TwitterRank: Finding Topic-sensitive Influential Twitterers

Jianshu Weng, Ee-Peng Lim, Jing Jiang
School of Information Systems
Singapore Management University
{jsweng,eplim,jingjiang}@smu.edu.sg

Qi He
College of Information Sciences and Technology
Pennsylvania State University
qhe@ist.psu.edu

ABSTRACT
This paper focuses on the problem of identifying influential users of micro-blogging services. Twitter, one of the most notable micro-blogging services, employs a social-networking model called “following”, in which each user can choose who she wants to “follow” to receive tweets from without requiring the latter to give permission first. In a dataset prepared for this study, it is observed that (1) 72.4% of the users in Twitter follow more than 80% of their followers, and (2) 80.5% of the users have 80% of their followers who are following them back. Our study reveals that the presence of “reciprocity” can be explained by phenomenon of homophily [14]. Based on this finding, TwitterRank, an extension of PageRank algorithm, is proposed to measure the influence of users in Twitter. TwitterRank measures the influence taking both the topical similarity between users and the link structure into account. Experimental results show that TwitterRank outperforms the one Twitter currently uses and other related algorithms, including the original PageRank and Topic-sensitive PageRank.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Information Search and Retrieval—information filtering, retrieval model;
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—indexing methods

General Terms
Algorithms, Design, Experimentation

Keywords
Twitter, influential, PageRank

1. INTRODUCTION
Micro-blogging is an emerging form of communication. It allows users to publish brief message updates, which can be submitted in many different channels, including the Web and text messaging service [1, 16]. One of the most notable micro-blogging services is Twitter. It allows twitterers to publish tweets (with a limit of 140 characters). Twitter also provides the “social-networking” functionality. Unlike other social network services that require users to grant friend links to other users befriending them, Twitter employs a social-networking model called “following”, in which each twitterer is allowed to choose who she wants to follow without seeking any permission. Conversely, she may also be followed by others without granting permission first. In one instance of “following” relationship, the twitterer whose updates are being followed is called the “friend”, while the one who is following is called the “follower”.

Twitter has gained huge popularity since the first day that it was launched [16, 4]. It has also drawn increasing interests from research community. There is previous work [8] to study the topological and geographical properties of the social network formed by the twitterers and their followers. In this paper, we are interested in identifying the influential twitterers. The benefit of solving this problem is multi-fold. First, it potentially brings order to the real-time web in that it allows the search results to be sorted by the authority/influence of the contributing twitterers giving a timely update of the thoughts of influential twitterers. Second, according to [16], Twitter is also a marketing platform. Targeting those influential users will increase the efficiency of the marketing campaign [9, 10]. For example, a handphone manufacturer can engage those twitterers influential in topics about IT gadgets to potentially influence more people. There are also applications that utilize Twitter to gather opinions and information on particular topics. Identifying influential twitterers for interesting topics can improve the quality of the opinions gathered.

Currently, Twitter and many other applications interpret a twitterer’s influence as the number of followers she has. However, is this really a good indicator of influence? In a dataset prepared for this study, it is observed that (1) 72.4% of the users follow more than 80% of their followers, and (2) 80.5% of the user have 80% of their friends follow them back. Two seemingly conflicting reasons can possibly...

1 Another similar service is Plurk.
2 Users in Twitter are usually dubbed twitterers, and the short message updates published by the users tweets.
3 In this paper, an influential twitterer is one with certain authority within her social network.
explain such “reciprocity”. First, the “following” relationship is so casual that each twitterer just randomly follows someone, and those being followed follow back just for the sake of courtesy. Second, it might be the opposite, i.e., the “following” relationship is a strong indicator of the similarity among users. In other words, a twitterer follows a friend because she is interested in the topics the friend publishes in tweets, and the friend follows back because she finds they share similar topic interest. This phenomenon is called “homophily”, which has been observed in many social networks [14]. The cause of such “reciprocity” has important implication here. If it is caused by the first reason, identifying the influential twitterers based on “following” relationship would be rendered meaningless since the “following” relationship itself does not carry strong indication of influence. On the other hand, the presence of homophily indicates that the “following” relationships between twitterers are related to their topical similarity.

Our study confirms that homophily does exist in the context of Twitter. This justifies that there are some twitterers who do seriously “follow” someone because of common topical interests instead of just playing a “number game”. Based on this observation, we propose a novel approach to measure the influence of twitterers, known as TwitterRank. The framework of the proposed approach is shown in Figure 1. First, topics that twitterers are interested in are distilled automatically by analyzing the content of their tweets. Based on the topics distilled, topic-specific relationship networks among twitterers are constructed. Finally, we apply our TwitterRank algorithm, which is an extension of PageRank, to measure the influence taking both the topical similarity between twitterers and the link structure into account.

![Figure 1: Framework of the Proposed Approach](image)

This paper improves the state-of-the-art by making two contributions. First, to the best of our knowledge, this paper is the first to report homophily in Twitter. Second, it introduces TwitterRank to measure the topic-sensitive influence of the twitterers. Prior to this, a twitterer’s influence is often measured by her node in-degree in the network, i.e., the number of followers. However, as observed in previous social network analysis studies [12, 3], in-degree does not accurately capture the notion of influence. PageRank improves over in-degree by considering the link structure of the whole network [3]. Nevertheless, Pagerank ignores the interests of twitterers, which affects the way twitterers influence one another. Our proposed approach addresses the shortcomings of in-degree and PageRank by taking into account both the link structure and topical similarity among twitterers.

The rest of this paper is organized as follows: A Twitter dataset has been prepared for the purpose of this study. Section 2 describes in detail how the dataset is prepared. Topic distillation and the phenomenon of homophily observed in the dataset is elaborated in Section 3, while TwitterRank is proposed in Section 4. Section 5 presents the experimental results, comparing TwitterRank with the benchmark method currently used by Twitter and other related algorithms. Section 6 briefly summarizes related work. Finally, Section 7 concludes with directions for further research.

2. TWITTER DATASET

For the purpose of this study, a set of Twitter data about Singapore-based twitterers was prepared in April, 2009 as follows:

1. We obtained a set of top-1000 Singapore-based twitterers\(^5\) from twitterholic.com\(^6\). Denote this set as $\mathcal{S}$. As four of the top-1000 twitterers were not available when the dataset was being prepared, $|\mathcal{S}| = 996$.

2. We then crawled\(^7\) all the followers and friends of each individual twitterer $s \in \mathcal{S}$ and stored them in set $\bar{\mathcal{S}}$.

3. Let $\mathcal{S}' = \mathcal{S} \cup \bar{\mathcal{S}}$, and $\mathcal{S}^* = \{s | s \in \mathcal{S}'$, and $s$ is from Singapore$. $|\mathcal{S}^*| = 6748$. For each $s \in \mathcal{S}^*$, we crawled\(^7\) all the tweets she had published so far. Denote the set of all the tweets obtained as $\mathcal{T}$. $|\mathcal{T}| = 1,021,039$.

2.1 Tweet Distribution

The latest tweet in the dataset was published on April 25, 2009, while the earliest one was on July 18, 2006. Numbers of tweets by month during the time period captured in the dataset are plotted in Figure 2. It shows that Twitter started to attract substantial attention from Singapore-based twitterers from March 2008 onwards.

![Figure 2: Number of Tweets per Month](image)

Out of the 6748 twitterers in the dataset, only 5686 publish at least one tweet. For those 5686 twitterers, the average number of tweets each publishes is 179.57. The distribution of the tweets per twitterer is shown in Figure 3. If we do not consider the “outliers” indicated by the red circle\(^8\), it follows a power-law distribution. The presence of “outliers” in the dataset is caused by a restriction implemented by Twitter, which limits the maximum number of tweets visible to be 3200 even a twitterer has published more than 3200 tweets.

There are 30 such active twitterers in the dataset. Four of them are bots that publish tweets directly obtained from Twitter.
Count of tweets published
Number of twitterers
Figure 3: Distribution of Tweets per Twitterer

RSS feeds (usually more than one feed) they have subscribed to. We excluded two bots, one always re-publishing followers' tweets and another publishing only numbers. A spammer frequently publishing his username and URL in tweets was also excluded.

2.2 Friends/Followers

There are in total 49872 "following" relationships among the twitterers in $S^*$. Among the 6745 twitterers, 957 have no friends, while 1782 have no followers. The distribution of the numbers of the friends/followers each twitterer has are plotted in Figures 4(a) and 4(b) respectively. They again follow power-law distribution.

2.3 Reciprocity in Following Relationships

Reciprocity in following relationships is prevalent in Twitter. We examine this reciprocity by showing the correlation between number of friends and number of followers for each twitterer in Figure 5. It shows that the more friends a twitterer has, the more followers she has, and vice versa. A closer examination of the dataset reveals that there is high chance of “reciprocity” presented in the “following” relationships.

- 72.4% of the twitterers follow more than 80% of their followers,
- and 80.5% of the twitterers have 80% of their friends follow them back.

Figure 5: Number of Friends vs. Number of Followers

3. HOMOPHILY IN TWITTER

As mentioned in Section 1, two conflicting reasons can possibly explain such a “reciprocity”, i.e., twitterers’ casual “following” behaviors versus homophily. Homophily is a phenomenon showing that people’s social networks “are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics” [14]. In the context of Twitter, homophily implies that a twitterer follows a friend because she is interested in some topics the friend is publishing, and the friend follows back because she finds they share similar topical interest.

Although it is beyond the scope of this paper to find the real cause of the “reciprocity” in the “following” relationships for each twitterer, the presence of homophily implies that there are twitterers who are serious in choosing friends to follow. This implication is important in that identifying the influential twitterers based on the “following” relationships would be rendered meaningless if no twitterer is serious in “following” others. Two questions would help to verify whether homophily presents in the context of Twitter:

Question 1: Are twitterers with “following” relationships more similar than those without according to the topics they are interested in?

Question 2: Are twitterers with reciprocal “following” relationships more similar than those without according to the topics they are interested in?

To answer these questions, we need to know the topics that twitterers are interested in and to measure the topical similarity between pairs of twitterers. However, topic interests are not explicitly expressed by twitterers. A possible

---

3 In fact, our experimental results show that the two reasons may co-exist in the context of Twitter.
solution is to use the “#hashtag” uttered by twitterers\textsuperscript{10}. Nevertheless, there is a very low usage of “#hashtag” in the dataset, which makes “#hashtag’s not appropriate to be used as topics. To overcome this challenge, topic modeling, which is commonly used to analyze large volumes of unlabeled contents, is applied to automatically distill topics.

3.1 Topic Distillation

The goal of the topic distillation is to automatically identify the topics that twitterers are interested in based on the tweets they published. For this purpose, Latent Dirichlet Allocation (LDA) model\textsuperscript{2, 18, 6} is applied, which is an unsupervised machine learning technique to identify latent topic information from large document collection. It uses a “bag of words” assumption, which treats each document as a vector of word counts. Based on this assumption, each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. It also assumes a generative process for generating each document as follows:

1. for each document, pick a topic from its distribution over topics,
2. sample a word from the distribution over the words associated with the chosen topic,
3. the process is repeated for all the words in the document.

More formally, each of a collection of $D$ documents is associated with a multinomial distribution over $T$ topics, which is denoted as $\theta$. Each topic is associated with a multinomial distribution over words, denoted as $\phi$. $\theta$ and $\phi$ have Dirichlet prior with hyper-parameters $\alpha$ and $\beta$ respectively. For each word in one document $d$, a topic $z$ is sampled from the multinomial distribution $\theta$ associated with the document, and a word $w$ from the multinomial distribution $\phi$ associated with topic $z$ is sampled consequently. This generative process is repeated $N_d$ times ($N_d$ is the total number of words in document $d$) to form document $d$\textsuperscript{2, 18, 6}. This generative process can be graphically represented using commonly-used plate notation in Figure 6. In this figure, shaded and unshaded plates indicate observed and latent variables respectively. An arrow corresponds to a conditional dependency between two variables and boxes indicate repeated sampling with the number of repetitions given by the variable in the bottom of the corresponding box.

![Figure 6: Graphical Representation of LDA Model](image)

The model has two parameters to be inferred from the data, i.e. document-topic distributions $\theta$, and the $T$ topic-word distributions $\phi$. By learning these two parameters, information can be obtained about which topics authors typically write about as well as a representation of the content of each document in terms of these topics. In this study, Gibbs sampling is applied for model parameter estimation\textsuperscript{11}.

To distill the topics that twitterers are interested in using LDA, documents should naturally correspond to tweets. However, since the goal is to understand the topics that each twitter is interested in rather than the topic that each single tweet is about, we aggregate the tweets published by individual twitterer into a big document. Thus, each document essentially corresponds to a twitterer.

The result is represented in three matrices:

1. $DT$, a $D \times T$ matrix, where $D$ is the number of twitterers and $T$ is the number of topics. $DT_{ij}$ contains the number of times a word in twitterer $s_i$’s tweets has been assigned to topic $t_j$.
2. $WT$, a $W \times T$ matrix, where $W$ is the number of unique words used in the tweets and $T$ is the number of topics. $WT_{ij}$ captures the number of times unique word $w_i$ has been assigned to topic $t_j$.
3. and $Z$, a $1 \times N$ vector, where $N$ is the total number of words in the tweets. $Z_i$ is the topic assignment for word $w_i$.

3.2 Hypothesis Testing

Among the three matrices in the result of topic distillation, matrix $DT$ is of particular interest. It contains the number of times a word in a twitterer’s tweets has been assigned to a particular topic. We can row normalize it as $DT'$ such that $||DT'_i|| = 1$ for each row $DT'_i$. Each row of matrix $DT'$ is basically the probability distribution of twitterer $s_i$’s interest over the $T$ topics, i.e. each element $DT'_{ij}$ captures the probability that twitterer $s_i$ is interested in topic $t_j$. Given this, the topical difference between twitterers can be measured as follows.

**Definition 1.** Topical difference between two twitterers $s_i$ and $s_j$ can be calculated as:

$$\text{dist}(i, j) = \sqrt{2 * D_{JS}(i,j)}$$ (1)

$D_{JS}(i,j)$ is the Jensen-Shannon Divergence between the two probability distributions $DT'_{i}$ and $DT'_{j}$, which is defined as:

$$D_{JS}(i,j) = \frac{1}{2} (D_{KL}(DT'_{i}||M) + D_{KL}(DT'_{j}||M))$$ (2)

$M$ is the average of the two probability distributions, i.e. $M = \frac{1}{2}(DT'_i + DT'_j)$. $D_{KL}$ in Eq (2) is the Kullback-Leibler Divergence which defines the divergence from distribution $Q$ to $P$ as: $D_{KL}(P||Q) = \sum P(i) \log \frac{P(i)}{Q(i)}$.

\textsuperscript{10}Twitter allows each twitterer to associate a hashtag to the tweets. Hashtags in tweets are equivalent to tags typically found in content sharing services, e.g. Youtube and Flickr.

\textsuperscript{11}It has been proved $\sqrt{2 * D_{JS}(i,j)}$ is a metric for probability distributions which fulfills the triangle inequality [5]. Another reason it is used here is that it reduces the non-normality of the data, which will potentially influence the robustness of the t-test.
With the topical difference measured, the two questions listed in the beginning of Section 3 can be answered by statistical hypothesis testing. It is noted that, in this study, hypothesis testing, and topic distillation as well, is applied on a set of twitterers who publish more than 10 tweets in total. We denote this set as $S^*$, and $|S^*| = 4050$.

3.2.1 Question 1

**Question 1** can be formalized as a two-sample $t$-test:

Let $μ_{\text{follow}}$ be the mean topical difference of the pairs of twitterers with “following” relationship, and $μ_{\text{nofollow}}$ the mean topical difference of those without.

The null hypothesis is $H_0 : μ_{\text{follow}} = μ_{\text{nofollow}}$, and the alternative hypothesis is $H_1 : μ_{\text{follow}} < μ_{\text{nofollow}}$.

Ideally, individual statistical hypothesis testing shall be conducted for each twitterer. Nevertheless, most of the twitterers (3785 out of 4050) have less than 30 friends, which is not statistically significant. Therefore, two cases are considered when answering **Question 1**.

**Case 1.**

Denote the set of twitterers with more than 30 friends as $S_{\text{U}}$, and $|S_{\text{U}}| = 265$. Individual statistical hypothesis test is conducted for every twitterer $s_i \in S_{\text{U}}$. First, calculate the topical difference between $s_i$ and each of her friends, based on which $μ_{\text{follow}}$ is calculated. Then, choose some twitterers uniformly at random from those $s_i$ does not follow, and the number of the chosen non-friends is same as the number of $s_i$’s friends. Calculate the topical difference between $s_i$ and each non-friend, based on which $μ_{\text{nofollow}}$ is calculated. Finally, a two-sample $t$-test (under the assumption of unequal population variances) is conducted on the two populations formed with the above approach.

Results shows that for 232 out of the 265 twitterers with more than 30 friends, the null hypothesis is rejected at significant level $α = 0.01$\(^{13}\).

**Case 2.**

Denote the set of twitterers with less than 30 friends as $S_{\text{L}}$, and $|S_{\text{L}}| = 3785$. For this set of twitterers, the hypothesis testing is conducted on the twitterer congregation. First of all, calculate the topical difference for all the pairs of twitterers whose “following” relationship are initiated by any twitterer $s_i \in S_{\text{L}}$, based on which $μ_{\text{follow}}$ is calculated. Then, for each $s_i \in S_{\text{L}}$, choose some non-friends uniformly at random, the number of the chosen non-friends is same as the number of $s_i$’s friends. Congregate all the pairs of twitterers, and calculate the difference between each pair, based on which $μ_{\text{nofollow}}$ is calculated. Finally, a two-sample $t$-test (under the assumption of unequal population variances) is conducted on the two populations formed with the above approach. The test outcome is that the null hypothesis is rejected at significant level $α = 0.01$ with a $p$-value of $4.5 \times 10^{-6}$.

Together with the results in **Case 1**, the answer to **Question 1** is clearly that with very high probability, twitterers with “following” relationships are more similar than those without according to the topics they are interested in.

3.2.2 Question 2

**Question 2** is also formalized as a two-sample $t$-test:

Let $μ_{\text{sym}}$ be the mean topical difference of the pairs of twitterers with reciprocal “following” relationship, and $μ_{\text{asym}}$ the mean topical difference of those without.

The null hypothesis is $H_0 : μ_{\text{sym}} = μ_{\text{asym}}$, and the alternative hypothesis is $H_1 : μ_{\text{sym}} < μ_{\text{asym}}$.

There are in total 11505 pairs of twitterers with reciprocal “following” relationship. However, there are only 67 twitterers with more than 30 reciprocal and non-reciprocal friends respectively. Hence, we conduct the two-sample $t$-test on the twitterer congregation. First of all, calculate the topical difference for all the pairs of twitterers with reciprocal “following” relationship, based on which $μ_{\text{sym}}$ is calculated. Then, for each twitterer, choose some non-reciprocal friends uniformly at random such that the number of the chosen non-reciprocal friends is same as the number of reciprocal friends she has. Congregate all the non-reciprocal relationships, and calculate the topical difference for each non-reciprocal relationship, based on which $μ_{\text{asym}}$ is calculated. With the above two populations, the null hypothesis is rejected at significant level $α = 0.01$ with a $p$-value of $1.2 \times 10^{-6}$. This outcome gives a positive answer to **Question 2** that with very high probability, twitterers with reciprocal “following” relationships are more similar than those without according to the topics they are interested in.

Positive answers to both **Question 1** and **Question 2** provide evidences to the existence of the homophily phenomenon in the Twitter dataset. Based on this finding, a novel approach to measure twitterers’ influence is proposed in the next section.

4. TOPIC-SENSITIVE INFLUENCE MEASURE

Intuitively, the influence of a twitterer can be interpreted similar to the “authority” of a web page: a twitterer has high influence if the sum of influence of her followers is high; at the same time, her influence on each follower is purely based on the relative amount of content the follower received from her. This similarity motivates the use of PageRank in measuring influence.

Although the “authority” of web page and influence of twitterer shares certain similarities, there are also major differences. The influence on each follower is purely based on relative amount of content the follower receives as the latter may not read content with topics less interesting even when the relative content is large. Since twitterers generally have different expertise and/or interests in various topics, influence of twitterers also vary in different topics. Given this, a topic-sensitive TwitterRank is proposed to measure the influence of twitterers.

4.1 Topic-specific TwitterRank

First of all, a directed graph $D(V, E)$ is formed with the twitterers and the “following” relationships among them. $V$ is the vertex set, which contains all the twitterers. $E$ is the edge set. There is an edge between two twitterers if there is “following” relationship between them, and the edge is directed from follower to friend.
A random surfer model on graph $D$ computes the TwitterRank as follows: the random surfer visits each twitterer with certain probability by following the appropriate edge in $D$. TwitterRank differentiates itself from PageRank in that the random surfer performs a topic-specific random walk, i.e. the transition probability from one twitterer to another is topic-specific. By doing so, we are essentially constructing a topic-specific relationship network among twitterers.

The transition matrix for topic $t$, denoted as $P_t$, is defined as follows.

**Definition 2.** Given a topic $t$, each element of matrix $P_t$, i.e. the transition probability of the random surfer from follower $s_i$ to friend $s_j$, is defined as:

$$P_t(i, j) = \frac{|T_j|}{\sum_{s_i \text{ follows } s_a} |T_a|} \cdot \text{sim}_t(i, j)$$

where $|T_j|$ is the number of tweets published by $s_j$, and $\sum_{s_i \text{ follows } s_a} |T_a|$ sums up the number of tweets published by all of $s_i$’s friends. $\text{sim}_t(i, j)$ in Eq. (3) is the similarity between $s_i$ and $s_j$ in topic $t$, which is defined as:

$$\text{sim}_t(i, j) = 1 - |D_{ij}^t - D_{ji}^t|$$

This definition captures two notions. Assume twitterer $s_i$ follows a number of friends. Those friends publish different numbers of tweets, all of which will be directly visible to $s_i$. The more a friend $s_j$ publishes, the higher portion of tweets $s_i$ reads from $s_j$. Generally, this leads to a higher influence on $s_i$, which corresponds to a higher transition probability from $s_i$ to $s_j$. This intuition is captured in the first term in the RHS of Eq. (3). Figure 7 shows an example about three twitterers. $s_a$ follows $s_b$ and $s_c$, who publish 500 and 1000 tweets respectively. In this case, $s_a$’s influence on $s_c$ is two times of that of $s_b$ when the topical similarity among the three twitterers is not taken into account.

**Figure 7:** Example of Transition Probability Calculation

Second, $s_j$’s influence on $s_i$ is also related to the topical similarity between the two as suggested by the homophily phenomenon discussed in Section 3. Row-normalized matrix $DT^t$ is one of the results in the topic distillation. A row $DT^t_{ij}$ contains the probability of twitterer $s_j$’s interest in different topics. The similarity between $s_i$ and $s_j$ in topic $t$ can be evaluated as the difference between the probability that the two twitterers are interested in the same topic $t$, which is basically the second term in the RHS of Eq. (3). The more similar the two twitterers are, the higher the transition probability from $s_i$ to $s_j$.

It is possible that some twitterers would “follow” one another in a looping manner without “following” other twitterers outside the loop. Such loop will accumulate high influence without distribute their influence. To tackle this, a teleportation vector $E_t$ is also introduced, which basically captures the probability that the random surfer would “jump” to some twitterers instead of following the edges of the graph $D$. $E_t$ is defined as follows.

**Definition 3.** The teleportation vector of the random surfer in topic $t$ is defined as:

$$E_t = DT^t_{ii}$$

$DT^t_{ii}$ is the $t$-th column of matrix $DT^t$, which is the column-normalized form of matrix $DT$ such that $\|DT^t\|_1 = 1$. $DT$ is one of the results obtained during the topic distillation, each entry of which contains the numbers of times words in a twitterer’s tweets has been assigned to a specific topic.

With the transition probability matrix and teleportation vector defined, the topic-specific TwitterRank can be calculated.

**Definition 4.** The topic-specific TwitterRank of the twitterers in topic $t$, denoted as $\overrightarrow{TR}_t$, can be calculated iteratively by:

$$\overrightarrow{TR}_t = \gamma P_t \cdot \overrightarrow{TR}_t + (1 - \gamma)E_t$$

$P_t$ is the transition probability matrix defined in Eq. (3), $E_t$ is the teleportation vector defined in Eq. (5). $\gamma$ is a parameter between 0 and 1 to control the probability of teleportation. The lower $\gamma$ is, the higher probability the random surfer will teleport to twitterers according to $E_t$, and vice versa.

**4.2 Aggregation of Topic-specific TwitterRank**

The approach presented in Section 4.1 generates a set of topic-specific TwitterRank vectors, which basically measure the twitterers’ influence in individual topics. An aggregation of TwitterRank can also be obtained to measure twitterers’ overall influence.

**Definition 5.** Twitterers’ general influence can be measured as an aggregation of the topic-specific TwitterRank in different topics, which is calculated as:

$$\overrightarrow{TR} = \sum_t r_t \cdot \overrightarrow{TR}_t$$

$\overrightarrow{TR}_t$ is the TwitterRank vector for topic $t$, while $r_t$ is the weight assigned to topic $t$ and associated $\overrightarrow{TR}_t$.

Depending on the applications, different set of weights can be assigned to derive the influence of twitterers in different scenarios.

**General influence:** $r_t$’s can be set as the probabilities of different topics’ presence, which are calculated according to the number of times unique words have been assigned to corresponding topics as captured in matrix $WT$. In this case, the aggregation of TwitterRank is essentially the twitterers’ general influence.

**Perceived general influence:** $r_t$’s can also be set as the probabilities that a particular twitterer $s_i$ is interested in different topics, which are calculated according to
the number of times words in $s_i$'s tweets have been assigned to corresponding topics as captured in matrix $DT$. In this case, the aggregation of TwitterRank becomes $s_i$'s personal perception of twitterers' general influence.

5. EMPIRICAL EVALUATION

Section 4 proposes TwitterRank, which measures different twitterers’ influence by taking into account the topical similarity among twitterers as well as the link structure. This section shows the results of applying TwitterRank in our Twitter dataset. We also elaborate on an evaluation procedure for effective comparison with other related algorithms.

5.1 Influential twitterers identified in the Twitter dataset

We first compare the most influential twitterers identified by TwitterRank with the most active twitterers identified during topic distillation.

As mentioned in Section 3.2, topic distillation is applied on a set of twitterers who publish more than 10 tweets in total. We denote this set as $S^*_n$, and $|S^*_n| = 4050$. All the experiments in the rest of this paper is conducted on this set of twitterers and their tweets. The tweets in the dataset are written with a mixture of different languages including Chinese, English, French, German, Japanese, etc. We removed from tweets those words containing non-English characters, stopwords, punctuations, numbers, URLs, words with less than 3 characters, and words in the form “@username”. These words do not help in topic modeling. The remaining words are stemmed. LDA is conditioned on three parameters, i.e. Dirichlet hyper-parameters $\alpha, \beta$, and topic number $T$. In this paper, they are set as $T = 50, \alpha = 50/T, \beta = 0.1$.

Table 1 lists the top-5 active and influential twitterers in the five top topics. Top topics are identified in the order of the probabilities of topic presence, which are calculated according to the number of times unique words have been assigned to corresponding topics as captured in matrix $WT$ (see Section 3.1). It is observed that the active twitterers are not necessarily influential in each topic.

The results in Table 1 are reasonable. Twitterers “mrbrown”, “moby74”, and “kormmandos” are among the top-5 influential twitterers in all the five top topics identified in the dataset. “mrbrown” mainly tweets about Singapore citizen life and IT-related news. He also tweets often about things happened in his office or during his trips, as well as those in his bicycle ride from home to office. The words frequently used in expressing these topics are captured in the five top topics as shown in Tables 1. Additionally, “mrbrown” has the highest number of followers (as captured in the dataset), including some influential ones like “AngMoGirl”, “claudia10”, and “moby74”.

“moby74” tweets mainly about work, family life, food, and IT-related topics (such as the features of Twitter, website design, and Internet connection speed). Although “moby74” has much fewer followers than “mrbrown”, “moby74” has

5.2 Comparison with related algorithms

In this section, we study quantitatively the effectiveness of the proposed TwitterRank. Comparisons against related algorithms are also conducted. The related algorithms studied include:

- **In-degree**, which measures the influence of twitterers by the number of followers. This is the measurement currently employed by Twitter and many other third-party services, such as twitterholic.com and wefollow.com.

- **PageRank**, which measures the influence with only link structure of the network taken into account [3];

- **Topic-sensitive PageRank**, which measures topic-specific influence by calculating PageRank vector for each topic. Nevertheless, unlike TwitterRank, same relationship network, i.e., same transition probability matrix is used for different topics, but with a topic-biased teleportation vector [7].

For ease of presentation, the proposed TwitterRank is denoted as TR, and the three related algorithms are abbreviated to InD, PR, and TSPR respectively.
5.2.1 Correlation

We first study the correlation between the rank lists generated by the different algorithms. The correlation is measured as the Kendall’s τ [11]. τ takes value in the range of [−1, 1]. If the two lists are exactly the same, τ = 1; whereas τ = −1 if one list is the reverse of the other. For other values in the range, a larger value of τ implies higher agreement between the two lists.

Table 2(a) lists the τ values between the rank lists generated by various algorithms studied. For TR and TSPR, we apply the across-topic aggregation mentioned as “general influence” in Section 4.2. It is observed that TR generates ranked list different from those generated by other algorithms since τ ≠ 1. It is also observed that TR has higher agreement with both TR and TSPR than with InD and PR. This is because both TR and TSPR consider the topical dimension while InD and PR do not. We have also studied the correlation between the four algorithms in different topics, the same trend is observed. Table 2(b) lists τ values between the rank lists by the four algorithms in the 5 top topics listed in Tables 1.

Table 2: Correlation between Rank Lists by Different Algorithms

(a) General Rank

<table>
<thead>
<tr>
<th>Topic #</th>
<th>Associated Words</th>
<th>Active Twitterers</th>
<th>Influential Twitterers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>work morn time</td>
<td>nikipaniki annusberry</td>
<td>mrbrown moby74</td>
</tr>
<tr>
<td></td>
<td>night home</td>
<td>chblake slightlyfamous</td>
<td>kormmandos singapornews</td>
</tr>
<tr>
<td>2</td>
<td>people world</td>
<td>ennn PatchouliW</td>
<td>mrbrown moby74</td>
</tr>
<tr>
<td></td>
<td>life word time</td>
<td>balaăiăutt PoohPiPi</td>
<td>kormmandos singapornews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FunksMokk Mon</td>
<td>AngMoGirl</td>
</tr>
<tr>
<td>3</td>
<td>time twitter</td>
<td>maynecic arsenav</td>
<td>mrbrown moby74</td>
</tr>
<tr>
<td></td>
<td>hope work friend</td>
<td>stuarttan balaăiăutt</td>
<td>singapornews singinfomap</td>
</tr>
<tr>
<td></td>
<td></td>
<td>derrickwaa</td>
<td>kormmandos</td>
</tr>
<tr>
<td>4</td>
<td>google design</td>
<td>balaăiăutt BoltClock</td>
<td>mrbrown moby74</td>
</tr>
<tr>
<td></td>
<td>twitter web site</td>
<td>fabrikade flashmech</td>
<td>kormmandos claudia10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>erwammance</td>
<td>AngMoGirl</td>
</tr>
<tr>
<td>5</td>
<td>love feel</td>
<td>highpriestess tstar</td>
<td>mrbrown moby74</td>
</tr>
<tr>
<td></td>
<td>eat hot hair</td>
<td>nikipaniki killerpussy</td>
<td>kormmandos hana77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>moby74</td>
<td>benkoe</td>
</tr>
</tbody>
</table>

5.2.2 Performance in Recommendation Task

Tangible benefit can be realized when applying it to some tasks. In this paper, we evaluate the usefulness of TwitterRank in the twitterer recommendation task. The recommendation task is designed as Figure 8 shows.

Step 4. Given this, the quality of the recommendation is measured as the number of twitterers in S_i who have a higher rank than s_f. More formally, it is defined as follows:

Definition 6. Assume L is a ranked list recommended by any of the algorithms, and s_i is a twitterer. Let l(s_i) be the rank of s_i in l (a higher rank corresponds to a low-numbered rank in l). The quality of the recommendation Q(l) is measured as Q(l) = |{s_i | s_i ∈ S_i, and l(s_i) < l(s_f)}|. s_f is the friend removed in Step 3 of Figure 8. The lower the value of Q(l) is, the higher the quality of corresponding algorithm is.

Different L’s based on various criteria have been used to study the proposed TwitterRank’s performance as comprehensively as possible. Currently, there are in total four criteria based on which L is generated:

(a): Two L’s denoted by L_{fh} and L_{fl} are generated based on the number of followers that s_f has: L_{fh} has s_f with high follower count, while L_{fl} has s_f with low follower count. s_f’s follower count is considered high if it is larger than FH, and low if smaller than FL. FH and FL are set as the 90th and 10th percentile of all the follower counts of the twitterers in S_i. To generate L_{fh} (or L_{fl}), |L| = 30 “following” relationships are chosen uniformly at random among all the existing relationships in which s_f fulfills the criteria described above.
difference in the performance of all the algorithms. Yet, InD achieves the best performance. This is probably because, in the dataset, tweeters’ “following” behaviors have already been biased toward those with more followers, since InD is essentially the algorithm applied in Twitter to recommend friends.

In scenarios where $L_{th}$ is used, TR’s performance is the worst among all. This is because the quality of topics distilled for $s_f$ is not as good since LDA-based topic distillation is less accurate with little content available. Consequently, this impacts the performance of TR which takes into account the topical similarity when measuring the tweeters’ influence.

In scenarios where $L_{dh}$ is used, TR outperforms all the other algorithms except InD. This phenomenon, together with the one observed in scenarios where $L_{fh}$ is used, shows that there still exist some tweeters who do not “follow” based on topical similarity, although homophily is observed.

TR performs the best in all the other scenarios, though the improvement is not significant in most of cases. It is noted that in scenarios where $L_{fh}$ is used, TR outperforms the other algorithms significantly, especially InD and PR. This is because friends of $s_f$ in the “following” relationships in $L_{fh}$ are with lower numbers of followers. Consequently, the corresponding $s_o$ would have lower chance to be biased by the recommendation made by Twitter, which is essentially made with InD. In such cases, the chance that the “following” relationship is formed due to topical similarity is higher. Therefore, TR outperforms InD and PR, which do not take into account topical similarity. Furthermore, TR outperforms TSPR. This is because TSPR uses identical transition probability matrix when calculating the topic-specific ranks. By doing so, TSPR basically propagates a tweeter’s influence in one topic to her friends in different topics with equal probabilities.

6. RELATED WORK

Currently, Twitter measures a tweeter’s influence as the number of followers she has. The more followers she has, the more impact she appears to make in the Twitter context, because she seems more popular. The underlying assumption here is that every tweet published by a tweeter is read by all her followers. A similar metric relies on the ratio between the number of one’s followers and the number of friends. Another metric proposed by the Web Ecology project [13] measures the influence based on the ratio of attention (including retweet, reply, and mention) a tweeter received to the tweets she published.

These three metrics do not utilize the global link structure among tweeters. There are attempts which take into account the global link structure when measuring influence in the Twitter context, e.g. TunkRank\footnote{TunkRank is originally proposed by Daniel Tunkelang in http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/. An implementation of the idea is available at http://tunkrank.com/}. TunkRank extends PageRank and calculates the influence of tweeter recursively as:

$$Influence(X) = \frac{1 + p \cdot \sum_{Y \in Followers(X)} Influence(Y)}{|Followers(Y)|}.$$

Here, $p$ is the constant probability that tweeters retweet a tweet. TunkRank measures tweeter $X$’s influence as the expected number of tweeters who will read a tweet that
she publishes. In this respect, TunkRank is similar to the proposed TwitterRank (see the first term in the RHS of Eq. (3)). However, TunkRank (and the above-mentioned three metrics as well) ignores the possibility for twitterers to interact with the content in Twitter.

The proposed TwitterRank acknowledges such possibility and extends PageRank with the consideration of topical similarity between twitterers. The most similar work in this aspect is Topic-sensitive PageRank (TSPR) proposed by Haveliwala [7]. It is also this work that the performance of TwitterRank is compared against. TSPR uses identical transition probability matrix when calculating the topic-specific influence. By doing so, TSPR basically propagates a twitterer’s influence in one topic to her friends in different topics with equal probabilities. In contrast, TwitterRank applies different transition probability matrices for different topics, which is validated by the experimental results to capture the topic-specific influence better.

7. CONCLUSIONS AND FUTURE WORK

This paper focuses on finding influential twitterers in Twitter. This paper is the first to report the phenomenon of homophily in a community of Twitter. By making use of this phenomenon, a Pagerank-like algorithm, called TwitterRank, is proposed to measure the topic-sensitive influence of the twitterers. The experimental results show that the proposed TwitterRank outperforms other related algorithms. Nevertheless, as an early attempt to bring order to Twitter, TwitterRank still has space for improvement.

First, as the experimental results show, there are still some twitterers “follow” not because of the topical similarity between them and their friends. We plan to classify different categories of twitterers by studying their “following” behaviors more closely, and apply TwitterRank on those with more serious “following” behaviors. Second, the current design of TwitterRank takes into account number of tweets a twitterer publishes (see Eq. (3)). This makes it susceptible to manipulations if a twitterer deliberately publishes a large number of tweets. In the future, we plan to improve this by incorporating other interactions between two twitterers, e.g., reply/mention between two twitterers. Third, we also plan to validate the “homophily” phenomenon and TwitterRank in a larger dataset. To collect a larger dataset, we are currently monitoring the Twitter public timeline using Twitter streaming API16. At the same time, we crawl the “following” relationship among twitterers using Twitter API17. Last but not least, currently the topic distillation is conducted on a snapshot of Twitter and the numbers of twitterers and topics are fixed. Nevertheless, Twitter is a platform for free and open conversations among twitterers. An incremental approach to topic distillation in Twitter is still a topic deserves further study.

8. REFERENCES


17 Twitter API: http://apiwiki.twitter.com/Twitter-API-Documentation.