Analysing the Duration of Trending Topics in Twitter Using Wikipedia

Tuan Tran, Mihai Georgescu, Xiaofei Zhu, Nattiya Kanhabua
L3S Research Center / Lebniz University Hannover. Hannover, Germany
{ttran, georgescu, zhu, kanhabua}@l3s.de

ABSTRACT

Analysis trending topics in Twitter is the goldmine for a variety of studies and applications. However, the contents of topics vary greatly from daily routines to major public events, enduring from a few hour to weeks or months. It is thus helpful to identify trending topics related to real-world events or those originated within virtual communities. In this paper, we analyze trending topics in Twitter using Wikipedia as the reference for studying provenance of trending topics. We show that among different factors, duration of a trending topic characterizes well Twitter exogenous trending topics compared with the endogenous ones.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: [Information networks]

General Terms

Algorithms, Experimentation

Keywords

Twitter, Wikipedia, timeseries, temporal analysis

1. INTRODUCTION

Recently recognized as an important channel for instant updates about real-world incidents, information in Twitter often exhibits spikes during prominent events such as Super Bowl. Existing methods detect and track real-world events reported in Twitter typically through the volume of posts [3, 2]. However, the lack of contextual information from resources other than Twitter sphere makes these methods unable to identify whether trending topics truly reflect real-world events, or just a "virtual" topic that stay within Twitter only (e.g. spontaneous meme such as "#uFromLAif") . This makes systems misled with spam topics, while miss other potential events In this work, we propose a new framework to analyse the provenance of Twitter trending topics, exploiting background information from Wikipedia. Intuitively, information in Wikipedia is more creditable and focused as compared with Twitter. As previous work [3, 2], we propose using hashtag to predict the future behaviour of trending topics in Twitter. Unlike previous work, we predict how long it takes for a trending topic to saturate after the peak. Our exploratory analysis shows that duration is a stronger signal to indicate the long-term influence of a hashtag than peak volumes, and distinguish better endogenous and exogenous topics.

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.
ACM 978-1-4503-2622-3/14/06..

2. METHODOLOGY

Datasets. For the Wikipedia data, we obtained the English revision history dump on 30 Nov. 2012 (380 million updates of 4 million articles), and the Wikipedia page view log. We used TREC Tweets2011 corpus[1], which has 16 million public tweets sampled from 23 Jan. to 8 Feb. 2011. Volumes were aggregated to daily level.

Burst and Duration. To define a trending hashtag, we employed the simplified strategy to detect bursts in a time series as follows. For each time point t with the value $n(t)$, we look back at preceding k values, and claim t a peak if the current value is l-time standard deviations higher than the mean value of the preceding window: $n(t) \geq l\sqrt{(n(t-i)-\mu)^2} + \mu, i = \overline{1,k}$ where μ is the mean of k variables $n(t-i)$. We measure the duration as the distance (number of days) from the first peak to the closest day where the hashtag volume goes under a threshold τ. If the hashtag has several peaks, duration is the averaged out. We observe that $k = 3, l = 3$ and $\tau = 10$ give the most intuitive peak outcomes in Tweets2011.

We get only hashtags with more than 40 tweets in at least one day, and choose 628 random hashtags, amounting for 672,580 tweets. For each hahstag, assessors are displayed with the set of peak days and top 50 tweets on each day. The assessors then use keywords, mentions, abbreviations, etc. in the tweets and use the published days to issue to a search engine and Wikipedia. Each hashtag is annotated as whether the related information can be found on the Web (exogenous), and further whether it is found on Wikipedia (ongoing, otherwise breaking event). To prevent including future information in the Wikipedia dataset, we start from matching articles as seeds, process their revisions of the corresponding days, and include all outgoing linked articles to the study. In the end, we have 275 hashtags about endogenous topics, 353 about exogenous topics, in which 231 are

[1] `http://trec.nist.gov/data/tweets`

	FullSet		Endo		Breaking		Ongoing	
	WM	TwiNER+AIDA	WM	TwiNER+AIDA	WM	TwiNER+AIDA	WM	TwiNER+AIDA
Baseline	0.5865	0.5865	0.4167	0.4167	**0.6333**	**0.6333**	0.5444	0.5444
wstatic	0.7242	0.6912	0.6234	0.6190	0.6292	0.5801	0.5667	0.5590
wview	0.7284	0.6976	**0.7382**	0.7146	0.6333	0.5711	0.5667	0.5616
wedit	0.7383	0.6882	0.7355	0.7192	0.6333	0.5825	0.5731	0.5684
wstatic+wview	0.7355	0.7012	0.5612	0.6018	0.6250	0.5804	0.5625	0.5718
wstatic+wedit	**0.7411**	0.7134	0.6345	0.6129	0.6250	0.5727	**0.5778**	0.5645
wview+wedit	0.7346	0.7035	0.7337	**0.7682**	0.6333	0.5705	0.5670	0.5691
wstatic+wview+wedit	0.7374	**0.7276**	0.4333	0.4212	0.6250	0.5793	0.5767	**0.5792**

Table 1: Accuracy of hashtag staturation prediction in Tweets2011

breaking events (information found on the Web but not in Wikipedia in the peak day) and 122 ongoing topics.

Duration Prediction. We propose a framework that given a hashtag h peaked on day t_0, can predict the saturation length $L(t, h)$. As finding an exact value of L is difficult and often not necessary, we propose to classify the range which L falls in. For the Tweets2011 dataset (spanning 3 weeks), the range is defined as: [*1*] (last only 1 day), [*2-3*], [*4-7*] (last longer than 3 days to 1 week), [*7-14*] (last longer than 1 week to 2 weeks), [*14-21*] (last longer than 2 but less than 3 weeks), [*0 or 22+*] (last more than 3 weeks or the hashtag is continuous).

Entity Linking. For each hashtag h and the peak day t_0, we concatenate all the tweets in the order of published time, and use existing tools to link to a set of Wikipedia entities. For supervised approach, we use WikipediaMiner [2], and for the unsupervised approach, we use TwiNER [1] to identify entities in tweets, and AIDA[3] to disambiguate the entities.

Type	Features
Hashtag	(1) Hashtag length, (2) No. of segmented words in the hashtag, (3) (binary) if it has digits, (4) if it collocate with other hashtags, (5) no. of collocating hashtags, (6) fraction of capitalized characters in the hashtag
Tweets	(1)-(4) fraction of tweets having URLs/hashtags/ mentions/emoticons, (5)-(8) fraction of URLs/hashtags/ mentions/emoticons over tokens, (9) no. of distinct users, (10) average token length per tweet, (11) fraction of retweets, (12) 3-d emoticon vectors of tweets
Wiki static	(1) no. of matching Wikipedia articles, (2) no. of persons, (3) no. of locations, (4)-(5) maximal/average authority score of Wikipedia pages
Wiki Temporal	(1)-(4) if the edit/view count increase in all/any Wikipedia articles that match the hashtags, (5)-(8) minimum/maximal length of increase chains in view/edit count, (9) fraction of Wikipedia revisions that have URLs

Table 2: Features used for prediction

Model Features. We define 40 features, grouped in four categories as described in Table 2. The Hashtag and Tweets types are derived from previous work [3, 2] and used as the baseline. We propose several features extracted from matching Wikipedia entities to enhance the contextual knowledge. For instance, the authority score of an entity measures how importance it is w.r.t. to other entities: $authority(w) = \frac{|IN(w)|}{|OUT(w)|}$, with IN and OUT are incoming and outgoing link sets of the snapshot of article w on day t.

3. EXPERIMENTS

We conducted experiments on both the entire set of 48,803 hashtags (FullSet) in Tweets2011 and the annotated sample sets, each with the baseline and with Wikipedia feature types (static, view, edit) incrementally added. We used LibSVM[4] to train the classification model.

Result. Table 1 summarizes the accuracy of the classification in different feature settings. In the FullSet, for both entity linking systems, we see a clear improvement when incorporating Wikipedia information as feafures. Wikipedia edit history and Wikipedia structure information contribute the most to the increase in accuracy. Moreover, the performance of TwiNER+AIDA system is lower. This is explained by the fact that TwiNER is unsupervised and has inferior quality, and that AIDA is backed by the YAGO knowledge base, which only contains a subset of Wikipedia articles. Again, this emphasizes the importance of adding more information from Wikipedia to improve the prediction.

The performance varies in different kinds of trending topics. For endogenous topics, the result is unstable with both entity linking outcomes; adding different Wikipedia features sometimes harm the performance (although it does improve in general). This is because endogenous hashtags merely diffuse information within Twitter communities, and mentioned entities in tweets will thus not correlate well with the main content of the Twitter topic. For breaking topics, both systems do not gain any improvements with Wikipedia features, this confirms the fact that breaking events in Twitter are spreaded quicker than in Wikipedia. For ongoing topics, incorporating Wikipedia information does effectively improve the performance of the prediction in both entity linking settings. Method based WikipediaMiner performs best with Wikipedia static and edit features, and method based on TwiNER+AIDA performs best on the full combination.Last but not least, the general prediction performance of the systems can gain significant benefit when we increase the size of our data (from sample sets to FullSet). This positively supports the idea that despite the small size of the annotated dataset, our system does not overfit and has a good generability.

Acknowledgement. This work was partially funded by the European Commission for the FP7 projects CUbRIK and ForgetIT (under grants No. 287704 and No. 600826 respectively), and the ERC Advanced Grant ALEXANDRIA under grant No. 339233.

4. REFERENCES

[1] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: named entity recognition in targeted twitter stream. In *SIGIR*, pages 721–730, 2012.

[2] Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in Twitter. *JASIST*, 64(7):1399–1410, 2013.

[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.

[2] `http://wikipedia-miner.cms.waikato.ac.nz`

[3] `https://github.com/yago-naga/aida`

[4] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`