Journal of Web Engineering, Vol. 13, No.5&6 (2014) 405-429
© Rinton Press

FINDING NEWS-TOPIC ORIENTED INFLUENTIAL TWITTER USERS BASED ON TOPIC RELATED HASHTAG COMMUNITY DETECTION

FENG XIAO, TOMOYA NORO, TAKEHIRO TOKUDA

Department of Computer Science, Tokyo Institute of Technology

Meguro, Tokyo 152-8552, Japan
{xiao, noro, tokuda}@tt.cs.titech.ac.jp

Received July 2, 2013
Revised May 20, 2014

Recently, more and more users would like to collect and provide information about news topics in Twitter, which is one of the most popular microblogging services. Virtual communities defined by hashtags in Twitter are created for exchanging information about the news topic. Finding influential Twitter users in these communities related to a news topic would help us understand why some opinions are popular, and get valuable and reliable information for the news topic. In this paper, we propose a new approach to detect news-topic-related user communities defined by hashtags based on characteristic co-occurrence word detection. We also propose RetweetRank and MentionRank to find two types of influential Twitter users from these news-topic-related communities based on user's retweet and mention activities. Experimental results show that our characteristic co-occurrence word detection methods could detect words which are highly relevant to the news topic. RetweetRank could find influential Twitter users whose tweets about the news topic are valuable and more likely to interest others. MentionRank could find influential Twitter users who have high authority on the news topic. Our methods also outperform other related methods in evaluations.

Key words: Social Network Analysis, Twitter, hashtag, PageRank, characteristic co-occurrence word
Communicated by: G.-J. Houben & L. Olsina

1 Introduction

Microblogging [20] is a new way for users to collect and provide information on the Web. One of the most famous microblogging services is Twitter, which attracts more than 200 million active users creating over 400 million messages, called tweets, everyday [5]. Most of these tweets often concern topics of headline news or persistent news [14], making Twitter an important data source for news.

Functions provided by Twitter help users easily share news with each other. A user could follow other users who have the same interest. He can repost interesting tweets, called retweet, when he would like to share them with his followers. Mention and reply[a], prefixing user name with @ symbol, are used for purposes such as direct communication with others, or referring to users who are relevant. Hashtags (the # symbol prefixed to a short character string) are widely used by users to categorize and joint tweets together for a certain topic. Virtual user communities defined by hashtags in tweets are

[a] "Reply" is taken as a special case of "mention". We refer to both of them as "mention" in this paper.

formed to exchange information with others in these communities [10] [32]. We refer to a group of users who use the same hashtag in their tweets as "hashtag community", and these users are members of the hashtag community.

Although Twitter is a good platform to share news, a recent survey conducted on ordinary users reveals that 92% of users choose to go directly to news websites and 85% of users would do a specific keyword search for their interested news topics. Getting news from social media like Twitter is supplemental for news consumption [29]. However, it is difficult for a user to find those supplemental contents for their interested news topics. Suppose, for example, a user who is interested in U.S. presidential election sends a query "Obama" to a news search engine to get news articles containing the query word. However, it is difficult for him to get tweets related to this news topic by sending the same query to Twitter. That's because the tweet has the length limitation of 140 characters. Tweets related to this news topic do not necessarily contain the query word. Another option is to get tweets by following others. However, it is still difficult to find users worth following for the news topic. Tweets posted by some users about the news topic are valuable and more likely to interest other users while tweets posted by others, even related to the news topic, are unattractive and more likely to be ignored. Following those users whose tweets are paid close attention to by others would help to get attractive contents and understand why some opinions are popular for the news topic. However, measuring the value of tweets posted by a Twitter user is a non-trivial task. Also, Twitter users could post tweets freely while it is hard to know whether contents of these tweets are reliable or not. Following those Twitter users who have high authority on the news topic (e.g. a political journalist reporting the presidential election) would help us get more reliable information. However, for ordinary users, especially those users who are novices for the news topic, professionals of the news topic might be unknown to them.

The purpose of our research is to help ordinary users find influential Twitter users worth following for a news topic after they search for the news topic by a keyword (we refer to the keyword as target word in this paper). Two new methods are proposed to find two types of influential Twitter users for the news topic in which ordinary users are interested. One type of influential Twitter user often posts tweets containing valuable information for the news topic. Their tweets are more likely to interest others (e.g. get retweeted). We refer to this type of Twitter user as content-based influential Twitter user. Following this type of Twitter user could get tweets which are very attractive and help us understand why some opinions for the news topic are popular. The other type of influential Twitter user has high authority on the news topic so that other Twitter users would be more likely to communicate (e.g. mention) with him. We refer to this type of Twitter user as authority-based influential Twitter user. Following this type of Twitter user could get tweets which are reliable because these users have high authority on the news topic. For the news topic of U.S. presidential election, one good example of content-based influential Twitter user is "@PatDollard", a famous filmmaker in the U.S. who often shares his opinions about the election and attracts many others, especially Republican supporters. One good example of authority-based influential Twitter user is "@andersoncooper", the Twitter account of a famous American journalist. His tweets for the presidential election are reliable due to his special social position. To find these two types of influential Twitter users for a news topic, tweets related to the news topic are needed. However, due to the length limitation of tweets, ordinary Information Retrieval methods are no longer effective in collecting tweets related to the news topic. In this paper, we collect tweets related to the news topic by detecting hashtags which are relevant to the

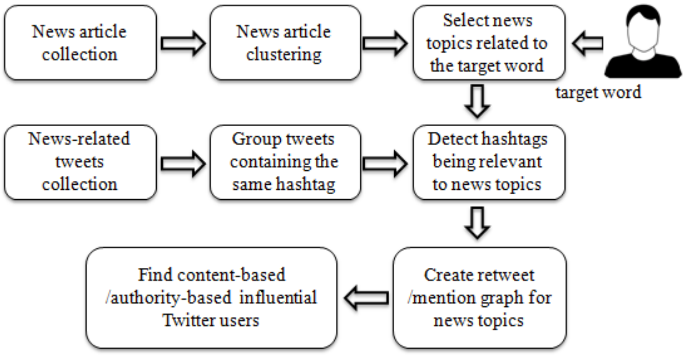

News article collection
News article clustering
Select news topics related to the target word
target word
News-related tweets collection
Group tweets containing the same hashtag
Detect hashtags being relevant to news topics
Create retweet /mention graph for news topics
Find content-based /authority-based influential Twitter users

Figure 1 System structure

news topic. A hashtag which is often used to share contents about the news topic are considered to be relevant to the news topic. We refer to this hashtag as a news-topic-related hashtag, and the hashtag community defined by this hashtag as news-topic-related hashtag community. Tweets containing news-topic-related hashtags are taken as tweets related to the news topic. Two types of influential Twitter users could be found from users who posted these tweets.

The whole system structure of our approach is shown in Figure 1. We first collect news articles and tweets related to news published in a certain period of time (for example: one day) concurrently. Then news articles are clustered into topics. Tweets containing the same hashtag are grouped together. After a user provides the target word for searching, news topics related to the target word are selected. For each of these news topics, news-topic-related hashtags are detected based on two newly proposed characteristic co-occurrence word detection methods. For users in hashtag communities defined by these news-topic-related hashtags, two user activity graphs are created. One is retweet graph created based on user's retweet activities. The other is mention graph created based on user's mention activities. Content-based and authority-based influential Twitter users could be found from these two user activity graphs by using newly proposed RetweetRank and MentionRank methods.

Contributions of this paper are two-fold. One contribution is that we propose two new methods to detect characteristic co-occurrence words with the target word/hashtags from news articles/tweets. They are basic components for detecting news-topic-related hashtag communities. Characteristic co-occurrence words are words which provide important information for news topics and hashtags. Our methods are not only query-and-topic dependent to detect/weight words in news articles, but also effective in detecting/weighting words in tweets. The other contribution is that, since a user retweets a tweet of others because he is interested in tweet contents and he mentions other users because mentioned users are relevant to the topic he is talking about [6], we treat user's retweet and mention activities differently and propose RetweetRank and MentionRank to find content-based and authority-based influential Twitter users from news-topic-related hashtag communities. Experimental results show that our methods to detect characteristic co-occurrence words outperform other methods using TF-IDF, Jaccard coefficient and Log Likelihood Ratio. To find content-based and authority-based influential Twitter users in news-topic-related hashtag communities, RetweetRank and MentionRank outperform other methods using tweet number, in-degree, and PageRank.

Although our study focuses on members of hashtag communities in this research, we believe that influential Twitter users found by our methods are quite helpful. After invented in 2007, hashtags become more and more widely used in Twitter to form conversations about a topic among Twitter users globally without following each other. Other conversations formed by functions like reply are restricted by follow relations. A user is less likely to reply to other users who are not followed by him because their tweets will not appear in his Twitter timeline.

The rest of this paper is organized as follows. We discuss related work in Section 2. In Section 3, we describe our approach to detect news-topic-related hashtag communities based on characteristic co-occurrence word detection. RetweetRank and MentionRank are proposed in Section 4 to find content-based and authority-based influential Twitter users from hashtag communities detected in Section 3. Experimental results and evaluations are described in Section 5. Finally, we make the conclusion with directions for future research in Section 6.

2 Related Work

Our research presented here relates to two research fields about social networking services. One is hashtag retrieval/recommendation, and the other is finding influential Twitter users.

2.1 Hashtag Retrieval/Recommendation

Hashtag retrieval/recommendation has been studied recently. Lehmann et al. [17] classified hashtags in four classes based on their activity profiles over time. However, our purpose is to detect the relevance between hashtags and a news topic based on the content they relate to. Popularity variation of hashtag in its activity profile could not reflect this relevance. Weng et al. [27] proposed methods to model the interestingness of hashtags by studying how hashtags are used within and across communities, but they do not correlate hashtags with user’s interested topics. Efron [7] proposed a new approach to retrieve relevant hashtags after a keyword is given. However, one keyword may relate to more than one topic. All hashtags related to different topics might be mixed together. Zangerle et al. [34] recommended hashtags for a newly input tweet by calculating similarity between the new tweet and old tweets based on TF-IDF. Hashtags which are frequently used in old tweets being similar to the new one get recommended. Mazzia and Juett [19] also proposed use of Bayesian model to recommend hashtags based on newly input tweet. Kywe et al. [15] considered not only newly input tweet, but also similarities between users to recommend hashtags. Experimental results showed that their method yields better performance than other methods only considering tweet contents. Although these researches seem to be reasonable, there are still some problems. Firstly, similarities between tweets in researches above simply rely on common words in tweets. However, due to the length limitation, two tweets may refer to the same topic while both of them have no common word. Secondly, TF-IDF is no longer a good choice for short text [26] like tweets. Due to the huge number of tweets, the IDF part would dominate the final score, assigning too large a score to the word which appears scarcely (e.g. misspelling). Lastly, purposes of researches above are different from ours. They try to recommend hashtags for a newly input tweet. However, our purpose is to detect hashtags which are relevant to a news topic searched by the target word.

2.2 Finding Influential Twitter Users

Finding influential users in social networking services has been focused by researchers recently. Many methods have been proposed for measuring user's influence in Twitter. These methods could be mainly classified into two classes based on user's relation type.

One class of these methods measures user's influence based on user's follow relation. The most intuitive way to measure a user's influence is to count the number of followers the user has. It is based on the assumption that more followers a user has, more impact he could make on other users. Another similar method uses the ratio of the number of user's followers to the number of users he follows. However, follow relation is not a good indicator for user's influence. A Twitter user could follow a large number of other users, wishing them to follow back for courtesy. Moreover, only considering follow relation ignores user's interaction with other users. The user whose tweets are ignored by most of his followers has less influence on the others even if many users follow him.

The other class of method measures user's influence based on his interactive activities like mention, reply, and retweet. Cha et al. [6] analyzed three influence measures based on user's retweet, mention, and follow relations independently. They found that the number of user's followers reveals little about his influence. Retweet represents the value of tweets, and mention represents user's name value. Leavitt et al. [16] defined Twitter user's influence as the potential of a user's action to initiate a further action by other users. They measured user's influence by the ratio of attentions he received (being mentioned, replied, and retweeted) to the number of tweets he posted. Anger and Kittl [1] proposed another influence measure based on the ratio of user's tweets which are retweeted and the ratio of user's followers retweeting his tweets or mentioning him. Hajian and White [9] proposed Influence Rank, a variant of PageRank, to quantify user's influence in Twitter. The difference between Influence Rank and PageRank is the way in which the teleportation vector is defined. The teleportation vector in Influence Rank is calculated based on a combination of user's follow, like, comment, and retweet activities. Romero et al. [23] proposed another Influence-Passivity algorithm to measure the influence and passivity of Twitter users based on retweet activity. They proposed methods to define two transition matrices to measure the amount of influence each user accepts/rejects from others. Then HITS algorithm [13] is applied to these two transition matrices to determine the influence of each user (hub score in HITS). Although researches presented above seem to be reasonable, an influential Twitter user in general might not be influential for a specific news topic. Our methods could find those Twitter users who are influential for the news topic in which ordinary users are interested. Also, existing researches do not consider different purposes of user activities. They use only one type of activity (retweet activity in [23]), or take all activities as the same relation type [1, 9, 16]. We take this difference into account and propose methods to find two types of influential Twitter users based on retweet and mention activities respectively.

Finding topic related influential Twitter users has also been explored. Ye and Wu [33], and Bigonha et al. [2] found influential Twitter users for a manually selected topic (Michael Jackson's death and soda brands) based on user's activities like reply, and retweet. However, they ignore the link structure among users. A user should be more influential if he is retweeted/mentioned by other influential Twitter users rather than users with less influence. Noro et al. [21] proposed a new approach to find influential Twitter users related to a query word. However, one query word might correspond to multiple topics. Influential Twitter users for different topics might get mixed together. Weng et al. [28]

found high follow reciprocity among Singapore Twitter users and proposed TwitterRank method to find influential Twitter users for topics based on user's follow relation. They defined a new transition matrix with teleportation vector, taking into account the number of tweets posted and the topical similarity between users. However, results from [6] contradict the high follow reciprocity after analyzing near-complete data from Twitter. Also, the definition of topic in TwitterRank is different from the definition in our methods. The topic from TwitterRank is distilled by Latent Dirichlet Allocation as a distribution over a fixed vocabulary. Our news topic is defined as a group of news articles published in a period of time (for example: one day) reporting about the same recent event in the world. Cano et al. [4] also proposed Topic-Entity PageRank to find influential Twitter users for both topic and entity. Tweets are categorized into predefined topics by OpenCalais[b]. Then a transition matrix is defined for each topic based on retweet activity. PageRank algorithm is applied to this transition matrix to find influential Twitter users for the topic. However, topics from Topic-Entity PageRank are predefined while our news topics are automatically clustered from news articles. Also, an influential Twitter user for one topic from OpenCalais, for example Politics, might not be always influential for all political issues in the world.

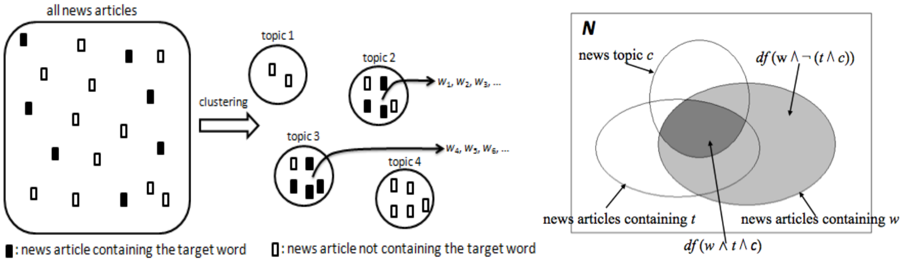
In this paper, we are not trying to recommend hashtags to users. Instead, we measure the relevance between hashtags and news topics searched by the target word. Hashtags which are highly relevant to the news topic are detected. Then content-based and authority-based influential Twitter users for this news topic could be found from these hashtag communities based on user's retweet and mention activities.

3 News-Topic-Related Hashtag Community Detection

In this section, we will describe how to find hashtag communities being relevant to a news topic searched by the target word. After a user provides the target word, news topics related to the target word are selected. A news topic is a group of news articles published for a period of time (for example: one day) reporting about the same recent event in the world. To detect hashtag communities being relevant to a news topic, relevance between the news topic and hashtags should be measured. Traditional way to measure the relevance is to calculate the cosine similarity between the news topic and the hashtag, which are both represented by vectors under the Vector Space Model [24]. Each dimension of the vector corresponds to a separate term in news articles or tweets, and the term weight is calculated by the TF-IDF [12]. Although TF-IDF works well in many tasks like the Information Retrieval, it is no longer the best choice in our approach. Firstly, TF-IDF is a query-independent method. No matter what the target word is, all term weights do not change. Secondly, TF-IDF is a topic-independent method. Terms which often appear in news articles of a topic should be weighted higher while TF-IDF could not reflect this idea. Lastly, TF-IDF is no longer effective in weighting terms in tweets because the length of tweet is extremely short [26].

In order to solve these problems, we propose two new methods to detect characteristic co-occurrence words with the target word/hashtags from news articles/tweets. Characteristic co-occurrence words are words which provide important information for news topics or hashtags. Our methods to detect characteristic co-occurrence words are based on two assumptions:

[b] OpenCalais, http://www.opencalais.com/


all news articles
clustering
topic 1
topic 2
topic 3
topic 4
$w_1, w_2, w_3, \ldots$
$w_4, w_5, w_6, \ldots$
▮: news article containing the target word
▯: news article not containing the target word
N
news topic c
$df(w \wedge \neg(t \wedge c))$
news articles containing t
news articles containing w
$df(w \wedge t \wedge c)$

(a) Procedure of characteristic co-occurrence word detection

(b) PIOLog method

Figure 2 Characteristic co-occurrence word detection for news topics

- Characteristic co-occurrence word *w* should often co-occur with the target word *t* /hashtag *ht* in news articles/tweets. We take it as the Inside part.
- Characteristic co-occurrence word *w* should not always appear in news articles/tweets without the target word *t* /hashtag *ht*. We take it as the Outside part.

Characteristic co-occurrence words detected from news articles of a news topic or from tweets containing the same hashtag are selected to create the vector for representing the news topic or hashtag. Relevance between the news topic and hashtag could be measured by the cosine similarity.

3.1 Characteristic Co-occurrence Word Detection for News Topics

Based on these two assumptions, words often co-occurring with the target word in news articles while being less likely to appear in news articles without the target word are taken as characteristic co-occurrence words. However, one target word may relate to multiple news topics. Characteristic co-occurrence words detected from different news topics may get mixed together. Also, general words which often co-occur with the target word (e.g. "White House" for "Obama") should be excluded since they co-occur with the target word regardless of the news topic.

To solve these problems, all news articles are clustered into topics. Characteristic co-occurrence words could be detected from each news topic related to the target word. The procedure is shown in Figure 2 (a). Characteristic co-occurrence words are detected from news topics related to the target word respectively without mixing words from different news topics. Additionally, news articles in topics other than the topic we focus on are treated as the "Outside part" regardless of existence of the target word. This will exclude general words which often co-occur with the target word regardless of the news topic if news articles containing the target word are separated into two or more than two news topics.

To reflect our ideas, we propose Probabilistic Inside-Outside Log method (PIOLog) to detect characteristic co-occurrence word *w* with the target word *t* for a news topic *c* as follows:

$$\mathrm{PIOLog}(w,t,c) = \log \frac{(1-s_p)\mathrm{P}(w \mid t \wedge c) + s_p}{(1-s_p)\mathrm{P}(w \mid \neg(t \wedge c)) + s_p} \tag{1}$$

$$\mathrm{P}(w \mid t \wedge c) = \frac{df\,(w \wedge t \wedge c)}{df\,(t \wedge c)} \tag{2}$$

$$\mathrm{P}(w \mid \neg(t \wedge c)) = \frac{df\,(w \wedge \neg(t \wedge c))}{df\,(\neg(t \wedge c))} = \frac{df\,(w) - df\,(w \wedge t \wedge c)}{N - df\,(t \wedge c)} \tag{3}$$

where *df* (*w*) is the number of news articles containing the word *w*. *df* (*w*∧*t*∧*c*) is the number of news articles containing both *w* and *t* in the news topic *c*. *N* is the total number of news articles (Figure 2(b)). s_p is a smoothing parameter ranging from 0 to 1. *df(t∧c)* is taken as the Inside part. Words which often co-occur with *t* in news articles of the news topic *c* would get a large score in Equation 2, reflecting our idea of the first assumption. *N – df(t∧c)* is taken as the Outside part. Words which are less likely to appear in news articles without *t* or unrelated to *c* will get a small score in Equation 3, reflecting our idea of the second assumption. Words whose PIOLog scores calculated by Equation 1 are large would be more likely to be characteristic co-occurrence words with the target word *t* for the news topic *c*.

3.2 Characteristic Co-occurrence Word Detection for Hashtags

In order to find news-topic-related hashtags, one intuitive way is to retrieve tweets related to a news topic and select hashtags which are frequently used in these tweets. However, tweet length is limited within 140 characters, which means there is not enough information in a single tweet to decide whether the tweet relates to a news topic or not. Also, traditional way to weight terms like TF-IDF is no longer effective for short text, which has been pointed out in [26].

To solve these problems, we concatenate on tweets containing the same hashtag and a hashtag vector is created based on concatenated contents of these tweets. Each dimension of the hashtag vector corresponds to a characteristic co-occurrence word with the hashtag from tweets. Here, words which often co-occur with the hashtag in tweets while they are less likely to be used in tweets with other hashtags are taken as characteristic co-occurrence words with the hashtag. To reflect this idea, we extend our PIOLog method and propose Probabilistic Inside-Outside Log method for Hashtag (PIOLogH) to detect characteristic co-occurrence word *w* with hashtag *ht* as follows:

$$\mathrm{PIOLogH}(w, ht) = \log \frac{(1 - s_p)\mathrm{P}(w \mid ht) + s_p}{(1 - s_p)\mathrm{P}(w \mid \neg ht) + s_p} \tag{4}$$

$$\mathrm{P}(w \mid ht) = \frac{\#\mathrm{Tweet}(w \wedge ht)}{\#\mathrm{Tweet}(ht)} \tag{5}$$

$$\mathrm{P}(w \mid \neg ht) = \frac{\#\mathrm{Tweet}(w) - \#\mathrm{Tweet}(w \wedge ht)}{TN - \#\mathrm{Tweet}(ht)} \tag{6}$$

where #Tweet(*w* ∧ *ht*) gives the number of original tweets containing both *w* and *ht*. Original tweets are tweets posted by users excluding retweeted tweets. Since official retweet function does not allow users to revise tweet contents, hashtags in retweeted tweets could not reflect original ideas of hashtag usage of users. So these retweeted tweets are excluded here. *TN* is the total number of original tweets containing hashtags in our dataset. #Tweet(*ht*) is taken as the Inside part. Words which often co-occur with *ht* in tweets will get a large score in Equation 5, reflecting our first assumption. *TN* - #Tweet(*ht*) is taken as the Outside part. Words which are less likely to appear in tweets with other hashtags will get a small score in Equation 6, reflecting our second assumption. Words whose PIOLogH scores calculated in Equation 4 are large would be more likely to be characteristic co-occurrence words with the hashtag.

3.3 Detecting News-Topic-Related Hashtag Community

To measure the relevance between the news topic and hashtags, we create the news topic vector and the hashtag vector. Top-*n* characteristic co-occurrence words with their PIOLog or PIOLogH scores which are larger than the rest are selected. Cosine similarity between the news topic vector and the hashtag vector is calculated to measure the relevance between news topic *c* related to the target word *t* and the hashtag *ht* as follows:

$$\text{HTRelevance}(ht, c, t) = \cos(\overrightarrow{hv}(ht), \overrightarrow{nv}(c, t)) \quad (7)$$

$$\overrightarrow{hv}(ht) = < hw_1, \cdots, hw_n > \text{and } \overrightarrow{nv}(c, t) = < nw_1, \cdots, nw_n > \quad (8)$$

where HTRelevance(*ht, c, t*) gives the relevance score between *ht* and *c* calculated by the cosine similarity between hashtag vector $\overrightarrow{hv}(ht)$ and news topic vector $\overrightarrow{nv}(c, t)$.

When a user uses a hashtag in his tweets to share contents about a news topic, some words related to the news topic would be used in his tweet. Also, when the hashtag is widely used by other users for that news topic, more words related to that news topic would be used in their tweets. Both of them will result in a high cosine similarity between $\overrightarrow{hv}(ht)$ and $\overrightarrow{nv}(c, t)$. Hashtags which have large relevance scores with *c* are news-topic-related hashtags, and hashtag communities defined by these hashtags are news-topic-related hashtag communities.

4 Finding News-Topic Oriented Influential Twitter Users

In this section, we introduce our new methods to find content-based and authority-based influential Twitter users for news topics related to the target word. These two types of influential Twitter users could be found based on retweet and mention activities. Our methods to find influential Twitter users for a news topic are based on two assumptions:

- More users a user gets retweeted/mentioned from, more influence the user would have.
- A user has high influence if other users who retweet/mention him are influential.

Based on these two assumptions, we extend the PageRank method and propose RetweetRank and MentionRank to measure the content-based and authority-based influences of Twitter users.

4.1 RetweetRank: Finding Content-based Influential Twitter Users

Hashtags which have high relevance scores with a news topic *c* detected in Section 3 are taken as news-topic-related hashtag set, which is denoted by H_c. Hashtag communities defined by these hashtags contain Twitter users who used any hashtag $ht \in H_c$ in their tweets. A directed graph $G_{RT}(V_{RT}, E_{RT})$ is created among these Twitter users based on their retweet activities. We refer to this retweet graph as G_{RT} in this paper. V_{RT} is the vertex set. It contains Twitter users who retweeted tweets or got retweeted by others in hashtag communities defined by hashtags in H_c. E_{RT} is the edge set. If user u_a retweets a tweet containing any $ht \in H_c$ from user u_b, there is an edge between them, directing from u_a to u_b.

RetweetRank uses a model of random surfer in G_{RT}. The random surfer follows edges in E_{RT} to visit the next Twitter user based on retweet activities of the former one. The random surfer would also jump to any Twitter user with certain probability even if there is no edge between them. Unlike PageRank whose random surfer visits the next vertex uniformly, the random surfer of RetweetRank visits the next vertex based on user's retweet activities and hashtag preference for *c*. In RetweetRank, the random surfer would be more likely to visit the next user whose tweets containing news-topic-related hashtags are often retweeted by the former user, and these two users often use common hashtags for the news topic. We refer to the transition matrix of RetweetRank for a news topic *c* as A_{RR}. The transition probability from u_a to u_b is calculated as follows:

$$\mathrm{A}_{\mathrm{RR}}(u_a,u_b)=\frac{\#\mathrm{RT}(u_a,u_b\mid\forall ht\in \mathrm{H}_c)}{\sum_{u_i\in \mathrm{V}_{\mathrm{RT}}}\#\mathrm{RT}(u_a,u_i\mid\forall ht\in \mathrm{H}_c)}\times \mathrm{HSim}(u_a,u_b) \tag{9}$$

$$\mathrm{HSim}(u_a,u_b)=\vec{\mathrm{H}}(u_a)\cdot\vec{\mathrm{H}}(u_b) \tag{10}$$

$$\vec{\mathrm{H}}(u_i)=<\#\mathrm{Tweet}(u_i,ht_1),\cdots,\#\mathrm{Tweet}(u_i,ht_m)>\text{and}\ u_i\in \mathrm{V}_{\mathrm{RT}},\mathrm{H}_c=\{ht_1,\cdots,ht_m\} \tag{11}$$

where $\#\mathrm{RT}(u_a, u_b \mid \forall ht \in \mathrm{H}_c)$ gives the number of tweets u_a retweeted from u_b containing any $ht \in \mathrm{H}_c$. $\mathrm{HSim}(u_a, u_b)$ gives the similarity of hashtag preference between u_a and u_b. $\vec{\mathrm{H}}(u_i)$ is the hashtag preference vector of user u_i. Each dimension of this vector is $\#\mathrm{Tweet}(u_i, ht_j)$, which is the normalized number of original tweets containing ht_j posted by u_i. Similar hashtag preference of two users indicates similar interest of them for *c*. Transition probability between two users in retweet graph is large if one user often retweets tweets containing any $ht \in \mathrm{H}_c$ from the other user, and two users have similar hashtag preference for *c*. Finally A_{RR} is made to be stochastic so that sum of entries in each row equals one.

4.2 MentionRank: Finding Authority-based Influential Twitter Users

Similar assumptions are also applied to find authority-based influential Twitter users. A directed graph $G_{MN}(V_{MN}, E_{MN})$ is created among users in news-topic-related hashtag communities based on user's mention activities. We refer to this mention graph as G_{MN} in this paper. V_{MN} is the vertex set. It contains Twitter users who mentioned others or got mentioned by others in hashtag communities defined by hashtags in H_c. E_{MN} is the edge set. If user u_a mentions user u_b in his tweets containing any $ht \in H_c$, there is an edge between them, directing from u_a to u_b.

MentionRank also uses the random surfer model in G_{MN}. The random surfer visits the next user following edges between users. He would also jump to any user in G_{MN} with certain probability without any edge. The random surfer of MentionRank would visit the next user based on mention activities of the former one. He would be more likely to visit the next user often mentioned by the former one than other users mentioned in fewer times. We refer to the transition matrix of MentionRank for a news topic *c* as A_{MR}. The transition probability from u_a to u_b is defined as follows:

$$\mathrm{A}_{\mathrm{MR}}(u_a,u_b)=\frac{\#\mathrm{MN}(u_a,u_b\mid\forall ht\in \mathrm{H}_c)}{\sum_{u_i\in \mathrm{V}_{\mathrm{MN}}}\#\mathrm{MN}(u_a,u_i\mid\forall ht\in \mathrm{H}_c)} \tag{12}$$

where $\#\mathrm{MN}(u_a, u_b|\forall ht \in \mathrm{H}_c)$ gives the number of original tweets containing any $ht \in \mathrm{H}_c$ and mentioning u_b by u_a. Transition probability from u_a to u_b would be large if u_a often mentions u_b in tweets containing $ht \in \mathrm{H}_c$ while u_a is less likely to mention others. We do not consider hashtag preference of users here because authority-based influential Twitter users are often mentioned due to their name value for the topic, not hashtags they use. A_{MR} is also made to be stochastic so that sum of entries in each row equals one.

4.3 Topic-Related Teleportation Vector

To guarantee that the probability distribution in PageRank would converge to a steady state, a teleportation vector is introduced to make the transition matrix be irreducible and aperiodic [3]. Also, in our retweet and mention graphs, some pairs of Twitter users retweet/mention each other in a looping manner without retweeting/mentioning others. These user pairs accumulate influence scores without propagating their influence outside. To solve these problems, a teleportation vector is introduced to allow the random surfer to jump to vertices without an edge in a certain probability instead of travelling along edges of the graph.

The random surfer in PageRank jumps to any vertex of the graph uniformly while it does not consider the relevance between vertices and the topic. Here we introduce a topic-related teleportation vector for all vertices (users) in retweet and mention graphs. It would make the random surfer be more likely to jump to the next user who is highly relevant to the news topic, making the final results more topic-sensitive. A user is highly relevant to a news topic *c* if he often posts tweets about *c* in news-topic-related hashtag communities, and hashtags used by the user are highly relevant to *c*. Since users who are highly relevant to the news topic are interested in this topic, those tweets they retweeted are more valuable than tweets randomly retweeted by others, and Twitter users mentioned by them are more likely to have high authority on that news topic. This could help to find out content-based and authority-based influential Twitter users more effectively and completely. The teleportation vector $\overrightarrow{TV_c}$ for a news topic *c* is defined as follows:

$$\overrightarrow{TV_c} =< \text{UserRelevance}(u_1, c), \cdots, \text{UserRelevance}(u_n, c) > \text{and } u_i \in \text{V}_{\text{RT}} \text{ or } u_i \in \text{V}_{\text{MN}} \quad (13)$$

$$\text{UserRelevance}(u_i, c) = \left[\sum\nolimits_{ht_j \in \text{H}_c} \frac{\#\,\text{Tweet}(u_i, ht_j)}{\#\,\text{Tweet}(u_i, \text{H}_c)} \times \text{HTRelevance}(ht_j, c, t) \right] \times \log[\#\ \text{Tweet}(u_i, \text{H}_c) + 1] \quad (14)$$

where each dimension of $\overrightarrow{TV_c}$ corresponds to a user in retweet/mention graph. The value of each dimension UserRelevance(u_i, *c*) measures user's relevance to *c*. HTRelevance(ht_j, *c*, *t*) is the relevance score between hashtag ht_j and *c* calculated in Section 3.3. #Tweet(u_i, ht_j) gives the number of original tweets containing $ht_j \in \text{H}_c$ posted by u_i. #Tweet (u_i, H_c) gives the total number of original tweets posted by u_i containing any $ht \in \text{H}_c$. A user who is interested in the news topic *c* and often shares contents in news-topic-related hashtag communities would get a large relevance score in his dimension. The random surfer would be more likely to jump to him. Finally the teleportation vector is normalized to make the sum of dimension values to be one.

4.4 Ranking Content-based and Authority-based Influential Twitter Users

With transition matrices for retweet and mention graphs and topic-related teleportation vector defined, RetweetRank and MentionRank can be calculated by using power iteration method as follows:

$$\overrightarrow{RR_c} = d(\text{A}_{\text{RR}})^T \cdot \overrightarrow{RR_c} + (1-d)\overrightarrow{TV_c} \text{ until} \| \overrightarrow{RR_c}(\text{k}) - \overrightarrow{RR_c}(k-1) \| < \varepsilon \quad (15)$$

$$\overrightarrow{MR_c} = d(\mathrm{A}_{\mathrm{MR}})^T \cdot \overrightarrow{MR_c} + (1-d)\overrightarrow{TV_c} \text{ until} \| \overrightarrow{MR_c}(k) - \overrightarrow{MR_c}(k-1) \| < \varepsilon \tag{16}$$

where $(\mathrm{A}_{\mathrm{RR}})^T$ is the transposed transition matrix for retweet graph. $(\mathrm{A}_{\mathrm{MR}})^T$ is defined in the same way. $\overrightarrow{TV_c}$ is the teleportation vector for the news topic c calculated in Equation 13. d is the damping factor. Computations for RetweetRank vector $\overrightarrow{RR_c}$ and MentionRank vector $\overrightarrow{MR_c}$ are done iteratively. These two vectors would converge to stationary probability vectors until 1-norm of the residual vector is less than a predefined threshold ε. Finally, value in each dimension of $\overrightarrow{RR_c}$ or $\overrightarrow{MR_c}$ indicates a user's content-based influence score in retweet graph, or his authority-based influence score in mention graph.

5 Experiments and Evaluation

5.1 Dataset Description

To evaluate the effectiveness of our methods, news article dataset and news-related tweet dataset are prepared for the experiment by crawling news articles and news-related tweets concurrently. News article collection is shown in Figure 3(a). We collect news articles written in English from 96 news sites in 21 countries/regions every day. However, it is not easy to collect news-related tweets because it is difficult to decide whether one tweet relates to a news topic or not due to the length limitation of tweet. Our solution is to manually select 54 active Twitter accounts of news providers and collect tweets containing mentioned/tagged screen name of these accounts (e.g. @CNN, #CNN) by using Twitter Streaming API[c]. Then hashtags excluding tagged screen name of these news providers and used in more than 10 collected tweets are selected. These hashtags are used as queries to search for more tweets by using Twitter Search API[d]. At last we combine tweets collected from these two APIs to create the news-related tweet dataset. The whole procedure is shown in Figure 3(b). We select news articles and news-related tweets collected on October 11th, 2012 for our experiment. There are 6,868 news articles and 1,496,420 news-related tweets collected on this day. Although there might be some other tweets related to news topics, collecting those tweets by using ordinary Information Retrieval technologies is no longer effective.

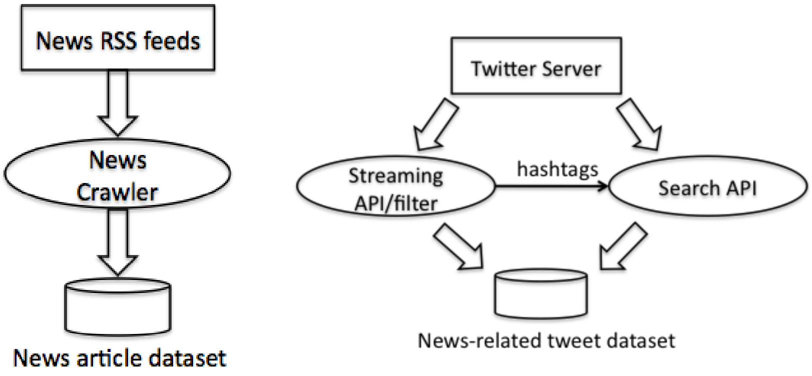
5.2 Experimental Setup

Collected news articles are parsed by TreeTagger [25] and Stanford Named Entity Recognizer (SNER) [8]. All nouns, proper nouns, foreign words, verbs, and adjectives are picked up to represent each news article under the Vector Space Model [24] as a term vector. Then all news articles are clustered into topics by using Hierarchical Agglomerative Clustering (HAC) [18] with a predefined similarity threshold of th_{news}.

After user provides the target word, a news cluster is taken as a news topic related to the target word if more than half of its news articles contains the target word. Then PIOLog method in Section 3.1 is used to detect characteristic co-occurrence words with the target word for the news topic. Top-*n* words with their PIOLog scores which are larger than the rest are selected to create the news topic vector. We refer to this news topic vector as $\overrightarrow{nv}_{\mathrm{PIOLog}}(c,t)$. Words which are highly relevant to the

[c] https://dev.twitter.com/docs/api/1.1/post/statuses/filter

[d] https://dev.twitter.com/docs/using-search



(a). News article collection

(b). News-related tweet collection

Figure 3 Experimental dataset collections

news topic are selected in the news topic vector with large term weights.

News-related tweets are also preprocessed. Firstly, non-English tweets from Twitter Search API and tweets having no hashtag or written by non-native English users whose language setting in their Twitter profile is not set to "en" are excluded. Also tweets from those 54 Twitter accounts of news providers are excluded. Secondly, original tweets are selected to create the hashtag vector. Retweeted tweets[e] are not used to detect news-topic-related hashtags because Twitter users are not allowed to revise the tweet contents when they use the official "Retweet" function. Hashtags in these retweeted tweets could not reflect original ideas of Twitter users. But retweeted tweets would be used in RetweetRank later. Thirdly, we create hashtag communities by grouping Twitter users with their original tweets containing the same hashtag. After excluding tagged screen name of news providers and hashtags used in less than 50 tweets, 2,772 hashtags with corresponding hashtag communities are selected. On average, there are 363.32 tweets posted by 235.46 users for each hashtag. Tweets containing the same hashtag are parsed into terms by using TreeTagger and SNER while mentions, URLs, and hashtags are excluded. We use PIOLogH method proposed in Section 3.2 to detect/weight characteristic co-occurrence words for each hashtag. Top-*n* words with their PIOLogH scores which are larger than the rest are used to create the hashtag vector. We refer to it as $\overrightarrow{hv}_{\text{PIOLogH}}(ht)$. For a news topic related to the target word, hashtags whose relevance scores calculated by Equation 7 are larger than a predefined threshold of th_{ht} are taken as news-topic-related hashtags, and hashtag communities defined by these hashtags are taken as news-topic-related hashtag communities.

We also do several preliminary experiments to estimate parameters (th_{news}, n, s_p, and th_{ht}). To estimate the value of th_{news} in news article clustering, 6,322 news articles are collected on October 9th, 2012 for the preliminary experiment. We select five news topics and estimate the value of th_{news} based on average precision of these topics. We observe that average precision of these five news topics have a sharp decrease when th_{news} is below 0.26. So the value of th_{news} is set to 0.26. To estimate the value of n which is the number of dimensions of news topic vector and hashtag vector, we also select these five news topics and set n to 100, 200, 300, 400, 500, and 600. The value of n should be small to decrease computation while there should be little variation of detected hashtags when the value of n increases.

[e] Here, we take tweets which begin with "RT @username:" as retweeted tweets.

Table 1 Summary of news topics related to the target word

ID	Summary
Target Word = “Obama”	
O_1	U.S. presidential election
Target Word = “Syria”	
S_1	Syria crisis and conflictions
Target Word = “game”	
G_1	American Major League Baseball news
G_2	News for England football match

So for each value of n ($n \geq 200$), we select top-100 detected hashtag list whose relevance scores calculated by Equation 7 are large and compare with the hashtag list detected when n is set to (n – 100). Variation of detected hashtag list is measured by the number of common hashtags and the Kendall’s Tau coefficient. As we observe that when the value of n is above 400, the number of common hashtags and the value of Kendall’s Tau coefficient reach high values and do not vary greatly. So the vector dimension (n) is set to 400. The value of s_p and th_{ht} are estimated from our previous researches [30, 31]. Smoothing parameter s_p is used to make the denominator of Equation 1 and 4 be nonzero. It should be small enough while detected words should be stable, not varying greatly for different values of s_p with small interval. In [30], we range the value of s_p from 0.01 to 0.1 with the interval of 0.01 and measure the variation of detected words along with the increase of s_p. We also use the number of common words and the Kendall’s Tau coefficient to measure variation. We observe that detected words with their ranking orders reach a stable status after s_p is above 0.05. We also observe the same situation in characteristic co-occurrence word detection from tweets. So the value of s_p is set to 0.05. Datasets from [31] are used to estimate th_{ht}, which is the threshold for detecting news-topic-related hashtags. The value of th_{ht} ranges from 0.1 to 0.2 with the interval of 0.01. Mean Average Precision (MAP) of detected hashtags reaches a large value without great variation when th_{ht} is above 0.17. So th_{ht} is set to 0.17 here.

We choose “Obama”, “Syria”, and “game” as target words. News topics related to each target word are selected. Summaries of these topics are described in Table 1. There are four news topics selected. They are denoted by O_1, S_1, G_1, G_2, including two of them (G_1 and G_2) relate to the target word of “game”. After detecting hashtags which are relevant to each of these news topics, retweet graph G_{RT} and mention graph G_{MN} are created for each news topic among Twitter users in these news-topic-related hashtag communities. Retweeted tweets and tweets containing mentions of other users are selected to create edges in G_{RT} and G_{MN}. Topic-related teleportation vector for each news topic are also created. The damping factor d is set to 0.85, the same value as in PageRank. The threshold ε for stopping iteration is set to 0.00005 since it does not affect results too much.

Table 2 shows detailed information about G_{RT} and G_{MN} for each news topic. $|c|$ gives the number of news articles clustered into the news topic. $|H_c|$ gives the number of news-topic-related hashtag communities. $|V_{RT}|$ and $|V_{MN}|$ show the number of vertices (users) in retweet and mention graphs respectively. $|E_{RT}|$ and $|E_{MN}|$ show the number of edges in these two graphs. RetweetRank and MentionRank are used in retweet and mention graphs of each news topic respectively. Twitter users whose RetweetRank scores or MentionRank scores are larger than the others are taken as content-

Table 2 Information about retweet and mention graphs for news topic *c*

Topic	$\lvert c \rvert$	$\lvert H_c \rvert$	$\lvert V_{RT} \rvert$	$\lvert E_{RT} \rvert$	$\lvert V_{MN} \rvert$	$\lvert E_{MN} \rvert$
O_1	179	112	3691	6924	4307	6560
S_1	86	20	464	996	439	534
G_1	65	24	176	131	403	321
G_2	33	6	80	63	178	147

based or authority-based influential Twitter users.

5.3 Comparison with Related Methods

In this section, we discuss related methods to detect characteristic co-occurrence words for news topics and hashtags. We also explain existing methods to find content-based and authority-based influential Twitter users.

5.3.1 Comparison for characteristic co-occurrence word detection

We compare newly proposed PIOLog and PIOLogH methods with TF-IDF, Jaccard coefficient and Log Likelihood Ratio (LLR) [18]. Characteristic co-occurrence words should be query and topic dependent. Our methods could reflect this idea while TF-IDF can't. Also, our method is asymmetric while Jaccard coefficient is symmetric. Whether word w_1 is a characteristic co-occurrence word with word w_2 and whether w_2 is a characteristic co-occurrence word with w_1 depend on the news topic, and they should be different in general. Other asymmetric methods like the Log Likelihood Ratio are often used for different purpose. For example, LLR is always used to detect word collocation, which is an expression consisting of two or more words that correspond to some conventional way of saying things (e.g. hot dog). However, our method is used to find two words often co-occurring because they are highly related due to a specific topic, not a conventional way of word using.

To use TF-IDF in our experiment, we calculate the centroid vector of each news topic using TF-IDF after all news articles are clustered into news topics. Top-*n* words whose TF-IDF scores are larger than the rest are selected to create the news-topic vector. We refer to it as $\overrightarrow{nv}_{\text{TF-IDF}}(c,t)$. Since TF-IDF can't be applied directly to tweets containing the same hashtag, we also propose a variant of TF-IDF to weight words in these tweets. We refer to it as Term Frequency-Inverse Hashtag Frequency (TF-IHF). The calculation of TF-IHF is described as follows:

$$\text{TF-IHF}(w,ht) = \text{TF}(w,ht) \times \text{IHF}(w,\text{HT}),\ \text{HT} = \{ht_1,\cdots,ht_i,\cdots\} \tag{17}$$

$$\text{TF}(w,ht) = \frac{n_{w,ht}}{\sum_k n_{k,ht}} \tag{18}$$

$$\text{IHF}(w,\text{HT}) = \log\frac{|\text{HT}|}{|ht_i : \#\text{Tweet}(w,ht_i) != 0| + 1},\ ht_i \in \text{HT} \tag{19}$$

where *w* is the word from tweets containing the hashtag *ht*. HT is the hashtag set containing all hashtags from our news-related tweet set. $n_{w,ht}$ gives the number of times *w* appears in tweets containing *ht*. #Tweet(w,ht_i) gives the number of tweets containing both word *w* and the hashtag ht_i. TF(w,ht) gives a high value to the word *w* often co-occurring with the hashtag *ht*. IHF(w,HT) gives a low value to *w* co-occurring with many other hashtags because this word might be generally more common than other words. TF-IHF value ranges from 0 to log(|HT|/2). Top-*n* words whose TF-IHF scores are larger than the rest are selected to create the hashtag vector. We refer to it as $\overrightarrow{hv}_{\text{TF-IHF}}(ht)$.

Jaccard coefficient is also revised to detect and weight words from news articles and tweets. To detect words for news topics, the Jaccard score of word *w* is calculated as follows:

$$\text{Jaccard}(w,t,c)=\frac{df(w\wedge t\wedge c)}{df(w\vee(t\wedge c))}=\frac{df(w\wedge t\wedge c)}{df(w)+df(t\wedge c)-df(w\wedge t\wedge c)} \tag{20}$$

where *df* ($w\wedge t\wedge c$) gives the number of news articles in the news topic *c* that contain both word *w* and target word *t*. *df* ($w\vee(t\wedge c)$) gives the number of news articles containing *w*, or containing *t* and in *c*. We select top-*n* words whose Jaccard scores are larger than the rest and create the news topic vector. We refer to it as $\overrightarrow{nv}_{\text{Jaccard}}(c,t)$. Jaccard scores of words in tweets are calculated in a similar way. We also select top-*n* words from tweets containing the same hashtag to create the hashtag vector. We refer to it as $\overrightarrow{hv}_{\text{Jaccard}}(ht)$.

LLR is also extended in our experiment. For news topics, the null hypothesis is set as the occurrence of word *w* is independent of the target word *t* and the news topic *c*. Alternative hypothesis is set in opposite as *w* is dependent on *t* and *c*. These two hypotheses are described as below:

$$\text{Null hypothesis}\text{H}_0:\text{P}(w\mid t\wedge c)=p=\text{P}(w\mid\neg(t\wedge c)) \tag{21}$$

$$\text{Alternative hypothesis}\text{H}_1:\text{P}(w\mid t\wedge c)=p_1\neq p_2=\text{P}(w\mid\neg(t\wedge c)) \tag{22}$$

LLR assumes a binomial distribution b(*k; n, p*) for each word in news articles, then the LLR score is calculated as follows:

$$\text{LLR}(w,t,c)=-2\log\frac{\text{L}(\text{H}_0)}{\text{L}(\text{H}_1)} \tag{23}$$

$$\text{L}(\text{H}_0)=\text{b}(df(w\wedge t\wedge c);df(t\wedge c),p)\cdot\text{b}(df(w)-df(w\wedge t\wedge c);N-df(t\wedge c),p) \tag{24}$$

$$\text{L}(\text{H}_1)=\text{b}(df(w\wedge t\wedge c);df(t\wedge c),p_1)\cdot\text{b}(df(w)-df(w\wedge t\wedge c);N-df(t\wedge c),p_2) \tag{25}$$

where L(H_0) and L(H_1) are probabilities of having *df* (*w*), *df* ($t\wedge c$), and *df* ($w\wedge t\wedge c$) observed under hypotheses H_0 and H_1 respectively. Top-*n* words with their LLR scores calculated by Equation 23 are selected to create the news topic vector. We refer to this news topic vector as $\overrightarrow{nv}_{\text{LLR}}(c,t)$. LLR scores of words from tweets could be calculated in a similar way. We also select top-*n* words from tweets containing the same hashtag to create the hashtag vector. We refer to this hashtag vector as $\overrightarrow{hv}_{\text{LLR}}(ht)$.

5.3.2 Comparison for finding content-based and authority-based influential Twitter users

To find two types of influential Twitter users, comparisons with related methods are conducted. Other methods used to find these two types of influential Twitter users are described as follows:

(1). Tweet number. This method measures Twitter user's influence based on the number of tweets posted by the Twitter user containing news-topic-related hashtags. More tweets posted by the user, more influential the user would be.

(2). In-degree. This method measures Twitter user's influence based on the number of times the Twitter user get retweeted/mentioned by others in retweet/mention graph. More times the user get retweeted/mentioned, more influential the user would be.

(3). PageRank. This method measures Twitter user's influence in retweet/mention graph by using PageRank algorithm. However, relevance between Twitter users and news topic is ignored in its teleportation vector. User's retweet/mention preferences are not considered either when calculating transition probabilities. Larger the PageRank score of a user has, more influential the user would be.

For ease of presentation, RetweetRank and MentionRank are denoted by RR and MR. Method using the number of posted tweets is denoted by TN. In-degree is denoted by IND and PageRank is denoted by PR.

5.4 Evaluation

In this section, we show our evaluation results for characteristic co-occurrence word detection. We also compare different methods to find content-based and authority-based influential Twitter users.

5.4.1 Evaluation for characteristic co-occurrence word detection

Since our purpose is to use characteristic co-occurrence words to detect news-topic-related hashtags, we compare the quality of results achieved by the same news-topic-related hashtag detection approach in Section 3.3 when input characteristic co-occurrence words are detected by different methods. The method whose detected words are topic-specific and more relevant to the news topic could help to detect hashtags which are more relevant to the news topic.

For each news topic (O_1, S_1, G_1 and G_2), we set four experiments with different characteristic co-occurrence word detection methods to detect news-topic-related hashtags. These four experiments are described as follows.

- Exp. 1: $\overrightarrow{nv}_{\mathrm{PIOLog}}(c,t)\cdot\overrightarrow{hv}_{\mathrm{PIOLogH}}(ht)$. Words from news topic vector are detected by PIOLog and words from hashtag vector are detected by PIOLogH.
- Exp. 2: $\overrightarrow{nv}_{\mathrm{TF\text{-}IDF}}(c,t)\cdot\overrightarrow{hv}_{\mathrm{TF\text{-}IHF}}(ht)$. Words from news topic vector are detected by TF-IDF and words from hashtag vector are detected by TF-IHF.
- Exp. 3: $\overrightarrow{nv}_{\mathrm{Jaccard}}(c,t)\cdot\overrightarrow{hv}_{\mathrm{Jaccard}}(ht)$. Words from news topic vector and hashtag vector are detected by Jaccard coefficient.
- Exp. 4: $\overrightarrow{nv}_{\mathrm{LLR}}(c,t)\cdot\overrightarrow{hv}_{\mathrm{LLR}}(ht)$. Words from news topic vector and hashtag vector are detected by Log Likelihood Ratio (LLR).

In each experiment, top-*n* words whose scores are larger than the rest are selected to create news topic vector and hashtag vector. Methods which outperform others would rank those topic-specific informative words higher and hashtags detected by these methods should be more relevant to the news topic.

Table 3 News-topic-related hashtags from four experiments for O_1

Exp. 1	Exp. 2	Exp. 3	Exp. 4
romneyryan2012	dateline	tcot	romney
politics	libyagate	teaparty	p2
tcot	activismrocks	tlot	tcot
election2012	debate2012	romneyryan2012	libyagate
mitt2012	2012election	p2	gop
romney	teamfollback	romney2012	romneyryan2012
romney2012	ia	obama2012	mitt
gop	joebiden	nobama	politics
p2	fourmoreyears	ocra	obama
teaparty	therealromney	lnyhbt	obama2012
obama	mostrecent	politics	debate
debate	rr2012	gop	2012election
mittromney	obamaisntworking	mitt2012	etchasketch
election	nobama2012	obama	mitt2012
mitt	etchasketch	twisters	election2012
Highly relevant hashtags (HR)	election2012, mitt2012, mittromney, romneyryan2012, romney, obama2012, debate2012, 2012election, romney2012		
Relevant hashtags (R)	2012election, romney2012, mitt2012, mitt, teaparty, nobama, romneyryan2012, obama2012, politics, election2012, election, mittromney, debate, romney, gop, obama, joebiden, debate2012, fourmoreyears, therealromney, rr2012, obamaisntworking		

To evaluate results of four experiments, we ask two assessors to judge the relevance between detected hashtags and the news topic. To help our assessors better understand the news topic, they could search for any information if they need to make a proper decision. The whole procedure is shown as below.

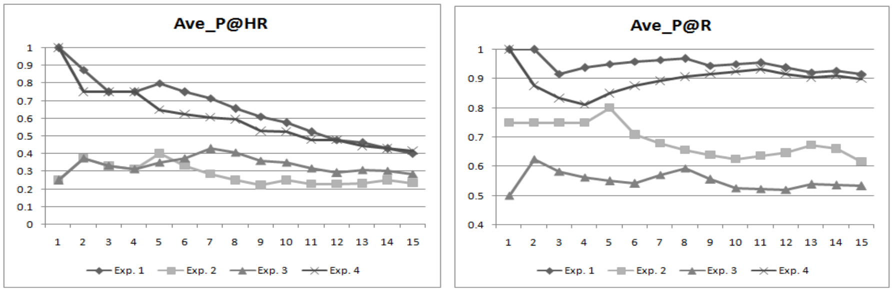
1. Two assessors are asked to read at least ten news articles which are carefully selected for each news topic so that these news articles can cover the main contents of the news topic to help them understand the news topic.

2. Top-15 hashtags with largest similarities detected by each of four experiments are mixed to form a hashtag set for each news topic. Assessors judge the relevance of each hashtag in this set to the news topic on a three-point scale: highly relevant, relevant and irrelevant. They can use any tool (e.g. TagDef[f] or Google) to find definitions of hashtags.

3. For each news topic, hashtags which are judged as highly relevant by two assessors are defined as highly relevant hashtags. We also define relevant hashtags as those which are not judged as irrelevant by any assessor. Notice that highly relevant hashtags are a subset of relevant hashtags.

For example, for the new topic of O_1 about U.S. presidential election (Table 3), “#election2012” is judged as the highly relevant hashtag because tweets containing this hashtag mainly relate to the

[f] TagDef: http://tagdef.com/



(a). Average P@HR curves

(b). Average P@R curves

Figure 4 Average curves of four experiments

presidential election. However, "#politics" which is often used to mark tweets about political issues is judged as relevant hashtag because it is not only for O_1, but also other political topics. Hashtags like "#mostrecent" used for other purposes or news topics are judged as irrelevant hashtags.

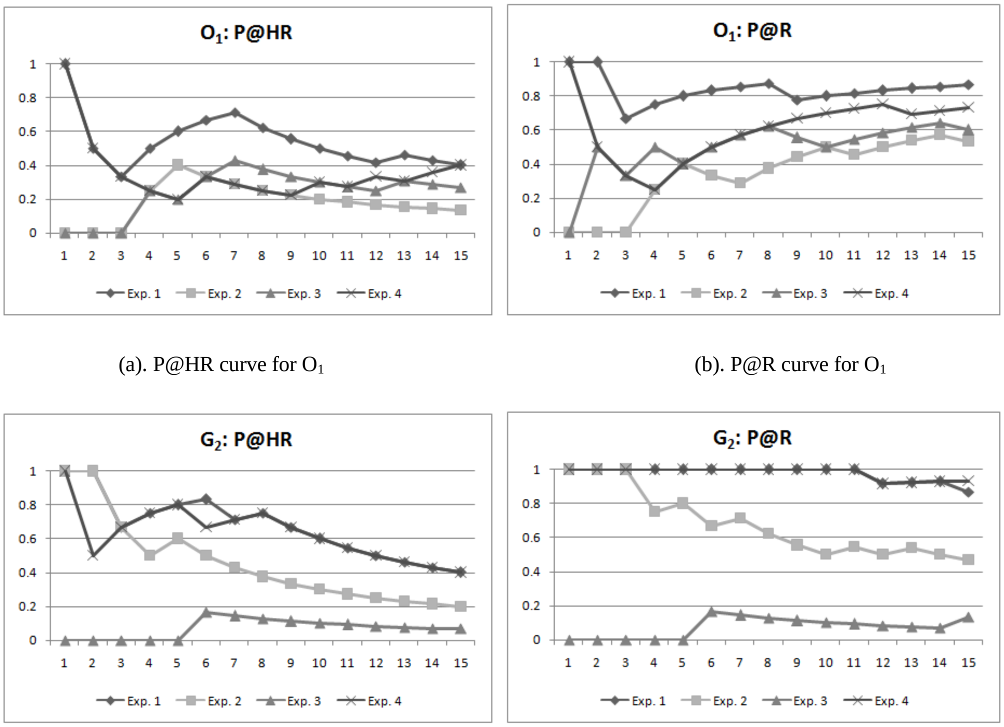
To evaluate performances of four experiments for these news topics, we use precision as the evaluation metric under two-levels:

- **Precision at highly relevance curve** (**P@HR curve**): each point on this curve indicates the fraction of top-*r* detected hashtags that are highly relevant hashtags for the news topic.
- **Prevision at relevance curve** (**P@R curve**): each point on this curve indicates the fraction of top-*r* detected hashtags that are relevant hashtags for the news topic.

We range the value of *r* from 1 to 15. For example, in Table 3, there are two highly relevant hashtags ("romneyryan2012" and "election2012") in top-4 detected hashtags from Exp. 1, so the P@HR when $r = 4$ is 0.5. P@R is calculated in the same way. Experiment whose P@HR and P@R curves locate higher than the others performs the best to detect news-topic-related hashtags. Methods used in this experiment outperform other methods for detecting characteristic co-occurrence words.

For each experiment, we average P@HR and P@R curves of four news topics. Figure 4 (a) and (b) show the average P@HR curve (Ave_P@HR) and average P@R curve (Ave_P@R) for four experiments. As we can observe that curves of Exp. 1, using our newly proposed PIOLog and PIOLogH methods, locate higher than curves of other experiments. This indicates that hashtags detected in Exp. 1 are more relevant to the news topic than hashtags detected in other experiments. PIOLog and PIOLogH methods used in this experiment are more likely to detect characteristic co-occurrence words for news topics and hashtags. Exp. 2 using TF-IDF with its variant as TF-IHF performs not well. As we have pointed out in the start of Section 3, TF-IDF is a query and topic independent method, which is not suitable for detecting characteristic co-occurrence words with the target word for a news topic. Also, TF-IHF does not consider the number of tweets containing both the word and the hashtag, which might bias towards words appearing many times in a few tweets.

As we can also observe, asymmetric methods (LLR, PIOLog and PIOLogH) outperform symmetric method (Jaccard coefficient) in our experiments. As we have pointed out before, when we take two words be w_1 and w_2, whether w_1 is a characteristic co-occurrence word with w_2 and whether w_2 is a characteristic co-occurrence word with w_1 depend on the topic and they should be different in general. Asymmetric method could reflect this idea while symmetric method can't.



(a). P@HR curve for O_1

(b). P@R curve for O_1

(c). P@HR curve for G_2

(d). P@R curve for G_2

Figure 5 P@HR and P@R curves for O_1 and G_2

For asymmetric methods, PIOLog and PIOLogH methods still perform better than LLR though their average curves are close. LLR considers the appearance of word *w* is independent/dependent on the target word *t* or the hashtag *ht* that seems to be similar to our assumptions. However, LLR is still not suitable for detecting characteristic co-occurrence words. LLR is often used to detect word collocation, which is a different purpose compared with ours. Our characteristic co-occurrence word detection is to detect words strongly related to the target word due to a specific news topic, not as a grammar unit constantly. We also manually check results for each news topic and observe that our methods perform better than LLR in large news topics, but they perform very similar in small news topics. Figure 5 (a) – (d) show P@HR and P@R curves for news topic O_1 and G_2. There are 179 news articles in O_1 and our methods (Exp. 1) outperform LLR (Exp. 4) in Figure 5 (a) and (b). However, for the news topic of G_2 containing 33 news articles, their results are very similar in Figure 5 (c) and (d). This is because LLR is more appropriate for sparse data [18]. That is to say, characteristic co-occurrence words co-occurring with the target word in news topics of small size are more likely to be detected by LLR while in large size of news topics, words co-occurring with the target word in less news articles are preferred by LLR. Although being appropriate for sparse data is an advantage of LLR to detect word collocation, it is a big disadvantage to detect characteristic co-occurrence words because characteristic co-occurrence words should co-occur with the target word in more news articles of the news topic. The same situation also happens for hashtags. This feature of LLR contradicts the definition of characteristic co-occurrence word.

Table 4 Content influential score manually assigned for content-based influential Twitter users of news topic *c*

Score	Description
2	The user often posts tweets for *c* while most of them often get retweeted by many users.
1	The user posts many tweets for *c* while only part of them get retweeted. The user's tweets for *c* get retweeted while his tweets for other topics get retweeted more times.
0	The user posts tweets unrelated to *c*. The user's tweets for *c* do not interest others.

Table 5 Authority influential score manually assigned for authority-based influential Twitter users of news topic *c*

Score	Description
2	The user's tweets are highly trustable for *c*. The user is highly relevant to *c* in the real world
1	The user is supported by some users about *c*. The user has high authority on other related topics while he also posts tweets for *c*.
0	The user posts tweets unrelated to *c*. The user's tweets are ignored by most of users.

5.4.2 Evaluation for finding content-based and authority-based influential Twitter users

To evaluate the effectiveness of our newly proposed RR and MR, we apply TN, IND, PR, and RR to the retweet graph of each news topic to find content-based influential Twitter users. We also apply TN, IND, PR, and MR to mention graph of each news topic to find authority-based influential Twitter users.

To evaluate content-based and authority-based influential Twitter users found by different methods for the news topic *c*, we select top-15 Twitter users from each method and ask two assessors to manually assign content influential score or authority influential score for each user on a three-point scale. Definitions of the content influential score and authority influential score are described in Table 4 and Table 5. Table 6 gives examples of top-15 content-based and authority-based influential Twitter users found by RR and MR for news topic O_1. For example, for content-based influential Twitter users, @PatDollard is assigned a content influential score of 2 by both assessors. That's because he is a famous Twitter user who often shares his opinions about the presidential election and attracts many others who often retweeted his tweets, especially Republican supporters. @Norsu2 posted many tweets about the news topic while only some of them got retweeted by a few users. He is assigned a score of 1 for his content influential score. @redostoneage posted a huge amount of tweets about O_1 while few of them got retweeted by others. He is more likely to be a robot rather than an ordinary Twitter user. So he is assigned a content influential score of 0. For authority-based influential Twitter users, @MittRomney is assigned an authority influential score of 2 because it is the verified Twitter account of presidential election nominee from Republican Party. @rotolo is a professor from Syracuse University whose major is Information Science. Although his major is different from the news topic, assessors still assign him an authority influential score of 1 because he has a high social position and his tweets about O_1 are still reliable. Other users who posted unrelated tweets are assigned zero.

After assessors finish assigning scores for all users, we calculate the Discounted Cumulative Gain (DCG) [11] of top-15 users found by each method. The DCG is calculated as follows:

Table 6 Top-15 Content-based and authority-based Twitter users found by RR and MR for news topic O_1

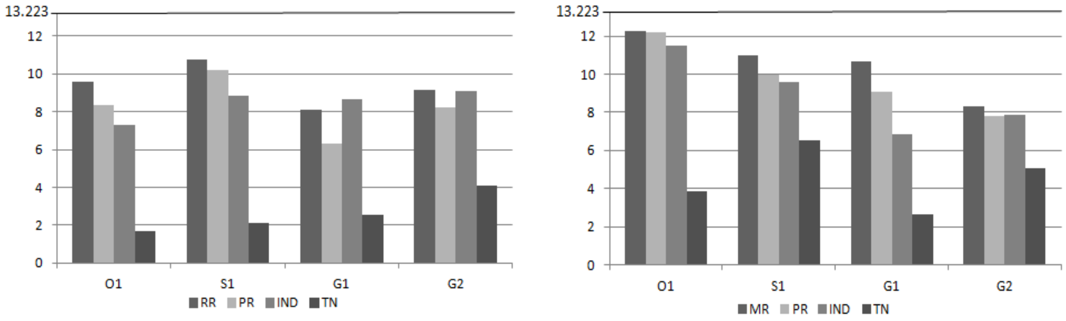
	RR	MR
	screen_name	**screen_name**
1	@PatDollard	@MittRomney
2	@LeftsideAnnie	@PaulRyanVP
3	@PaulRyanVP	@MarthaRaddatz
4	@NETRetired	@140elect
5	@maxnrgmike	@AC360
6	@BlueDuPage	@edshow
7	@NathanHale1775	@rotolo
8	@redostoneage	@InesMergel
9	@Norsu2	@andersoncooper
10	@CoffeeBean26	@rickklein
11	@ConNewsNow	@jonkarl
12	@chasepolitics	@FlakeforSenate
13	@retfado	@GOP
14	@Conservativeind	@cspan
15	@DarrellIssa	@DarrellIssa

$$\mathrm{DCG}_{15} = score_1 + \sum_{i=2}^{15} \frac{score_i}{\log_2 i} \tag{26}$$

where $score_i$ is the averaged content influential score or authority influential score manually assigned by assessors for the i-th user. DCG value ranges from 0 to 13.223. It considers not only user's influential scores, but also their ranking position. Method whose DCG value is larger could rank users often posting valuable tweets or having high authority on the news topic higher and outperform other methods whose DCG values are small.

Before showing evaluation results, someone may think that tweets posted by authority-based influential Twitter users may also be valuable and get retweeted many times because these tweets are highly trustable. However, as we can observe in Table 6, there are few users which are taken as both content-based and authority-based influential Twitter users. That's because most of tweets posted by authority-based influential Twitter users often concern the latest evolvement of the news topic while tweets from content-based influential Twitter users are more opinionated and more likely to attract user's interest since users in Twitter often share opinions on variety of topics and discuss current issues [22].

Evaluation results are shown in Figure 6 (a) and Figure 6 (b). As we can observe, DCG_{15} values of RR and MR for these news topics are larger than the others in most cases, which means RR and MR outperform other related methods. TN performs the worst compared with other methods, which means the number of tweets posted by the user is not a good indicator for his influence because these tweets might be ignored by his followers. IND seems to be reasonable to measure the influence. However, notice that retweet and mention are often used for campaigns, e.g. marketing campaign, in Twitter to gain reputation. These retweets/mentions are not suitable for measuring user's influence. Also, IND



(a) DCG for content-based influential Twitter users

(b) DCG for authority-based influential Twitter users

Figure 6 DCG for top-15 content-based/authority-based influential Twitter users about four news topics

ignores link structure among users. The link structure of user's retweet/mention activities is helpful to find influential Twitter users, which has been proved by better performances of PR, RR, and MR. Newly proposed RR and MR outperform PR because they consider user's retweet/mention preference for the topic and user's topic relevance. PR ignores these, causing negative affection to its results.

As we can also observe that RR and MR are not always better than the others in some cases. One explanation for this is that due to the rate limit of Twitter API, it is hard to collect all tweets related to news topics. For some news topics, vertices (users) in G_{RT} and G_{MN} are not well connected. For example, in retweet graph of G_1, the average in-degree of vertex is 0.744, which is the lowest in all retweet and mention graphs. This means that users are not well connected. RR does not perform better than IND in this retweet graph. However, RR and MR give better results in other graphs having higher average in-degree per vertex.

6 Conclusion

In this paper, we proposed RetweetRank and MentionRank to find content-based and authority-based influential Twitter users from hashtag communities which are relevant to a news topic searched by an input target word. As basic components of our research, we proposed PIOLog and PIOLogH methods to detect/weight characteristic co-occurrence words with the target word/hashtag from news articles/tweets. News-topic-related hashtags with corresponding hashtag communities are detected based on these characteristic co-occurrence words. For users in news-topic-related hashtag communities, RetweetRank and MentionRank are used to find content-based and authority-based influential Twitter users for the news topic. Experimental results showed that our PIOLog and PIOLogH methods are more likely to detect characteristic co-occurrence words for news topics and hashtags. Newly proposed RetweetRank and MentionRank could find two types of influential Twitter users from news-topic-related hashtag communities and outperform other related methods.

In the future, we are planning to improve our data collection method. More tweets related to news topics should be collected within the rate limit of Twitter API. Also, not only news topics, but also other topics discussed by Twitter users should be considered. After we manually checked contents of some tweets posted by influential Twitter users, we found that tweets posted by some users, especially content-based influential Twitter users, are highly opinionated and are strongly supported/opposed by other Twitter users. Mining opinions from tweets posted by influential Twitter users could help us understand why some opinions are popular and widely accepted, which is another research direction we are considering.

References

1. Anger, I. and Kittl, C., Measuring Influence on Twitter. In Proc. of the 11th Int. Conf. on Knowledge Management and Knowledge Technologies, pp.31:1-31:4, ACM (Graz, Austria. 2011).
2. Bigonha, C., Cardoso, T. N. C., Moro, M. M., Almeida, V. A. F. and Goncalves, M. A., Detecting Evangelists and Detractors on Twitter. In Proc. of the Brazilian Symposium on Multimedia and the Web, pp. 107-114 (Belo Horizonte, Brazil, 2010).
3. Brin, S. and Page, L., The Anatomy of A Large-scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems Vol. 30, Issue 1-7, pp. 107–117, Elsevier Science Publishers B. V. (1998).
4. Cano, A. E., Mazumdar, S., and Ciravegna, F., Social Influence Analysis in Microblogging Platforms - A Topic-sensitive based Approach. Special Issue on the Semantics of Microposts. Semantic Web Journal. pp. 1-5, IOS Press (2011).
5. Celebrating #Twitter7. http://blog.twitter.com/2013/03/celebrating-twitter7.html
6. Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K. P., Measuring User Influence in Twitter: The Million Follower Fallacy. In Proc. of the 4th Int. AAAI Conf. on Weblogs and Social Media, AAAI Press (Washington DC, USA, 2010).
7. Efron, M., Hashtag Retrieval in a Microblogging Environment. In Proc. of the 33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 787-788, ACM (Geneva, Switzerland, 2010).
8. Finkel, J.R., Grenager, T. and Manning, C., Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363-370, ACL (Ann Arbor, Michigan, 2005).
9. Hajian, B. and White, T., Modelling Influence in a Social Network: Metrics and Evaluation. In IEEE 3rd Int. Conf. on Social Computing, pp. 497-500, IEEE (Boston, MA, USA, 2011).
10. Huang, J., Thornton, K. M. and Efthimiadis, E. N., Conversational Tagging in Twitter. In Proc. of the 21st ACM Conf. on Hypertext and Hypermedia, pp. 173-178, ACM (Toronto, Canada, 2010).
11. Jarvelin K. and Kekalainen, J., Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems, Vol.20, Issue 4, pp. 422–446, ACM (2002).
12. Jones, K. S., A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation, Vol. 28, pp. 11-21. (1972).
13. Kleinberg, J. Authoritative sources in a Hyperlinked Environment. Journal of the ACM, Vol. 46, Issue 5, pp. 604-632, ACM (1999).
14. Kwak, H., Lee, C., Park, H. and Moon, S., What is Twitter, a Social Network or a News Media? In Proc. of the 19th Int. Conf. on WWW, pp. 591-600, ACM (Raleigh, North Carolina, USA, 2010).
15. Kywe, S. M., Hoang, T-A., Lim E-P. and Zhu, F., On Recommending Hashtags in Twitter Networks. In Proc. of the 4th Int. Conf. on Social Information, pp. 337-350, Springer-Verlag (Lausanne, Switzerland, 2012).
16. Leavitt, A., Burchard, E., Fisher, D. and Gilbert, S. (2009). The Influentials: New Approaches for Analyzing Influence on Twitter. Web Ecology Project. http://www.webecologyproject.org/wp-content/uploads/2009/09/influence-report-final.pdf
17. Lehmann, J., Gonçalves, B., Ramasco, J. J. and Cattuto, C., Dynamical Classes of Collective Attention in Twitter. In Proc. of the 21st Int. Conf. on WWW, pp. 251-260, ACM (Lyon, France, 2012).
18. Manning, C. D. and Schutze, H., Foundations of Statistical Natural Language Processing. MIT Press (1999).

19. Mazzia, A. and Juett, J. (2011), Suggesting Hashtags on Twitter. EECS 545 Project, http://www-personal.umich.edu/~amazzia/pubs/545-final.pdf
20. Microblogging. http://en.wikipedia.org/wiki/Microblogging
21. Noro, T., Ru, F., Xiao, F. and Tokuda, T., Twitter User Rank Using Keyword Search. In the 22nd European-Japanese Conf. on Information Modelling and Knowledge Bases, pp. 31-48, IOS Press (Prague, 2012).
22. Pak A. and Paroubek P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proc. of the 7th Conf. on Int. Language Resources and Evaluation. pp. 1320-1326, ELRA (Valletta, Malta, 2010).
23. Romero, D. M., Galuba, W., Asur, S. and Huberman, B. A., Influence and Passivity in Social Media. In ECML/PKDD, pp. 18-33, Springer-Verlag (Athens, Greece, 2011).
24. Salton, G., Wong, A. and Yang, C.S., A Vector Space Model for Automatic Indexing. Communications of the ACM, Vol. 18, Issue 11, pp. 613-620, ACM (1975).
25. Schmid, H., Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proc. of Int. Conf. on New Methods in Language Processing, pp. 44-49 (Manchester, UK, 1994).
26. Singhal, A., Buckley, C. and Mitra, M., Pivoted Document Length Normalization. In Proc. of the 19th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 21–29, ACM (Zurich, Switzerland, 1996).
27. Weng, J., Lim, E. P., He, Q. and Leung, C. W.-K., What Do People Want in Microblogs? Measuring Interestingness of Hashtags in Twitter. In Proc. of the 2010 IEEE Int. Conf. on Data Mining, pp.1121-1126, IEEE (2010).
28. Weng, J., Lim, E. P., Jiang, J. and He, Q., TwitterRank: Finding Topic-sensitive Influential Twitterers. In Proc. of the 3rd Int. Conf. on Web Search and Data Mining, pp. 261-270, ACM (New York, USA, 2010).
29. What Facebook and Twitter Mean for News. http://stateofthemedia.org/files/2012/03/Facebook-and-Twitter-Topline.pdf
30. Xiao, F., Noro, T. and Tokuda, T., Detection of Characteristic Co-occurrence Words from News Articles on the Web. In the 21st European-Japanese Conf. on Information Modelling and Knowledge Bases, Vol.237, pp.187-203, IOS Press (Tallinn, Estonia, 2011).
31. Xiao, F., Noro, T. and Tokuda, T., News-Topic Oriented Hashtag Recommendation in Twitter Based on Characteristic Co-occurrence Word Detection. In Proc. of the 12th Int. Conf. on Web Engineering, pp.16-30, Springer-Verlag (Berlin, Germany. 2012).
32. Yang, L., Sun, T., Zhang, M., and Mei, Q., We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In Proc. of the 21st Int. Conf. on WWW, pp. 261-270, ACM (Lyon, France, 2012).
33. Ye, S. and Wu, S. F., Measuring Message Propagation and Social Influence on Twitter.com. In Proc. of the 2nd Int. Conf. on Social Informatics, pp. 216-231, Springer-Verlag (Laxenburg, Austria, 2010).
34. Zangerle, E., Gassler, W. and Specht, G., Recommending #-Tags in Twitter, In Proc. of the Workshop on Semantic Adaptive Social Web. CEUR Workshop Proceedings, pp. 67-78 (2011).