Discover Breaking Events with Popular Hashtags in Twitter*

Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, Kuo Zhang
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
cuianqi@gmail.com, {z-m, yiqunliu, msp}@tsinghua.edu.cn,
zhangkuo@sogou-inc.com

ABSTRACT

In this paper, we utilize tags in Twitter (the *hashtags*) as an indicator of events. We first study the properties of hashtags for event detection. Based on several observations, we proposed three attributes of hashtags, including (1) *instability* for temporal analysis, (2) *Twitter meme possibility* to distinguish social events from virtual topics or memes, and (3) *authorship entropy* for mining the most contributed authors. Based on these attributes, breaking events are discovered with hashtags, which cover a wide range of social events among different languages in the real world.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement

Keywords

Twitter, hashtag, social media, burst detection

1. INTRODUCTION

Twitter has been one of the most popular microblogging services for users to share their topics of interests. Twitter messages (*tweets*) cover many social events all over the world. However, its length limitation, informal texts and different languages ask for non-linguistic features for textual analysis. To help identify the topics, tags are introduced as in-line terms in Twitter. They are denoted with a leading hash symbol (#), thus called *hashtags*. Hashtags are typically used to mark keywords or topics in a tweet. Within the limited (no more than 140-character long) length, hashtags greatly draw researchers' attentions, as they provide useful metadata for the topic and annotation of the tweet. Therefore, aggregated hashtags can be seen as the indicator of a topic in a large number of tweets. The bursting of hashtags reflects the bursting of events from two sides:

On one hand, a specific hashtag may refer to different objects, i.e., ambiguous. For example, the hashtag *#tea* may refer to either the beverage or the Tea Party Movement. No matter which topic it stands for, the frequency of *#tea* increases when either *tea*-related event occurs.

On the other hand, different hashtags may describe a same event. For example, in August 2009, the typhoon Morakot attacked Taiwan, causing floods and brought two new hashtags *#taiwanfloods* and *#morakot*. In our sampled dataset, they only appear in the event period with relatively high frequencies, as shown in the stacked bar chart (Figure 1). Meanwhile, many related tweets contain *#taiwan*, a long-lasting hashtag used wherever Taiwan-related news appears. No matter which hashtag is captured, we are able to tell that breaking events take place during the peak periods. As hashtags are sensitive to wide topics, the topics they indicate are not necessarily *real-world events*, i.e., they are not with **a specific time period**, **location** or **people** involved. Based on our observations, many of the trending topics are *Twitter Memes*: conversational topics that attract users to share their own personal feelings. For example, *#iaintafraidtosay* (I ain't afraid to say) and *#foramilliondollars* (for a million dollars) are not related to any specific events. These memes successfully attract user's attention and become much more popular than the other hashtags (as shown in Figure 1),

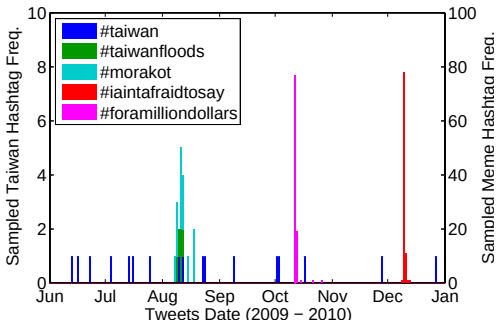

Figure 1: Frequencies of hashtags in a sampled dataset. The left y-axis is for #taiwan, #taiwanfloods, #morakot, the right axis is for #iaintafraidtosay and #foramilliondollars.

*Supported by Natural Science Foundation (60903107, 61073071) and National High Technology Research and Development (863) Program (2011AA01A207). This work has been done at the Tsinghua-NUS NExT Search Centre.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

though dying out very soon (called ephemeral in [8]). The memes are less valuable than the real events which take place at some locations during a time period. However, most of the previous studies fail to distinguish Twitter memes from breaking events. Instead, they treat them as trending topics.

In this paper, we study some event-related properties of hashtags, including temporal trends, authorships and pattern of texts. Then we examine the popular hashtags to discover breaking events and distinguish them from Twitter memes, spams, etc. An unsupervised algorithm is developed for estimations on large scale of tweets.

2. RELATED WORK

Tags are first brought to the Web applications by the social bookmarking website *del.icio.us* in 2003. Since then, many websites deploy tags to help users find content, such as Flickr (images) and YouTube (videos). These tags help topic retrieval effectively [9, 18]. Tags have been used for personalized recommendation and discovering users' interests as well [12], which is beyond the scope of this paper. Here we aim at the wisdom of the crowds (macroscopic) to look into the usage of hashtags, mainly on events detection.

Tags in Twitter (*hashtags*) are freely chosen by users, and attract researchers' attentions in diffusion of innovation [2, 11] or user interestingness [15]. When considered as symbols of trending topics [19], their evolutions and propagations are examined [5, 13]. However, their typology for different purposes varies from subjects [11] to functionalities [1].

The frequency of a hashtag (as a topic) is usually tracked for temporal analysis [8], whose patterns are divided into four categories [3]: whether the factor behind an event is "endogenous" or "exogenous", and whether a user can spread the news about the event to others or not ("critical" or "subcritical"). However, they only track a limited set of top hashtags in Twitter's trending topics. Another study [13] predicts the frequencies of hashtags by bringing in topological features, but they do not distinguish hashtags reflecting social events from Twitter memes.

Traditionally, bursts in text streams are detected with word-based methods [10], modeled by patterns [6, 16]. These studies are word-based which is limited with languages. To the opposite, if we discover the event from hashtags (then to words), it is more likely that it is a real event, since hashtags are mostly trusted annotations.

3. CHARACTERISTICS OF HASHTAGS

In this paper we use two sets of tweets, sampled from six months and three months respectively. The six-month set (hereinafter, *Tweets6*) is sampled from the Stanford's SNAP data [16], ranging from June to December 2009. The three-month set (hereinafter, *Tweets3*) is sampled by ourselves using the Twitter Streaming API from December 2011 to February 2012. The two corpora originally have more than two millions of tweets per day, but we sampled 10,000 daily. Both sets contain multilingual tweets. A data cleaning step is applied to remove invalid and duplicate tweets.

Table 1 lists some statistics of the datasets and hashtags. From the statistics we learn that (1) people use hashtags more often in 2011 (13.7%) compared with 2009 (10.5%), hence this indicator is becoming richer for analysis. (2) Hashtags hardly co-occur in tweets (only 2.4%) because the tweet length is limited. Hashtags are in-line terms; they

Table 1: Dataset Statistics

Dataset	*Tweets6*	*Tweets3*
Cleaned sampled tweets	2,028,794	778,395
Tweets containing hashtags	212,032 (10.5%)	106,327 (13.7%)
Tweets containing more than one hashtag	49,278 (2.4%)	18,654 (2.4%)
Tweets containing more than two hashtags	18,064 (0.9%)	6,293 (0.8%)
Unique hashtags	66,066	62,161
Total hashtag frequency	296,895	138,758

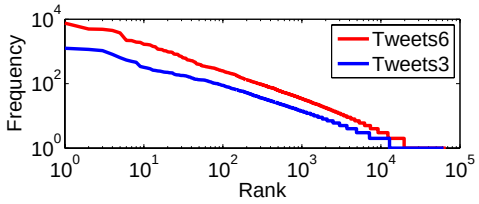

Figure 2: Frequency to the rank of the hashtags.

are more expensive to appear in one tweet. (3) The number of tweets with more than two hashtags are smaller (as 0.9% → 0.8%, or 8.5% → 5.9%). The individual hashtags appear to be more meaningful.

Sorting by the frequencies of the hashtags, their ranks to their corresponding frequencies is of power-law distribution [7] (Figure 2). As *Tweets6* covers a twice longer period than *Tweets3*, the frequencies of the most popular hashtags in *Tweets6* are much bigger than the ones in *Tweets3*, while the numbers of unique hashtags remain almost the same. This illustrates that, though the hashtags themselves change (to reflect different topics), their macroscopic characteristics are relatively stable over time. Hence we raise three questions on the three aspects of hashtags.

(1) Do popular hashtags reveal breaking?

Intuitively, popular hashtags are related to topics which most people concern. These topics include both breaking events and persistent discussions. Figure 3 demonstrates three popular hashtags in *Tweets3*: (1) *#sopa* short for "Stop Online Piracy Act", a controversial U.S. bill between online intellectual-property protection and free speech and innovation threats. The peak on January 18th is the date when many major Internet sites committed to an Internet blackout to protest against the bill. (2) *#ff* ("Follow Friday"), a Twitter meme to suggest whom to follow on Fridays. (3) *#nowplaying* as "now playing", to broadcast the music a user is now playing. These two aggregate higher frequencies but are less important for discovering the breaking events.

Under these observations, hashtags with respect to breaking events have unexpected changes (mainly increasing) of frequencies. The increment is almost impossible based on previous observations. Hence, we introduce *instability*, i.e., how likely the hashtag has a sudden increase or decrease.

(2) Do popular hashtags indicate events or memes?

Originally, hashtags are used from users' discussions in Twitter, which are mainly relevant to the real-world events. As time goes on, some memes come into being in Twitter, introduced in hashtags as distinguishable labels. For example, *#musicmoday* (suggest music to people on Mondays) and *#ff*, *#followfriday* are all conversational topics. Other examples include sentence leading phrases like *#cantlivewithout*, *#factsaboutme*, etc. This type of hashtags is referred to as Twitter Memes [17] or Idioms [11]. Although

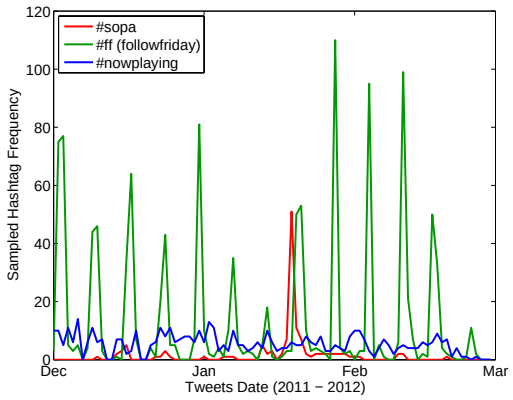

Figure 3: Trends of three hashtags in *Tweets3*.

Table 2: Popular Hashtags with Highly Contributed Authors in *Tweets6*. STF: Sampled Hashtag Freq., #TC: No. of Tweets by the Top Contributor

Hashtag	STF (#TC, %)	# Authors
#nieuws	198 (105, 53.0%)	10
#nl	110 (72, 65.5%)	8
#property	86 (66, 76.7%)	19
#praytweets	57 (56, 98.2%)	2
#indonesia	57 (35, 61.4%)	22

they are irrelevant to real events, the topics are introduced suddenly in Twitter, bringing in some quick rising.

Twitter memes are usually about personal feelings. These hashtags are usually formed by a concatenation of common words to make them unique, distinguishable from real-world events. Moreover, people join these topics when they see other people post them, i.e., they are motivated from external rather than internal (from their mind). Hence people tend to write the hashtag at the beginning of the tweet, to "join" this discussion proactively [14]. Considering these two aspects, we introduce the *Twitter meme possibility* to tell the difference between the memes and event topics.

(3) Are popular hashtags contributed by the crowds?

Some of the popular hashtags are contributed by only a few authors. These "top" authors contribute more than half of the tweets containing a certain hashtag. The hidden fact is that many of these users are spammers, i.e., automated agents who publish advertisements on Twitter. On the other hand, some of them are news publishers, who work online all day long posting news tweets. Table 2 lists some of the typical hashtags in this category. For example, the *#nieuws* ("news" in Dutch), *#nl* ("Netherlands") and *#indonesia* are all contributed by news sources.

Hence, the authorship of hashtags is helpful for tweet analysis, esp. for detecting automated agents and robots [4]. In our algorithm, we take the *authorship entropy* as a measurement on how concentrated the contributed authors are.

4. HASHTAG ATTRIBUTES AND CATEGORIZATION

Considering the previous three questions, we introduce three attributes to develop a categorizing algorithm.

Hashtag Instability is how unlikely the hashtag keeps a stable amount based on previous observation. In this paper, we derive a measurement from probability for *instability*.

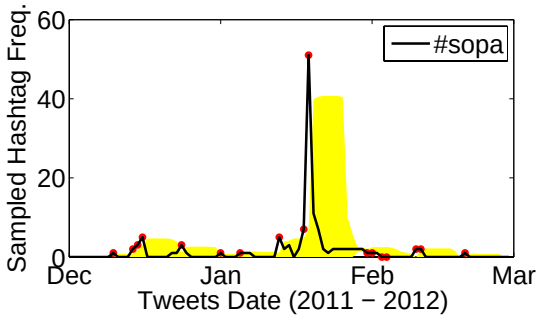

Figure 4: Trends of *#sopa* with a 90% confidence interval (yellow shade) in *Tweets3*. Points out of this interval are in red.

Assume a random variable X as the frequency of a hashtag H in a time period, with a Gaussian distribution. When we observe a sample x in one time period, the probability $Pr(X = x)$ can be estimated. The frequency is modeled around a mean μ, thus the probability $\tilde{P}$ of x away from μ (without loss of generality, let $x > \mu$) is computed as:

$$\tilde{P}(x) \quad = \quad Pr(X > x \bigvee X < 2\mu - x) \tag{1}$$

The $(2\mu - x)$ is the symmetric position of x away from μ.

We aim to discover the most instable (i.e., smaller probability than a threshold p) situations, hence we focus on the x's whose $\tilde{P}(x) < p$. The illustration of *#sopa* is shown in Figure 4, where the probability distribution is a Gaussian distribution on a daily period. Seven-day periods are used for estimating the probability parameters.

Then we define $Inst(\cdot)$ as the *instability* for each of the out-of-bound x_i's and the hashtag H (given p):

$$Inst(x) = -\log \tilde{P}(x), Inst(H) = \frac{1}{n} \sum_{\tilde{P}(x)<p} Inst(x) \tag{2}$$

where n is the days covered in the set for normalization.

Twitter Meme Possibility (TMP) of a hashtag is defined as how likely it indicates a Twitter meme. Based on our observations, the "word length ratio", i.e., the number of real English words N divided by the length of the hashtag L, and the probability of its appearing at the beginning of a tweet are both good measurements for this possibility.

The words are determined by word splitting. We use a dictionary containing more than 80,000 common English words from *12Dicts*. A dynamic programming method for word splitting is applied to obtain the minimal number of words N in the hashtag. Consecutive letters not in the dictionary are split into individual letters. For example, given a dictionary of {art, arthur, thursday, day}, the hashtag *#arthursday* is split into three words, either a/r/thursday or arthur/s/day, where the semantic information is not important. This measurement generates a probability $p_{\text{word}} = 1 - N/L$.

For hashtag's position, we make an estimation on the sampled tweet set, i.e., given a hashtag h,

$$p_{\text{pos}} = \frac{|\{\text{tweets starting with } h\}|}{|\{\text{tweets containing } h\}|} \tag{3}$$

Finally, $TMP(hashtag) = p_{\text{word}} \cdot p_{\text{pos}}$.

Authorship Entropy measures how concentrate the contributors are, and is defined similarly as the entropy in Information theory. Taking all the n tweets containing a hash-

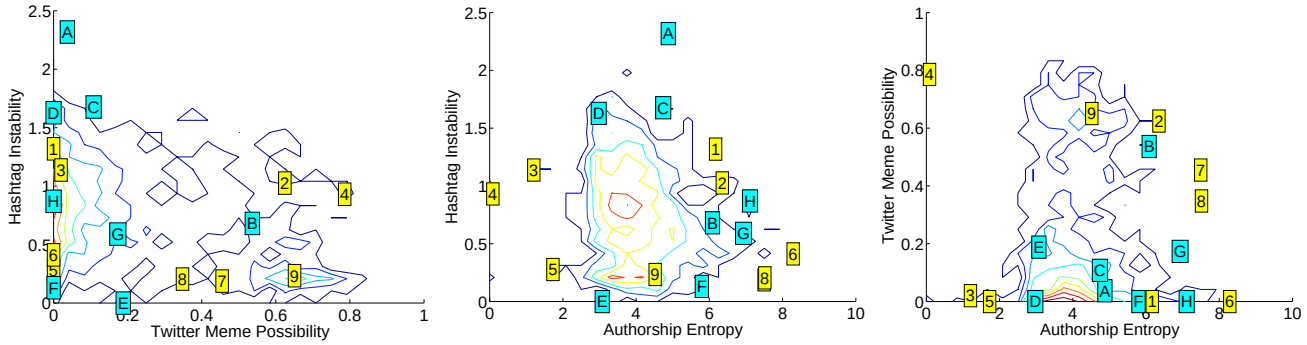

Figure 5: Contour of hashtag distributions. *Tweets6*: 1–#hcr, 2–#nowplaying, 3–#property, 4–#praytweets, 5–#abbeydawn, 6–#fb, 7–#musicmonday, 8–#followfriday, 9–#iaintafraidtosay. *Tweets3*: A–#sopa, B–#nowplaying, C–#halamadrid, D–#belieber, E–#bring1dtonyc, F–#fb, G–#teamfollowback, H–#ff

Table 3: Categories of Hashtag Subspaces. L=Low, H=High. A=Advertisements, M=Miscellaneous, T=Twitter Memes, B=Breaking Events

Inst.	TMP	Ent.	Cat.	Inst.	TMP	Ent.	Cat.
L	L	L	A	H	L	L	A
L	L	H	M	H	L	H	B
L	H	L	A	H	H	L	A
L	H	H	T	H	H	H	T

tag, the k authors of these n tweets contribute $c_1, c_2, \ldots, c_k$ ($\sum c_i = n$) times, respectively. The *authorship entropy* is:

$$Ent(hashtag) = -\sum_{i=1}^{k} \frac{c_i}{n} \cdot \log\left(\frac{c_i}{n}\right) \quad (4)$$

Categorization. The hashtag *instability*, *Twitter meme possibility* and *authorship entropy* are three orthogonal dimensions which are independent from each other. Considering each feature a lower or a higher value, the hashtag space is divided into eight subspaces, listed in Table 3. Example hashtags are plot in the contour map in Figure 5.

5. EVALUATIONS AND DISCUSSIONS

The experiment setup and results are shown in this section by comparing with ground truth and the baseline method.

5.1 Experiment Setup

As from Table 1, we consider the top 1% hashtags (about six hundred) as popular hashtags. Then we randomly sampled 250 hashtags from *Tweets6* and *Tweets3*, respectively. The judgement on the categories is from the observed tweet contents in the dataset. We create the ground truth by labeling the hashtags from two different annotators. A third annotator is introduced when the two labels are different. If there is still no majority, the hashtag is considered ambiguous and is not included. Finally, the evaluation set contains 191 and 200 hashtags in *Tweets6* and *Tweets3*, respectively.

As ambiguous hashtags are located in the center of the hashtag space, the mean values of each of the dimensions – this centric point is used to divide the space into subspaces.

The probabilistic distribution of *instability* is modeled by a Gaussian distribution whose mean and variance is estimated from a seven-day-before period, which reduces the affects brought by weekly topics. The probability p is set to 0.01 as the threshold for a 10% probability.

In addition to the ground truth, we look for other baseline methods. However, many recent literatures do not consider hashtags as indicators of events or classify them manually. There is no direct comparison available. Hence we implement the algorithm in [3, 8] which classifies hashtags into *popularity patterns*. They claim that the exogenous subcritical, exogenous critical and endogenous critical classes are mainly consisted of Twitter idioms (memes), breaking news, persistent news (or entities), respectively. However, this method is not able to discover advertising hashtags, since the author information is not taken into consideration.

5.2 Results and Discussion

The results of different algorithms on two sets are listed in Table 4. Best results are in **bold**. Note that a random classifier has an accuracy of 25%. The table shows that our subspace-based algorithm outperforms in most cases.

The ***Authorship Entropy*** helps discover popular hashtags contributed by spammers. Hashtags with a lower *Authorship Entropy*, i.e., with some highly contributed authors, are mostly promotional. In Table 3, four categories are all considered as advertisements (spams). This is also the first step of judgement in the *Subspace* algorithm. Some Twitter memes are classified to Ads due to their "marketer" like behaviors. No matter these hashtags are with a lower or higher *instability*, they are not related to real events.

Popular hashtags with higher *entropy* are coming from the crowd. The ***Twitter Meme Possibility*** feature come into help to distinguish Twitter memes from events. Note *#nowplaying* has different *instability* scores in *Tweets6* and *Tweets3*. The reason is that in *Tweets6* it suddenly increases in the last month, but in *Tweets3* it stays steadily (shown in Figure 3). This is also the best category our algorithm performs on. Some less popular Twitter memes may be polluted by other agents, hence some more features (URLs, co-occurred hashtags) may be helpful to discover all the memes.

Another type of the Twitter hashtags, who has lower *Twitter Meme possibilities*, are mainly "idioms" – some traditions on Twitter, e.g., suggesting whom to follow (*#followfriday*), what to listen (*#musicmonday*), etc. As traditions, they are popular but not breaking. Therefore, the ***instability*** feature helps us discover this type of memes. Only if the hashtag has a higher *instability* is it considered to indicate a breaking event, such as *#hcr* ("Health Care Reform", a social movement for health policy creation or changes in the

Table 4: Experiment Results (Precision, Recall and F-Measure) of Hashtag Categories

Dataset	*Tweets6*						*Tweets3*					
Algorithm	Popularity Pattern			Subspace			Popularity Pattern			Subspace		
Accuracy	17.8%			**40.0%**			31.5%			**38.0%**		
Breaking events	0.250	**0.231**	0.240	**0.333**	0.205	**0.254**	**0.192**	**0.192**	**0.192**	0.167	0.154	0.160
Twitter memes	0.000	0.000	0.000	**0.681**	**0.595**	**0.635**	**1.000**	0.060	0.113	0.725	**0.248**	**0.369**
Advertisements	–			**0.258**	**0.370**	**0.304**	–			**0.053**	**0.385**	**0.093**
Miscellaneous	**0.162**	**0.926**	**0.276**	0.125	0.148	0.136	**0.240**	**0.909**	**0.379**	0.220	0.205	0.212

U.S.), *#sopa*, and *#halamadrid* (for cheering up to the Real Madrid football club during the game, esp. on the goal moment). This category, occupying only one subspace in our division, is treated conservatively. All other popular but ambiguous hashtags are categorized as miscellaneous.

Setting a class of miscellaneous is not always good in a classification problem. However, many hashtags are indeed ambiguous, even after examining the related tweets. Algorithms may have different strategies of determining a hashtag as "miscellaneous" as well, hence their performances differ from each other.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we utilize the *instability*, *Twitter memes possibility* and *authorship entropy* of the hashtags to help classify them into different categories. Within the category of "breaking events", the hashtags are sensitive to the breaking events in the real world; while some other popular but not practical topics, i.e., Twitter memes and idioms are filtered out. This method is able to remove the advertising hashtags, which may cause false alarms on events.

As future work, we can investigate more features for better measurement of the attributes, to discover more on the ambiguous hashtags. The large amount of advertisement hashtags may be helpful to filter spam tweets. As a language independent method, it is capable to discover breaking events all over the world, regardless of the tweets' languages.

7. ACKNOWLEDGMENTS

We appreciate Cheng Luo, Bin Liang, Xin Li, Qi Fang, Shuai Huo and many other annotators. The NExT Search Centre is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

8. REFERENCES

[1] I. Cantador, I. Konstas, and J. M. Jose. Categorising social tags to improve folksonomy-based recommendations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):1 – 15, 2011.

[2] H.-C. Chang. A new perspective on twitter hashtag use: diffusion of innovation theory. In *ASIS&T '10*, pages 85:1–85:4, 2010.

[3] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. of the National Academy of Sciences*, 105(41):15649, 2008.

[4] A. Cui, M. Zhang, Y. Liu, and S. Ma. Are the urls really popular in microblog messages? In *CCIS '11*, pages 1–5, 2011.

[5] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *NAACL-HLT Workshop LSM '11*, pages 58–65, June 2011.

[6] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501, 2004.

[7] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW '07*, pages 211–220, 2007.

[8] 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, 2010.

[9] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su. Towards effective browsing of large scale social annotations. In *WWW '07*, pages 943–952, 2007.

[10] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *SIGMOD '10*, pages 1155–1158, 2010.

[11] D. M. 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*, pages 695–704, 2011.

[12] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys '08*, pages 259–266, 2008.

[13] O. Tsur and A. Rappoport. What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *WSDM '12*, pages 643–652, 2012.

[14] A. Wang, T. Chen, and M.-Y. Kan. Re-tweeting from a linguistic perspective. In *NAACL-HLT Workshop LSM*, pages 46–55, 2012.

[15] J. Weng, E.-P. Lim, Q. He, and C. W.-K. Leung. What do people want in microblogs? measuring interestingness of hashtags in twitter. In *ICDM '10*, pages 1121–1126, 2010.

[16] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM '11*, pages 177–186, 2011.

[17] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a twitter network. *First Monday*, 15(1):1–13, 2010.

[18] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *WWW '08*, pages 715–724, 2008.

[19] A. Zubiaga, D. Spina, V. Fresno, and R. Martínez. Classifying trending topics: a typology of conversation triggers on twitter. In *CIKM*, pages 2461–2464, 2011.