in a candidate set C_i^t. A user u belongs to the candidate set for topic t on d_i if and only if 1) u wrote at least one tweet relevant to t during the whole time period $[d_s, d_e]$, and 2) u follows some users who also wrote a tweet relevant to t during d_{i-h} and d_{i-1}. The first condition is because we are focusing on predicting **how many** relevant tweets a user would write, and the problem of determining *whether* a user would write relevant tweet has been addressed in our previous work [8]. This restriction also removes Twitter users who were simply offline or inactive during the period. The second condition is because social network users' behavior is greatly influenced by their social circles. Since tweets written by a user are broadcast to all followers, those followers will become aware of the topic discussed in the tweets and are likely to write about the topic, too.

Table 1 summarizes notations used in the paper.

Notation	**Meaning**
t	topic
u	Twitter user
d_i	day i
$vol_i^t u$	number of tweets relevant to t written by u on d_i
$[d_s, d_e]$	time period for prediction
h	number of past days used for modeling
U_i^t	authors that wrote tweets relevant to t on d_i
U^t	$\bigcup_{d_s \le d_i \le d_e} U_i^t$
$\Gamma(u)$	u's followers
$\Gamma^{-1}(u)$	u's friends (i.e. followees)
C_i^t	$\{u \mid u \in U^t \wedge (\exists v \in \bigcup_{d_{p-h} \le d_i < d_p} U_i^t$ s.t. $u \in \Gamma(v))\}$

Table 1: Notation Table

Feature Description

We introduce the set of features considered by our model, which can be divided into four categories: network, author, content and past activity.

Network Features (*NF*)

Network features are structural features which, on a coarse level, measure the amount of topic-related information a user u has received from his/her online friends on a day d_i.

- Connectivity: number of u's friends who wrote tweets relevant to t on d_i. According to the notations, it is $|\{v \mid v \in U_i^t \wedge v \in \Gamma^{-1}(u)\}|$.
- Highly-engaged Connectivity: number of u's *highly-engaged* friends. An author v is said to be highly-engaged on topic t on d_i if $vol_i^t v$ is within the top 3% of $vol_i^t w$ for all $w \in U_i^t$.
- Friends' total tweet volume: total number of relevant tweets written by u's friends on d_i.
- Interaction: total number of times u was *mentioned* (via the *@username* symbol) by friends in U_i^t on d_i. We use this feature to capture interactions between users and friends, which reflect much stronger ties than ordinary follower-friend relationships. Mention also includes *retweet* (via the *RT@username* convention), another type of user interaction.

Author Features (*AF*)

Author features are designed to capture the online influence of a user. Intuitively, the more influential and authoritative an author is, the more likely his/her followers will be "activated" and spread the information further. For an author u, the two author features are:

- Klout score: u's klout score[2], which is a unified third-party score composed of influence characteristics across multiple social networks.
- Logarithm of number of followers: $\log_{10}|\Gamma(u)|$. Though simple, this feature is often well correlated with a user's influence and network reach [1, 14].

Content Features (*CF*)

There exist intrinsic differences between 140-character tweets and traditional documents. First of all, tweets tend to have a more informal writing style, including heavy usage of acronyms and emoticons. Moreover, space constraints encourage the inclusion of hyperlinks that point to full-length articles or multimedia contents. For example, almost 50% of tweets relevant to the Occupy Wall Street movement contain at least one hyperlink[3]. Lastly, several features including retweet, mention and hashtag are Twitter-specific and do not appear on other platforms. In previous studies [11, 1], those content features were also shown to encourage spread of tweets. We define the following features for each tweet:

- Retweet count: number of *RT@username* patterns.
- Mention count: number of *@username* patterns.
- Hashtag count: number of *#phrase* patterns.
- Relevant URL count: number of relevant hyperlinks. To determine the relevancy of URLs, we used the *topic-*

[2] http://www.klout.com

[3] Check http://twitris.knoesis.org/ows/insights/

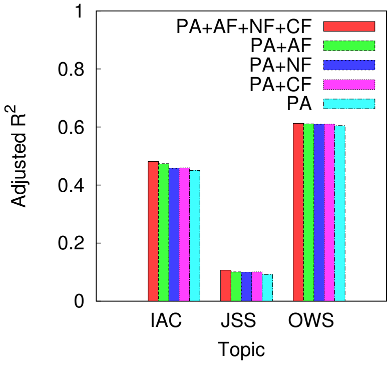

Figure 1: R_a^2 on Microscopic Prediction with Different Feature Groups, $h = 2$

Topic	max $P(f \geq F^* \mid H_0)$ Full vs. Strawman
IAC	2.094323×10^{-3}
JSS	1.142266×10^{-10}
OWS	$9.977034 \times 10^{-156}$

Table 3: Partial F-tests Results

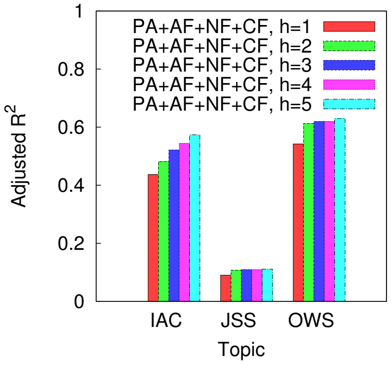

Figure 2: R_a^2 on Microscopic Prediction with Varying h Values

context set. If the hyperlinked content contains two or more of the concepts, the URL is regarded as relevant. Otherwise, the count is reduced by one.

- Multimedia URL count: number of hyperlinks to multimedia contents.
- Subjectivity score: weighted average of subjectivity scores of words[4], symbols and emoticons[5].

Past Activity (*PA*)

On the day d_i, the feature group of past activity contains $vol_j^t u$ for $d_j \in [d_{i-h}, d_{i-1}]$. It is included for two reasons. First, past activity is helpful in showing a user's inertia of writing topic-relevant tweets, which is in turn a good predictor for the user's future behavior. Second, it is easy to curate in practice.

Modeling Process

For each prediction day d_p, features described above are computed for all candidates and all days over $[d_{p-h}, d_{p-1}]$. A two-stage feature consolidation is then performed to materialize all features for a candidate user as a single vector. For a user u and a day $d_i \in [d_{p-h}, d_{p-1}]$, the first step is to sum up author features of u's friends in U_i^t as well as content features of tweets written by those friends. The assumption behind this operation is that the likelihood of a user generating relevant content is positively correlated to the amount of influence received from his/her friends. The second step is to, for each u, perform weighted summation of network, author and content features over the time period $[d_{p-h}, d_{p-1}]$, where the weight is exponentially decaying with time. For a day $d_i \in [d_{p-h}, d_{p-1}]$, the weight is calculated as $\alpha^{d_p - d_i}$, where $0 < \alpha < 1$. This is a common approach used in social network analytics in order to emphasize the importance of more recent information [15, 4, 8]. For this study we let α be 0.8. Past activity features are not aggregated, and each of them is treated as a separate feature.

We then build a linear regression model using feature vectors on d_p. All or a part of feature elements are used as regressors, and $vol_p^t u$, the number of u's relevant tweets on d_p is the regressand. Compared with other tools, linear regression has multiple advantages including higher efficiency, low storage overhead and statistical interpretability on model coefficients. We discuss the prediction performance of our model in detail in the following section.

Experiments

Datasets

In this section, we present experimental results following the methodology described above. We crawled tweets relevant to three topics frequently discussed in late 2011 for nearly a month: Anti-corruption movement in India (*IAC*), former football coach Jerry Sandusky's scandal (*JSS*) and Occupy Wall Street movement (*OWS*). Table 2 lists basic information of the three datasets.

Topic	Period	# Tweets	# Unique Authors
IAC	11/06 - 12/02	93,525	19,705
JSS	11/06 - 11/30	251,316	152,174
OWS	11/06 - 12/02	2,042,653	320,415

Table 2: Datasets Statistics

In this paper, we focus on the study of efficacy of each feature group as a unit. For brevity, studies on the effects of individual features is not presented due to the exponential number of possible combinations of them.

Microscopic (User-Level) Prediction

We first tested the prediction of individual users' relevant tweet volume. We use R language's linear regression package, and report adjusted R^2 value (R_a^2) for each model. For

[4] http://www.cs.pitt.edu/mpqa/subj_lexicon.html

[5] Compiled from emoticons listed on http://en.wikipedia.org/wiki/List_of_emoticons and their variants.

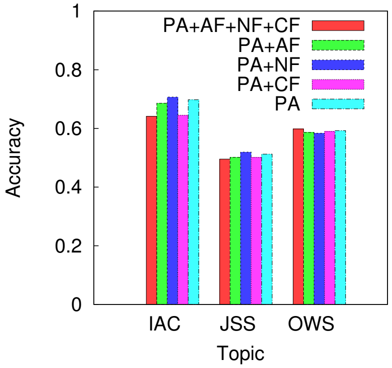

Figure 3: Accuracy on Macroscopic Prediction with Different Feature Groups, $h = 2$

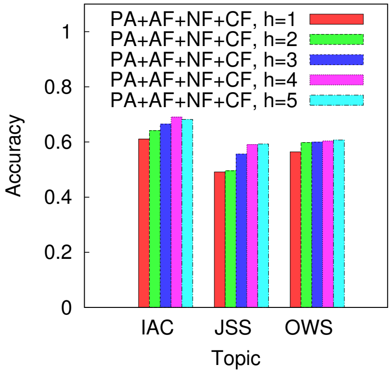

Figure 4: Accuracy on Macroscopic Prediction with Varying h

a regression model with n records and p regressors (i.e. features), R_a^2 it is defined as $R_a^2 = 1 - (\frac{n-1}{n-p})\frac{SSE}{SSTO}$, where $SSTO$ and SSE are sum of squares of regressands and residuals, respectively [6]. Higher R_a^2 value indicates that a larger proportion of total sum of squares is explained, thus a higher explanatory power of the model.

We used two days of past information (i.e. $h = 2$) to build models, and computed average R_a^2 value over the period. Figure 1 shows the results of five models using different selections of features. The name of each model indicates which feature groups it uses, where *PA* stands for past activity, *NF* for network features, *AF* for author features and *CF* for content features. For example, model $PA+AF$ includes past activity information and author features as regressors. As observed from the plot, higher R_a^2 values are obtained when extra features are added on top of past activity. Another finding is that author features introduce additional explaining power beyond network features and content features. Although there is no guarantee of causality, it may suggest that the motivation behind users' involvement in topical discussion is attributed more to the general influence of friends than the specific content.

R_a^2 value could be inflated when more regressors are included. To address this concern, we further performed *partial F-tests* on the full model ($PA+AF+NF+CF$) against the simple strawman (PA). The null hypothesis H_0 is that all additional features' coefficients are zero, and a statistic F^* will follow an F distribution if H_0 holds [6]. As shown in Table 3, for all topics the conditional probability $P(f \geq F^* \mid H_0)$ never exceeds 10^{-2} on any single day's data. Therefore, we reject H_0 and conclude that the additional explaining power from extra features is statistically significant.

For the JSS dataset we note that the overall user-level (microscopic) prediction accuracies are low. We should point out that on this dataset the average number of tweets per user is under 2. Thus, there is by and large insufficient information on most users to predict how much they will tweet on this topic. However, it is interesting to note that if we look at a subset of the users that tweet more frequently (> 5 tweets on this topic, results not shown) and also when one aims at predicting the output of the collection of users in its entirety, the accuracy increases significantly (see section *macroscopic prediction* and Figure 3).

Impact of Amount of Past Information

To investigate the impact of past information amount on model performance, we ran another set of experiments where parameter h was varied from 1 to 5. Figure 2 shows the result. The first observation is that the more past information is available, the higher R_a^2 values. A second observation is that improvement from additional past information is often diminishing, suggesting that recent information has larger influence than older. Such a finding is consistent with those from previous works.

Macroscopic prediction

Finally, we present the results on predicting the behavior of users *en masse*. For a prediction day d_p, we use the coefficients learned from previous day's regression model to fit the feature vectors on d_p. Then we compute the accuracy value as $1 - \frac{|\sum_{i \in C_p^t} \hat{vol}_p^t u - \sum_{i \in C_p^t} vol_p^t u|}{max(\sum_{i \in C_p^t} \hat{vol}_p^t u, \sum_{i \in C_p^t} vol_p^t u)}$, where $\hat{vol}_p^t u$ is the estimated total volume and $vol_p^t u$ is the real total volume.

Figures 3 and 4 show results with varying models and h values, respectively. For each topic, the average accuracy over days is reported. Compared with that on microscopic prediction, the performance of topic JSS has significant improvement. Again, the trend of diminishing return on the amount of past information is observed.

Conclusion

In this abstract, we introduce an effective framework for modeling and predicting the volume of topic-specific tweets that will be generated by Twitter users in the future. Experimental results show that including features based on content, network structure and online influence provides higher explaining power than using past activity information alone, and the benefit is statistically significant. We also find that newer knowledge contributes more to the prediction accuracy than older knowledge. Apart from being able to predict an individual user's future tweet volume, our model also obtains reasonable accuracy when modeling aggregated volume from all users. For future works, we would like to devise topic-specific influence measures of social network users and use the extracted features in a non-linear regression model, as the non-linear correlation between features and regressand could be higher. Finally, we are also interested in investigating the performance of each single feature to see the effect across the feature dimensions.

Acknowledgement

This work is supported by NSF SoCS Grant IIS-1111118, "Social Media Enhanced Organizational Sensemaking in Emergency Response".

References

[1] E. Bakshy, J. Hofman, W. Mason, and D. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM'11*, pages 65–74. ACM, 2011.

[2] E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user-created content. In *EC'09*, pages 325–334. ACM, 2009.

[3] J. Gomide, A. Veloso, W. Meira Jr, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *WebSci'11*. ACM, 2011.

[4] A. Goyal, F. Bonchi, and L. Lakshmanan. Learning influence probabilities in social networks. In *WSDM'10*, pages 241–250. ACM, 2010.

[5] S. Kairam, D. Wang, and J. Leskovec. The life and death of online groups: predicting group growth and longevity. In *WSDM'12*, pages 673–682. ACM, 2012.

[6] M. Kutner, C. Nachtsheim, and J. Neter. *Applied linear regression models, 4th Edition*. McGraw-Hill New York, NY, 2004.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW'10*, pages 591–600. ACM, 2010.

[8] H. Purohit, Y. Ruan, A. Joshi, S. Parthasarathy, and A. Sheth. Understanding user-community engagement by multi-faceted features: A case study on twitter. In *SoME'11, workshop in conjunction with WWW'11*, 2011.

[9] D. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *WWW'11*, pages 695–704. ACM, 2011.

[10] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW'10*, pages 851–860. ACM, 2010.

[11] B. Suh, L. Hong, P. Pirolli, and E. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom'10*, pages 177–184. IEEE, 2010.

[12] C. Thomas, P. Mehra, R. Brooks, and A. Sheth. Growing fields of interest-using an expand and reduce strategy for domain model extraction. In *WI-IAT'08*, volume 1, pages 496–502. IEEE, 2008.

[13] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM'10*, pages 178–185, 2010.

[14] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM'10*, pages 261–270. ACM, 2010.

[15] F. Wu and B. Huberman. Popularity, novelty and attention. In *EC'08*, pages 240–245. ACM, 2008.