doi:10.3772/j.issn.1006-6748.2016.02.006

Learning trendiness from twitter to web: a comparative analysis of microblog and web trending topics①

Wang Dong (王 栋)②*** , Xie Gaogang*
(*Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, P. R. China)
(**University of the Chinese Academy of Sciences, Beijing 100049, P. R. China)

Abstract

The development of microblog services has a considerable effect on the patterns of web access and Internet resources discovery. Understanding the interrelation between information diffusion in online social media and user web interests can help the web ecosystem stakeholders in developing new services and designing efficient systems with optimized resources. This paper explores whether or not one can infer the trends of topics in the web by observing the Twitter microcosm. Using datasets collected from Twitter and two representative web services (Google and Alexa), this work conducts a comparative analysis between trending patterns of topics in Twitter and in the web by considering both the temporal and spatial perspectives, and finds that individual topics in Twitter and in the web share similar trending patterns both from the temporal and spatial aspects. Nevertheless, the trendiness in Twitter can precede for a few hours and is highly unstable compared to the one in web. The application of these findings is also discussed on ad keywords planning in Search Engine Marketing.

Key words: microblog, web services, trendiness, comparative analysis

0 Introduction

Microblog services have dramatically changed the way that users discover content and consume information on the Internet[1]. Several studies proposed to exploit the popularity of URLs shared on social networks to predict the actual popularity of the linked content[2,3]. Given the importance of the prominent feature of microblog, this paper raises the question about whether or not the user interests in a web content can be inferred by observing the content trends in microblogs.

While the popularity of content reflects the long term importance of a content, trends and in particular positive trends (referred as trendiness) express arising and short-term interests. More specifically, this paper studies trendiness of topics in Twitter and compares it with the trendiness of contents in Google and Alexa, two representative services in the web sphere. This work focuses on highly positive trends and their corresponding trending topics which attract relatively higher interests within a short period of time.

As there is no absolute metric that captures the interests within the web spheres, web interests are defined as the extent to which web resources (i.e. webpages) are being used or searched in the Internet. Specifically, interests in the web sphere are measured as the relative number of users who search for a particular web content using a set of keywords through search engines (e.g. Google). In addition, web interest is also measured by the audience of webpages relying on statistics provided by Alexa[4]. On the other hand, a nature language processing approach is used to extract topics of interest in microblogs and measure trends in the microblog sphere based on a dataset of tweets collected from Twitter. Furthermore, official trend statistics provided by Twitter are also used.

This work compares the trendiness of topics in Twitter with the ones in Google, Alexa from both temporal and spatial aspects. In detail, the temporal evolution of trendiness in Twitter and their interrelation with web trends is first examined. The likelihood that a Twitter trending topic is also the trending in web (as illustrated by Google searches) is measured, and the temporal offset between trendiness in both spheres is characterized.

The evidences that trending topics share certain similar patterns within the two spheres are found in this work. It is observed that more than 70% of the tren-

① Supported by the Beijing Municipal Natural Science Foundation (No. 2015AA010201).
② To whom correspondence should be addressed. E-mail: wangdong01@ict.ac.cn
Received on Mar. 17, 2015

ding Twitter topics are likely to be also trending in the web and more than 72% of the web trending topics have been (or will be) also trending in the Twitter sphere. The results also suggest that trendiness seems to be in most cases originating from the Twitter sphere, with more than 65% of the topics trending in Twitter first for a few hours. A notable difference is that trendiness in Twitter is highly unstable as the topic rank stability changes frequently.

Secondly, the analysis is extended to spatial aspect of trendiness by observing trending topics across five countries. It is found that both in Twitter and in web, most trending topics obtain trends in not more than 2 countries and for a topic, the trending regions in the two spheres are similar, which advocates for a regional feature of trends.

Based on these observation, it can be concluded that it is possible to learn trending topics in web from Twitter. Even better, one can learn them a few hours earlier than the time they will get popular in web. This paper confirms this with a detailed experiment, in which the possibility of using trending topics from Twitter to infer ad keywords in Google AdWords, a widely-used online platform for Search Engine Marketing (SEM), is shown.

The structure of this paper is as follows. Section 1 describes the datasets used in this study. Section 2 analyzes the temporal interrelation between trends in microblogs and across the web. Section 3 studies the spatial dimension of such an interrelation between the two spheres. In Section 4, possible applications of the study are discussed. Section 5 introduces related work. Finally, Section 6 concludes this paper.

1 Methodology and dataset description

This section first describes the methodology used to infer the trending topics from tweets in Twitter as well as in Google and Alexa, and two popular sites provide trends in the web sphere. The metrics which are used to measure the trendiness of topics are also introduced. Finally, the datasets used for analysis are detailed.

1.1 Identifying trends

Trends describe the popularity dynamics of topics over a short time period, where topic c consists of a word or a sentence mentioned in tweets or queried using search engines. While a single-word topic might be easy to obtain from tweets or queries, multi-words topics should be learnt using some natural language processing methods, e.g. LDA.

1.1.1 Trending index volume

User interests in both microblog and web spheres have temporal dynamics. That is to say, the volume of mentions or searches for a particular topic naturally varies over time. The trending index volume $V_i(c)$ for topic c at a given time i is defined as the volume of the topic normalized by the maximum volume observed during an observation period of time and then scale the trending index volume to $[0,100]$, which is similar to the official definition provided by Google[5]. Over given period R, all trending index volumes $V_i(c)$ where $i \in R$ compose the trends of topic c during that period, $V(c)$, i.e. $V(c) = \{V_i(c), i \in R\}$. This study further uses $V^G(c)$ and $V^T(c)$ to represent the trends of topic c in Google and Twitter.

Extracting the trending index in Twitter is a challenging task, as it needs to extract the global trending topics over a particular period of time. Although Twitter offers an official trending service, the trends are determined by an "algorithm tailored for the user based on who [you] follow and [your] location. This algorithm identifies topics that are immediately popular, rather than topics that have been popular for a while or on a daily basis, to help [you] discover the hottest emerging topics of discussion in Twitter that matter most to [you]"[6]. In other words, Twitter official trending topics are personalized to user accounts. Therefore an alternative approach is adopted to extract global Twitter trends.

A topic consists of a single word or multiple words. For a single-word topic that includes only one word w, the trending index is measured as the word frequency based on the content of tweets. All tweets are binned into subsets S_i with a fixed time interval (daily and hourly in this study) and then for each subset S_i, the set of words W_i is extracted, and the word frequency $TF_i(w)$ for each word $w \in W_i$ is computed. Note that stop words (e.g. "a", "after", "that", etc.) which naturally appear with higher frequencies are ignored here[7]. A word is counted once per tweet even if it is repeated in the tweet. Since the number of tweets in each subset might vary greatly, the word frequency $TF_i(w)$ is normalized by the number of tweets in each subset, resulting in a relative topic frequency $RF_i(c) = TF_i(w)/|S_i|$, where $|S_i|$ is the number of tweets in subset S_i. Finally, all $RF_i(c)$s are scaled in $[0,100]$. The Twitter trending index volume for single-word topic c at time i in a period R can then be written as: $V_i(c) = RF_i(c)/\max_{j \in R}(\{RF_j(c)\}) \times 100$.

To obtain multi-words topics in Twitter, Latent

Dirichlet Allocation (LDA) is used[8]. LDA is a generative model that extracts statistical properties of text documents in a discrete dataset and models each document as a mixture of various latent topics. A topic created by LDA is always nameless and represents a cluster of words that tend to co-occur with a high probability within the topic. LDA learns the statistical relations among words and documents and then estimates the probability that a given document is related to a given topic. The total number of topics is denoted by k, a parameter of the LDA model. Supposing there are M documents in the copus and each document i includes N_i words, the topic distribution θ_i for each document i is described to follow a Dirichlet distribution $D(\vec{\alpha})$, where $\vec{\alpha}$ is a parameter vector of the Dirichlet prior with a size of k. In addition, the word distribution ϕ_z for a topic z also follows a Dirichlet distribution $D(\vec{\beta})$, where $\vec{\beta}$ is another parameter vector of the Dirichlet prior. Given the parameters and $\vec{\beta}$, the generative process for each document by LDA contains the following three steps:

1) Choose the topic distribution for a document θ_i from $D(\vec{\alpha})$, where $i \in \{1, \cdots, M\}$;

2) Choose the word distribution for a topic ϕ_z from $D(\vec{\beta})$, where $z \in \{1, \cdots, k\}$;

3) For each of word position j in document i, where $j \in \{1, \cdots, N_i\}$ and $i \in \{1, \cdots, M\}$:

3a) Choose a topic $z_{i,j}$ from $M(\theta_i)$ where $M(\theta_i)$ is a categorical random variable with parameter θ_i.

3b) Choose a word $w_{i,j}$ from $M(\phi_{z_{i,j}})$ where $M(\phi_{z_{i,j}})$ is a categorical random variable with parameter $\phi_{z_{i,j}}$.

Following the above process, the total probability of the model is:

$$P(\vec{W}, \vec{Z}, \vec{\theta}, \vec{\phi} \mid \vec{\alpha}, \vec{\beta}) = \prod_{i=1}^{M} P(\theta_i \mid \vec{\alpha}) \prod_{j=1}^{k} P(\phi_j \mid \vec{\beta}) \prod_{t=1}^{N} P(z_{i,t} \mid \theta_i) P(w_{i,t} \mid \phi_{z_{i,t}}) \tag{1}$$

where $\vec{W}$ is the set of words in all documents, $\vec{Z}$ is the set of topics in all documents, $\vec{\theta}$ is the distribution vector with size M of which item θ_i represents the topic distribution in document i, and $\vec{\phi}$ is the distribution vector with size k of which item ϕ_z represents the word distribution in topic z, N represents total number of words in all documents, that is, $N = \sum_{i=1}^{M} N_i$. The observable variable is while $\vec{\alpha}$, $\vec{\beta}$, $\vec{\theta}$ and $\vec{\phi}$ are latent variables. Note that Eq. 1 describes a parametric empirical Bayes model and one can derive various distributions (e. g. the associated word probabilities in a topic, the probability that a document belongs to a topic) uses Bayesian inference. Gibbs sampling is widely-used to recover the posterior marginal distribution of $\vec{\theta}$.

In the context of this paper, all tweets are binned into subsets S_i with a fixed time interval (hourly or daily). In the training process, each tweet is considered as a document and each subset S_i as a corpus of documents. For each corpus, LDA is used with 2,000 iterations of Gibbs sampling to extract 50 topics, each of which includes 20 relative words. For each training process over S_i, LDA model provides a probability vector for each tweet, the elements of which indicate the correlation between the tweet and the extracted topics. Based on this probability vector, a tweet can be considered to be related to the topic of which the corresponding probability is the highest in the vector, resulting in the relative topic frequency $RF_i(c)$ (i.e. the proportion of tweets related to c in S_i). Then, the Twitter trending index volume $V_i(c)$ for multi-words topic c at time i can be calculated by scaling $RF_i(c)$ within $[0, 100]$.

The trending index volumes for topics in Google is much easier to be obtained, as Google Trends provide the normalized search volume for both single-word and multi-words topics. These statistics can be used as the trending index volumes in Google directly.

However, it is hard to get the exact search volumes of topics from Alexa. Alternatively, the trendiness of topics in Alexa is estimated approximately with the assistance of topic rank information: the trend of topic c is considered in binary, that is, if topic c appears in the top trending list of Alexa at time i, then the trending index volume of c at i is 100, otherwise, it is 0. Clearly, a sharp rise can happen on Alexa at time i if c is in the top trending list at time i but not at time $(i-1)$.

1.1.2 Positive and negative trends

Topic c experiences a positive (resp. negative) trend at time i if its trending index value $V_i(c)$ is larger (resp. smaller) than $V_{i-1}(c)$. The corresponding increasing (resp. decreasing) trending index volume $V_i^+(c)$ (resp. $V_i^-(c)$) is $V_i(c) - V_{i-1}(c)$ (resp. $V_{i-1}(c) - V_i(c)$). For topic c, highly positive trend is defined as a positive trend that has an increasing trending index volume larger than a threshold α at the time of observation. The time of observation i is called highly positive time (day or hour) of the topic.

In this study, α is set to the 50$^{\text{th}}$ percentile, 75$^{\text{th}}$ percentile and 90$^{\text{th}}$ percentile of all positive trending index volumes in $V(c)$ respectively.

1.1.3 Trending topics

Trending topics are topics in which trending index

volume increases in a relatively higher proportion compared to others. In other words, a trending topic can be either a word, an expression (a set of concatenated words) or a tweet in which the immediate popularity is rapidly increasing, compared to other popular topics. The emergence of trending topics is either endogenously driven by users interests, or motivated by an exogenous event that prompts people's attention.

A trending topic at time i is identified as follows:

1) A discrete-time vector of trending index volumes for each topic c is derived, from which all positive trends can be extracted.

2) For each positive trend (of all topics), the corresponding increasing trending index V_i^+ is measured and then the average value of all increasing trending index volumes at time i, $\overline{V}_i^+$, is calculated.

3) If at particular time i, a positive trend of topic c is observed, $V_i^+(c) \geqslant \overline{V}_i^+$, then topic c is deemed trending at time i. Time i is called trending time (day or hour) of topic c.

Again, it is noteworthy that the notion of "trending" is different from "popular". The latter is highly dependent on the number of times the topic is mentioned, e.g. the number of relative tweets in Twitter or search volume in Google, across a rather long period of time, while trendiness focuses on the speed of increase in mentioning a topic within a short period of time. A topic that has been popular for a while is most likely to be not trending anymore, as the number of tweets mentioning this topic would become steady even though it is still high.

1.2 Datasets

For the purpose of this study, Twitter's tweets are used to extract the trends of topics in Twitter. The "official" trending topics as shown by the Twitter are also relied on for geographical pattern analysis. Google Trends and Alexa services are also used to obtain trends of the web sphere.

1.2.1 Twitter tweets

This paper uses a set of tweets T from Ref. [9] comprising 132,210,436 tweets published by 7,404,248 users over the period from August 1st, 2009 to August 31st, 2009. Two time granularities are considered: a daily topic analysis which matches the Google Trends service time granularity[10], and a topic extraction on an hourly basis which matches the Alexa trends analysis. As in Ref. [11], it is observed that the frequency of the top 5% popular words accounts for more than 95% of words count in the overall daily and hourly subsets of tweets T.

Daily (resp. hourly) single-word topics are extracted using simple term frequency statistics in order to extract the most relevant (top 5%) words on a daily (resp. hourly) basis. To extract multi-words topics, a LDA generative model is used as described above to classify them into different topics. In total, the daily set of topics, denoted as K_d^T, is composed of 76,760 single-word topics and 267 multi-words topics, while the hourly set of topics K_h^T is composed of 56,774 single-word topics and 372 multi-words topics.

1.2.2 Official Twitter trending topics

The tweets described above do not provide enough geographic information. In order to analyze the geographic patterns of Twitter trending topics, this work further collects for the period spanning from September 1st to October 31st, 2012, and every five minutes, the top 10 trending topics are suggested by Twitter for the following countries: U.S, U.K, Canada, France and Australia, which are abbreviated to *US*, *UK*, *CA*, *FR* and *AU* later. There are 6,858 unique trending Twitter topics, which compose a topic set H.

1.2.3 Google trends

For the purpose of the temporal analysis, the Google Trends statistics of the topics extracted from Twitter are collected. Using Google Trends, a dataset can be got, referred as G, which includes scaled and normalized daily Google search volumes for each topic $c \in K_d^T$ from August 1st to August 31st, 2009. In addition, in order to have a comparison study of geographical patterns in Twitter and Google Search, this work also collects the lists of top 10 countries where the topic $c \in H$ is the most frequently searched topic according to Google Trends from September 1st to October 31st, 2012.

1.2.4 Alexa rank lists

Although Google Trends provides the trending topics on Google[10], it is not suitable for the web trending topics collection mainly for two reasons. First, Google Trends service only offers the top 10 trending topics per day which are far from enough to compose a complete trending topics set. Second, Google Trends provides the daily trending topics of specific countries but not the global ones. Fortunately, some informative websites such as Alexa provide the information about global trending topics in the web with a fine granularity. Specifically, Alexa keeps track of the top 20 global trending topics (search keywords) in the web for any hour since July 26th, 2009. This provides an effective way to estimate the hourly trendiness of topics in the web sphere. This work collects the hourly top topic lists of Alexa during August 1st to August 31st, 2009. Totally, there are 898 unique trending topics, composing the web topic set K_h^A. This dataset includes information

about the topics' ranks in each hour as well.

2 Temporal analysis

This section investigates how the trending topics in microblog sphere behave in the web at first. Later, the section proceeds to analyze the reverse interrelation by studying the Alexa dataset compared to the collected Twitter dataset to examine how the trends of trending topics in the web look like in Twitter.

In summary, it is found that the trending topics are similar within the two worlds where at least 70% of the Twitter trending topics are likely to be also trending in the web and 72% of the web trending topics have been (or will be) also trending in the Twitter world. The results also suggest that although the trendiness in Twitter seems to be synchronous with the one in Google on daily granularity basis, most of the trends of these topics are actually driven by Twitter population in advance, and then spread in the web on a finer granularity (such as on an hourly basis). The notable observed difference is that trendiness in Twitter is highly unstable. It is also found that almost all Twitter trending topics exhibit a very low rank stability, which is opposed to the high stability observed for the web trending topics.

2.1 How do Twitter topics behave in the web?

As the topics extracted from tweets are used to collect their trends in Google, an analysis can be made on how accurately topics' trends in Twitter can approximate their trends in Google.

2.1.1 Trends similarity in Twitter and Google

The similarity between trends in Twitter and Google using are examined using Kullback-Leibler divergence (also called relative entropy), which is a measure of the difference between two probabilities X and Y[12]. The K-L divergence of Y from X, $D_{KL}(X \parallel Y)$, is the expected number of extra bits required to code samples from X when using a code based on Y, rather than using a code based on X, i.e. the information lost when Y is used to approximate X. Typically, the K-L divergence of Y from X is defined as follows:

$$D_{KL}(X \parallel Y) = \sum_i X(i) \log\left(\frac{X(i)}{Y(i)}\right) \quad (2)$$

The smaller the value is, the closer the two distributions are. In this paper, X and Y are related to the Twitter trends and Google trends of topic c, respectively. $X(i)$ (resp. $Y(i)$) is the ratio of trending index volume of c at time i in Twitter (resp. Google) to the total trending index volume of c observed in Twitter (resp. Google).

For each topic that has trends in both Twitter and Google, the K-L divergence of the topic trends in two spheres is observed. The K-L divergence of trends for randomly selected topic pairs from two spheres are also compared. This random selection is used as null hypothesis. Fig. 1 shows the cumulative distribution function (CDF) of K-L divergences. A notable difference between the two K-L divergence distributions for both single-word topics and multi-words topics can be observed. For example, more than 60% of topic pairs have a K-L divergence less than 1 for the same single-word topics, while this value is only 43% for random selection.

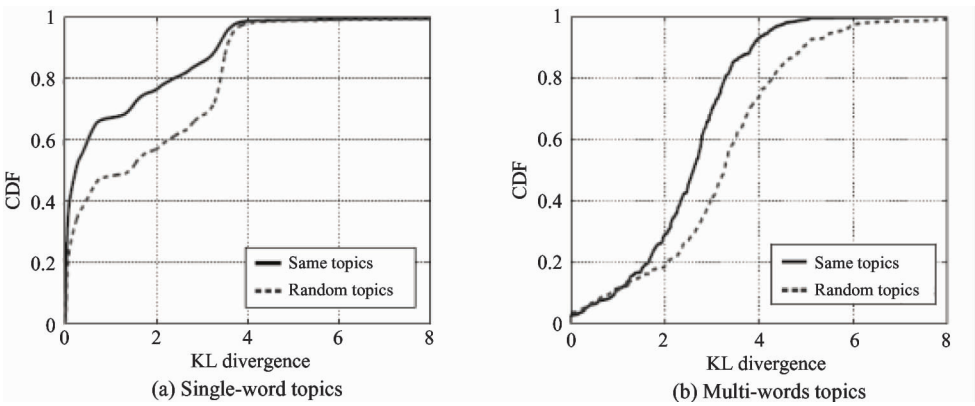

Fig. 1 Kullback-Leibler divergence in two spheres

2.1.2 Trending time analysis

Trending days for each topic (including single-word topics and multi-words topics) $c \in K_d^T$ are examined then. The number of trending days is defined as the number of days the topics have been tagged as trending (either in Twitter or in Google). Fig. 2 shows the distribution of the number of trending days for single-word topics and multi-words topics in Twitter and Google. About 10% of single-word/multi-words topics in Twitter have not been trending (i.e. with 0 trending

days). This is to be expected because only Twitter topics that represent a daily set of the most relevant and popular words used in tweets are considered. It can also be observed that about 20% of topics (either single-word or multi-words topics) in Google have not been trended. Recall that Twitter topics are used to crawl Google Trends service. This observation indicates that 20% of these Twitter topics have never been trended in Google.

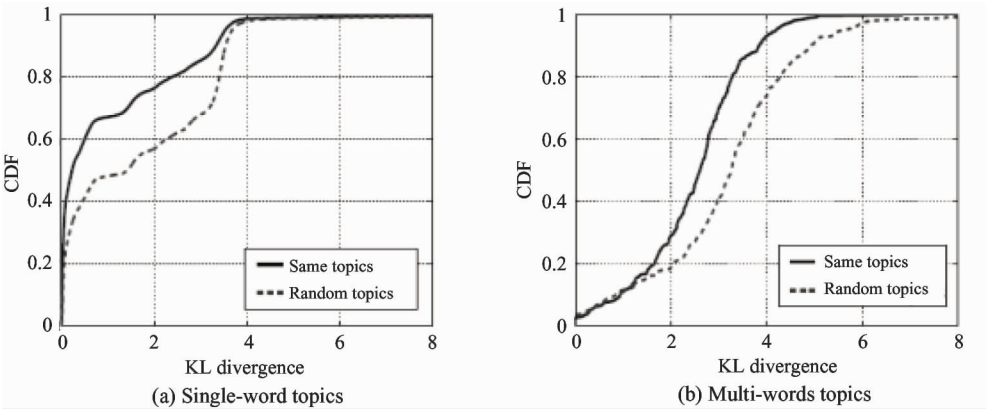

Fig. 2 Distribution of number of trending days for trends in Twitter and in Google

It is interested that compared with Google, topics in Twitter have a shorter trending time. For example, about 20% of the single-word topics are trending in Twitter for more than 3 days, while this proportion is 40% in Google and 20% of topics are even trending in Google for more than 6 days. This observation suggests that trendiness of topics in Twitter is much more volatile than in Google.

2.1.3 Highly positive trends analysis

This part examines for the topics $c \in K_d^T$, the number of highly positive trends they experience. Recall that a highly positive trend is one with the increasing trending index volume larger than a threshold α at a particular time. This typically captures a timely and particularly high interest in a specific topic. Here α varies with 3 values: 50^{th} percentile, 75^{th} percentile and 90^{th} percentile of positive trending index volumes.

Fig. 3 plots the distribution of the number of highly positive trends for topics in Twitter and Google. Depending on the value of α, the proportion of Twitter topics that do not exhibit any highly positive trend varies between 10% and 50% for single-word topics and between 10% and 70% for multi-words topics. Google shows a slightly larger number of highly positive trends than Twitter. For example, there are 30% of the single-word topics and 20% multi-words topics hitting more than 2 highly positive trends in Google with α = 75%, while this percentage in Twitter is about 20% for single-word topics and 10% for multi-words topics. The observation indicates that trending topics have a more stable impact on Google compared with in Twitter.

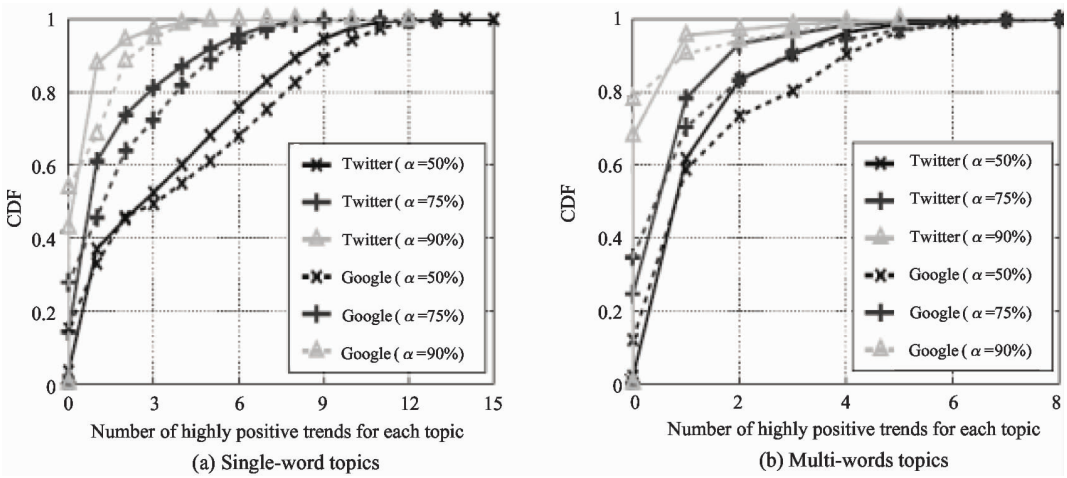

Fig. 3 Distribution of number of highly positive trends for trends in Twitter and in Google

The trending days and highly positive trends in Twitter and Google are compared further by checking that whether a similarity exists between them in Table 1, where the likelihood is computed as the probability that if a topic trending (resp. has highly positive trends) in Twitter is also trending (resp. has highly positive trends) in Google based on the crawled dataset.

Table 1 Comparison of trendiness likelihood in Twitter and in Google for all extracted topics

Metric		Similarity
Trending		65.51%
Highly positive trends	α:50%	69.58%
	α:75%	51.88%
	α:90%	33.90%

The likelihood that a Twitter trending topic is also trending in Google is 65%, and the likelihood for a Twitter topic that exhibits a highly positive trend with $\alpha = 50\%$ in Twitter to similarly show a highly positive trend in Google is 70%. However, when picking a Twitter topic experienced a very highly positive trend ($\alpha = 90\%$), there is only 30% of chances for that topic to experience the same highly positive trend in Google. While this lower number potentially stems from the high-selection of such topics in Twitter, it also suggests that Twitter trendiness is potentially more sensitive than Google. Given the different nature of usages of the two services, this is a reasonable explanation as Twitter users would potentially be more reactive to other users interests and topics.

2.1.4 Time offset analysis

The above results call for a deeper investigation of the time effect so that researchers can understand whether observed trends in one sphere can find their genesis in the other one. For this, the time offset is introduced to represent the difference between the trending times (resp. highly positive times) in Twitter and in Google for trending topics (resp. topics with highly positive trends). In this study, the time offset is defined as, based on a specific feature of trends (trending or highly positive trends), the difference between the first day this feature is observed in Twitter and the first day it is observed in Google. A positive value indicates that the feature happens first in Google and otherwise, it happens first in Twitter.

Fig. 4 and Fig. 5 depict the time offsets between trending and highly positive days for single-word topics respectively. It can be found that most of time offsets assemble around 0 where the proportion of time offsets in $[-1,1]$ interval is much larger than other intervals. In particular, more than half of the time offsets between trending days (resp. highly positive days) in Twitter and Google are in $[-1,1]$ interval, indicating that at most a one-day interval separates the trends in these two spheres. The similar results are observed on the multi-words topics and are omitted due to space limitation. These results show that the trendiness in Twitter is likely to be synchronous with the one in Google on daily granularity.

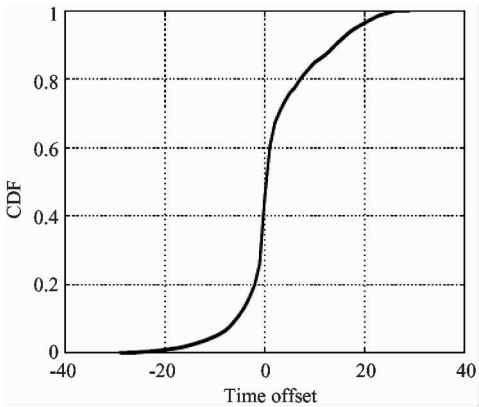

Fig. 4 Time offset between the trending days of single-word topics

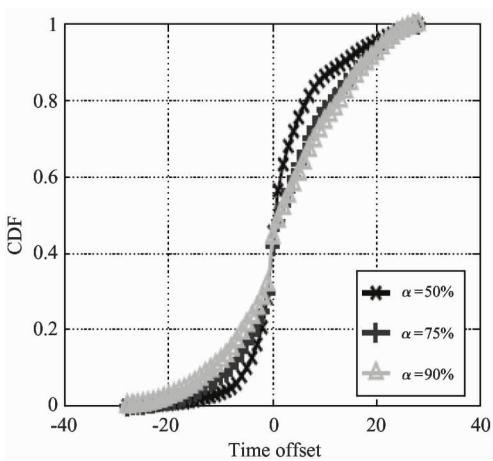

Fig. 5 Time offset between the highly positive days of single-word topics

2.2 How do Web topics behave in Twitter?

This section looks at the topics extracted from the web sphere, and analyzes their trendiness features in the microblog environment. As mentioned earlier in Section 1.2.4, the Google Trends service unfortunately does not provide enough information about the trending topics in web. As an alternative, a set of topics extracted from Alexa (trending topics) ranked lists, K_h^A, are used. This composes the set of trending web topics. This section focuses on the variation of the "trendiness rank" of topics both in Alexa (K_h^A) and in Twitter

(K_h^T) and is also able to conduct the analysis on a finer granularity, i. e. hourly as opposed to daily.

2.2.1 Trends Similarity in Alexa and Twitter

Similar to the analysis of Twitter topics in web, the K-L divergences between Alexa trends and Twitter trends are calculated at first. Recall that in Sec. 1, it is defined that if topic c is in the top trending list of Alexa at time i, then the trending index volume at i is 100; otherwise, it is 0.

As depicted in Fig. 6, if the topic pairs in Alexa and Twitter are randomly selected, the K-L divergences between the two trends are distinctly larger than the ones of same topics, which means the Twitter trends can also be related to the corresponding web trends.

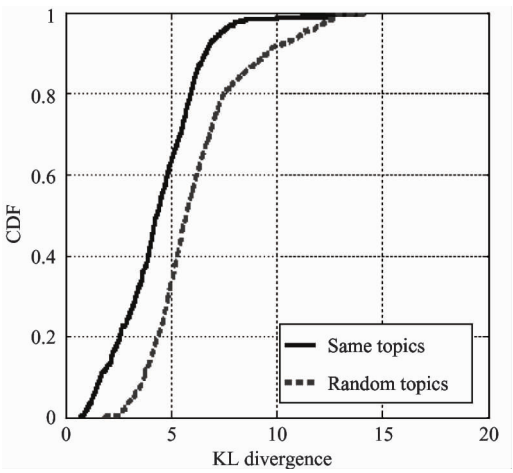

Fig. 6 Kullback-Leibler divergence between Alexa trends and Twitter trends

2.2.2 Trending time analysis

Fig. 7 shows the trending times (on hourly granularity) of 898 topics $c \in K_h^A$ in Alexa and in Twitter respectively, where the trending hours of topic c in Alexa are considered as the hours when c appears in the top trending list and the trending hours in Twitter are estimated using the method described in Section 1. There are two notable observations. First, only 28% of Alexa topics have not been trended in Twitter, which is another evidence that trending topics are similar within the two worlds. Second, topics are likely to be trended for a longer time in Alexa than in Twitter. For example, 16% of topics trending in Twitter for more than 10 hours while the corresponding number in Alexa is about 30%. This observation further confirms the volatility of trendiness in Twitter again.

2.2.3 Time offset analysis

The time offsets (in hour) between trending times of the same topics in Alexa and in Twitter are depicted in Fig. 8, where the positive value indicates that the trending feature happens first in Twitter and otherwise, it happens first in Alexa. Opposed to the results in Fig. 4, the distribution of time offsets in Fig. 8 is skewed towards the positive part, e. g. there are more than 65% time offsets are larger than 0 in Fig. 8. It can be concluded that although the trendiness in Twitter seems to be synchronous with the one in Google on daily granularity, most of trends of these topics are actually driven by Twitter population in advance, and then spread in web on a finer granularity (such as hourly granularity). This result is also in accordance with the reports in Ref. [13].

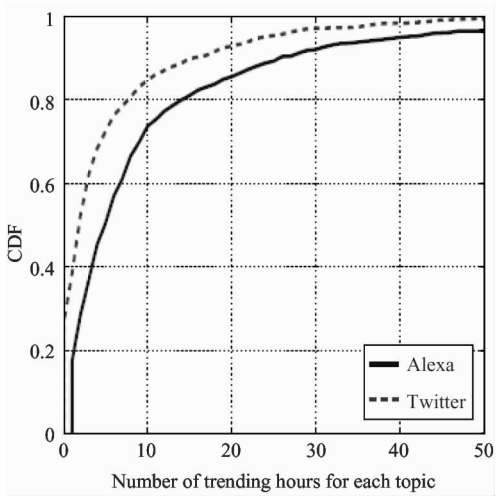

Fig. 7 Distribution of number of trending hours for Alexa trends and Twitter trends

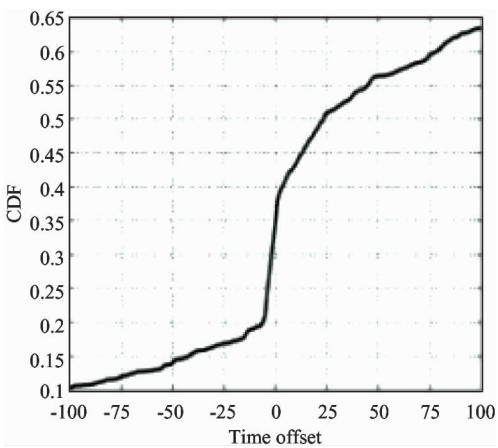

Fig. 8 Time offset between trending hours in Alexa and in Twitter

2.2.4 Rank stability analysis

The rank stability coefficient[14] of trends in Twitter and Alexa is calculated further in order to examine the volatility of trendiness within the two worlds. Given a time frame t, the rank stability coefficient for the top N trending topics in the i^{th} ($i > 1$) bin is defined as

$$R_N(i) = \frac{|S_N(i) \cap S_N(i-1)|}{N} \quad (3)$$

where $S_N(i)$ is the set of top N trending topics during the i^{th} time frame. The rank stability coefficient has values within $[0, 1]$, where 1 indicates no change and 0 means that all the topics in the list have changed.

Fig. 9 depicts the CDF of the rank stability coefficient of the top 20 (i.e. $N=20$) trending topics based on the topics extracted from Alexa (i.e. K_h^A) and the topics extracted from Twitter (i.e. K_h^T) on hourly granularity during the period of August 2009. A notable difference of rank stability coefficient in Twitter and Alexa can be observed. In particular, while there is a limited number of cases in Twitter experiencing a stability coefficients more than 0.5, as many as 90% of the cases in Alexa are more than 0.5. About half of the cases in Twitter have a 0 coefficient, indicating that all the trending topics have changed within one hour. The observations show the "ephemeral" trendiness in Twitter and much more stable web interests.

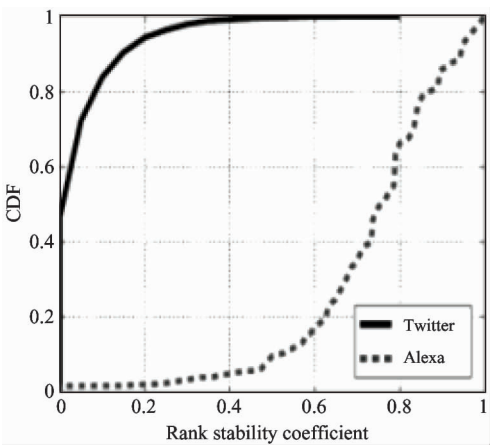

Fig. 9 Rank stability between Alexa top 20 trending topics and Twitter top 20 trending topics (hourly)

3 Spatial analysis

The interaction of information spreading in microblogs and web interests is not only reflected in time but also in the spatial dimension. It has been observed in Ref. [15] and Ref. [16] that both the topic's "original" location and the location of the receivers strongly affect the diffusion patterns of the information. This section analyzes the spatial/geographical dimension of the interaction between microblog trends and web interests.

In summary, it is found that large majority of trending topics appear concurrently in not more than 2 countries in both two spheres, which is a strong evidence of the existence of locality of interest in the trendiness of microblogs and web. Besides, it is also observed that more than 60% of the locality of interest of individual topics exhibit similar patterns in Twitter and in Google.

3.1 Locality of interest

The concept of locality of interest is introduced to characterize the geographic characteristics of trending topics. Five countries, US, UK, CA, FR and AU, are chosen to study whether or not topic c is trending in a specific location. The fewer number of different regions a topic is trending in, the more significant the locality of interest will be. To analyze the locality of interest, the trending topics provided by Twitter from Sept. 1st, 2012 to Oct. 31st, 2012 (dataset H) in these 5 countries are used, and the statistics provided by Google Trends for the same topics within the same period are considered.

Fig. 10 shows the trending topics overlap in the 5 different countries both in Twitter and Google. It can be observed that the Twitter's trending topics have a more notable geographical concentration effect compared to Google. About 80% of Twitter trending topics appear in only one country while this proportion in Google is only 47.5%. In both Twitter and Google, the majority of topics get trending in not more than 2 countries (95.6% in Twitter and 65.0% in Google). This indicates clearly that trendiness both in Twitter and in Google is geography-dependent.

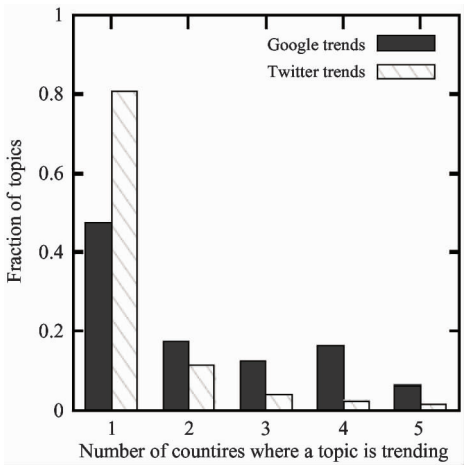

Fig. 10 The overlap of the trending topics in 5 different countries

3.2 Similarity of locality of interest

After confirming the existence of locality of interest in Twitter and Google, similarity of the two spheres in terms of such locality is checked further. To this end, the notion of interest vector is used. The interest vector of topic c is composed of 5 elements in order, $L_{US}(c)$, $L_{UK}(c)$, $L_{CA}(c)$, $L_{FR}(c)$ and $L_{AU}(c)$, each

of which is binary and 1 represents topic c trending in this country and otherwise the value is 0.

Google Trends provides the top 10 trending countries for each topic, so the appearance in the top list can be used to define the interest vector of Google. That said, $L_r(c)$ in Google is 1 if r is in the Google top country list of c; otherwise, it is 0. As to Twitter, whether a topic is in the top trending topic list for each country is considered. $L_r(c)$ in Twitter is 1 if c is in the Twitter top trending topic list of country r; otherwise, it is 0.

For each topic $c \in H$, there are two interest vectors: $\vec{L}^G(c)$ for Google and $\vec{L}^T(c)$ for Twitter. The Jaccard similarity index of these two vectors is computed for each topic to measure the similarity of Google and Twitter in terms of locality of interest. The Jaccard index is a statistic used for comparing the similarity and diversity of binary vectors. For two binary vectors $\vec{A}$ and $\vec{B}$, the Jaccard coefficient $J(\vec{A},\vec{B})$ is defined as:

$$J(\vec{A},\vec{B}) = \frac{\vec{A}\cdot\vec{B}}{|\vec{A}|^2 + |\vec{B}|^2 - \vec{A}\cdot\vec{B}} \tag{4}$$

where $\vec{A}\cdot\vec{B} = \sum_i A_i B_i = \sum_i (A_i \wedge B_i)$ and $|\vec{A}|^2 = \sum_i A_i^2 = \sum_i A_i$. For any pair of vectors $\vec{A}$ and $\vec{B}$, $0 \leqslant J(\vec{A},\vec{B}) \leqslant 1$. The closer this coefficient is to 1, the more similar the two vectors are. Fig. 11 presents the Jaccard similarity coefficient for individual topics. It can be observed that more than 60.5% of the topics exhibit a similarity value larger than 0.60 (i.e. at least 4 elements are the same between the two vectors), which suggests that locality of interest of individual topics exhibit similar pattern in Twitter and Google. In other words, trending topics have similar geographic trends in both Twitter and Google.

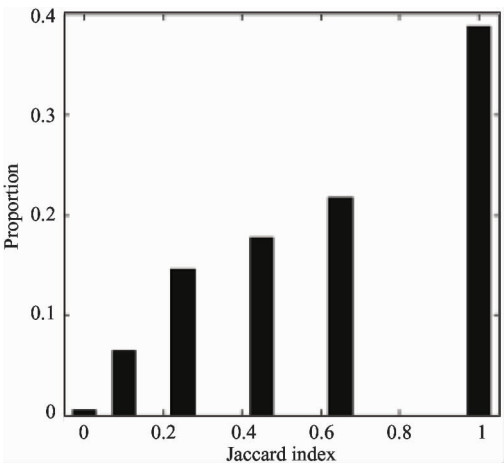

Fig. 11 Distribution of Jaccard index between interest vectors in Twitter and in Google

4 Application

The previous sections found that individual topics in the Twitter sphere and the web sphere share similar trending patterns from both temporal and spatial aspects. Nevertheless, the trendiness in Twitter can be leading for a few hours and is highly unstable compared to the web. The observations suggest the possibility of inferring trending topics from Twitter for the web sphere, which are traditionally provided by search portals like Google.

In fact, the estimation of trends of queries on search engines (such as Google, Bing etc.) is a crucial task in Search Engine Marketing (SEM) analysis. In a typical SEM scenario, advertisers publish their advertisements with the assistant of search engines. In the creation of their advertisements, advertisers choose a keyword or a sequence of keywords (i.e. topics in the context of this paper) relevant to their business, called "ad keywords", which will trigger the display of their advertisements in the returned search page of these ad keywords. As such, discovering the ad keywords searched frequently in search engine at a time (i.e. trending topics in web) is meaningful to capture high impressions and clicks of online advertisements[17-19]. This section shows that trending topics in Twitter could be used to discover superior Google ad keywords.

To this end, the top 10 trending topics of Twitter in US are sampled for every five minutes during two periods: from October 26th to November 2nd of 2013 and from February 2nd to February 8th of 2014. This results in a trending topic dataset T consisting of 1,175 unique trending topics. Twitter is also crawled to get the tweets from US during the same time periods of T using Twitter's streaming API. This results in 105,946 tweets randomly sampled by the Twitter API. Based on these tweets, 1,000 words are randomly chosen and are considered as a non-trending topics dataset N. This dataset is used as a reference for the comparison scenario.

For these 2,175 topics obtained from Twitter, the Google AdWords, which provides a "Keyword Planner" tool for helping users evaluate their ad keywords, is queried. The input of the tool is the chosen keyword and the output is the estimation of the number of impressions and clicks brought by this keyword based on the previous week statistics[20]. By querying this tool, the number of daily impressions and the number of daily clicks of each topic in US for the 10 following days after the topic is sampled from Twitter are obtained.

Fig. 12 shows the CCDF (complementary CDF) of the average estimated number of impressions and clicks returned by "Keyword Planner" for trending topics and non-trending topics in Twitter during the considered 10 days. A significant gap between the distribution functions in terms of both impressions and clicks can be observed. There are a high volume of impressions/clicks for the trending topics. For example, 2% of the trending topics have more than 200,000 estimated impressions while none of the non-trending ones can reach this volume. The results confirm that trendiness in Twitter can be used to infer adwords with high impressions and clicks in SEM.

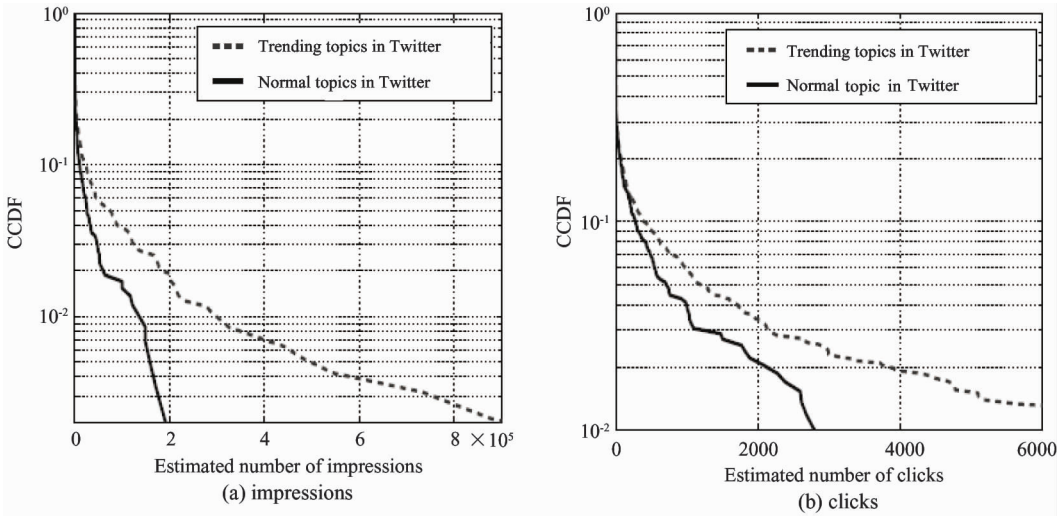

Fig. 12 The distribution of average estimated number of impressions/clicks on Google AdWords for trending and non-trending topics in Twitter during the 10 days

Notably, although "Keyword Planner" provides, based on previous week statistics, official estimates for impressions and clicks on Google AdWords platform, obtaining an up-to-date information about these values is challenging for advertisers simply because of the one week blackout period of "Keyword Planner". However, with the monitoring of Twitter, it can be shown here that advertisers can figure out the current market "status" of Google AdWords on a fine granularity (hours) basis.

5 Related work

Some studies focused on the temporal analysis, i. e. the co-occurrence in close time interval of popularity growth of resources and the diffusion of information in online social media. Sadikov, et al. in Ref. [21] used the features from online blogs and comments to predict the corresponding movies sales. Authors in Ref. [3] studied the correlation between the popularity of videos on a User Generated Content website and the spread of the video URLS by tweets. In Ref. [22], Teevan, et al. compared "simultaneous" search queries over microblogs platforms and on search engines. Kairam, et al. in Ref. [13] found that search and social media activity tend to follow similar temporal patterns. Giummole, et al. found that social trends fired by Twitter may lead to web hot trends derived from Google[23].

Other studies targeted the spatial dimension, i. e. the relationship between the location where a message is published and the scope of its diffusion. Brodersen, et al. found that social sharing generally widens the geographic reach of a video content[15]. Tsou, et al. in Ref. [16] introduced a new research framework for analyzing the spatial distribution of web pages and social media (Twitter) messages.

Among the previous work, the most similar one is the work in Ref. [13], where the researchers have found that social media activity around trending events on Twitter tends to lead query activity on search engines by 4 or 5 hours. However, the work of this paper has shown that the trendiness in Twitter can not only precede for a few hours but also highly unstable compared to the one in web, which indicates that the trending topics could be used as promising adwords in SEM and besides, and the data collected from Google AdWords is used to validate the conjecture. To the best of our knowledge, this study provides the first discussion about the usage scenarios in SEM based on such comparison.

6 Conclusion

This paper has compared the trending topics in Twitter and web (i.e. Google and Alexa) by considering both the temporal and spatial perspectives. It is found that the trending topics in Twitter and search in web tend to follow similar temporal patterns and that the trendiness in Twitter can precede by a few hours. However, trendiness is highly unstable in Twitter where top trending lists change more frequently. Besides, there is a geographical concentration effect of interest in both spheres. The trending "localities" are similar in the two spheres as well. Finally, the paper shows that these observations can be used for a "smart" predictive choice of adwords in SEM.

The ongoing work is to design a predictive statistical model. The latter should take into account the social graph structure and the multi-dimension of the topics to proactively react to prior observations from one of the spheres to accurately predict the future in the other-one.

Reference

[1] Rodrigues T, Benevenuto F, Cha M, et al. On word-of-mouth based discovery of the web. In: Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference, Berlin, Germany, 2011. 381-396

[2] Antoniades D, Polakis I, Kontaxis G, et al. Web: The web of short urls. In: Proceedings of the ACM International Conference on World Wide Web, Hyderabad, India, 2011. 715-724

[3] Wang Z, Sun L, Wu C, et al. Guiding internet-scale video service deployment using microblog-based prediction. In: Proceedings of the IEEE International Conference on Computer Communications, Orlando, USA, 2012. 2901-2905

[4] Alexa. Alexa website. https://www.alexa.com: Alexa, 2015

[5] Google. Helper. https://support.google.com/trends: Google, 2015

[6] Twitter. Twitter helper. https://support.twitter.com: Twitter, 2015

[7] Google. Stopword. http://code.google.com/p/stopwords: Google, 2015

[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *The Journal of machine Learning research*, 2003, 3: 993-1022

[9] Yang J, Leskovec J. Patterns of temporal variation in online media. In: Proceedings of the ACM International conference on Web Search and Data Mining, Hong Kong, China, 2011. 177-186

[10] Google. Googletrends. http://www.google.com/trends: Google, 2015

[11] Zipf G K. Human Behaviour and the Principle of Least-Effort. Cambridge: Addison-Wesley Publishers, 1949

[12] Kullback S, Leibler R A. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951, 22(22): 79-86

[13] Kairam S R, Morris M R, Teevan J, et al. Towards supporting search over trending events with social media. In: Proceedings of the AAAI International Conference on Web and Social Media, Cambridge, USA, 2013. 1-9

[14] Lempel R, Moran S. Rank-stability and rank-similarity of linkbased web ranking algorithms in authority-connected graphs. *Information Retrieval*, 2005, 8(2): 245-264

[15] Brodersen A, Scellato S, Wattenhofer M. Youtube around the world: geographic popularity of videos. In: Proceedings of the ACM International Conference on World Wide Web, Lyon, France, 2012. 241-250

[16] Tsou M H, Yang J A, Lusher D, et al. Mapping social activities and concepts with social media (twitter) and web search engines (yahoo and bing): a case study in 2012 US presidential election. *Cartography and Geographic Information Science*, 2013, 40(4): 337-348

[17] Joshi A, Motwani R. Keyword generation for search engine advertising. In: Proceedings of the IEEE International Conference on Data Mining, Hong Kong, China, 2006. 490-496

[18] Chen Y, Xue G R, Yu Y. Advertising keyword suggestion based on concept hierarchy. In: Proceedings of the ACM International Conference on Web Search and Data Mining, New York, USA, 2008. 251-260

[19] Thomaidou S, Vazirgiannis M. Multiword keyword recommendation system for online advertising. In: Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining, Taiwan, China, 2011. 423-427

[20] Google. Keywordplanner. http://adwords.google.com/ko/Keywor-dPlanner: Google, 2015

[21] Sadikov E, Parameswaran A, Venetis P. Blogs as predictors of movie success. In: Proceedings of the AAAI International Conference on Web and Social Media, San Jose, USA, 2009. 1-9

[22] Teevan J, Ramage D, Morris M R. Twittersearch: A comparison of microblog search and web search. In: Proceedings of the AAAI International Conference on Web and Social Media, Barcelona, Spain, 2011. 35-44

[23] Giummole F, Orlando S, Tolomei G. A study on microblog and search engine user behaviors: How twitter trending topics help predict google hot queries. *HUMAN*, 2013, 2(3): 195

Wang Dong, born in 1987. He is a Ph. D candidate at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). His research area includes network measurement and analysis of OSNs.