School of
Information Systems

TwitterRank: Finding Topic-sensitive Influential Twitterers

Jianshu Weng, **Ee-Peng Lim, Jing Jiang**

Singapore Management University

Qi He

Pennsylvania State University

WSDM 2010

Feb 5, 2010

Outline

- Introduction
- Dataset
- Topic Modeling and Homophily among Twitterers
- TwitterRank
- Experiments and Results
- Conclusions and Future Work

Introduction

- Given a set of twitterers, find the influential ones
  - for different topics
- Why the problem?
  - Identify opinion leaders, experts
  - Viral marketing, advertisement
- Challenges:
  - The relationship among twitterers seems to be non-serious
  - Topics unknown
  - Evaluation without ground truth

Outline

- Introduction
- Dataset
- Topic Modeling and Homophily among Twitterers
- TwitterRank
- Experiments and Results
- Conclusions and Future Work

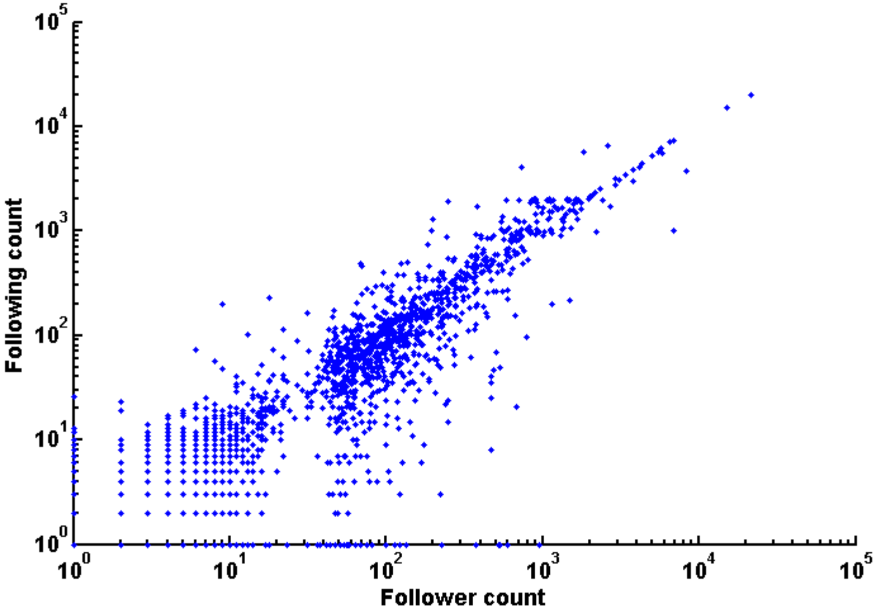
Data preparation

- Crawled $\mathcal{S}$ = a set of Singapore-based twitterers from twitterholic.com with highest number of followers.
- For each $s \in \mathcal{S}$, crawled its followers and friends $\bar{\mathcal{S}}$.
- $\mathcal{S}' = \mathcal{S} \cup \bar{\mathcal{S}}$ and $\mathcal{S}^* = \{s | s \in \mathcal{S}', \textit{ and } s \textit{ is from Singapore}\}$
- For each $s \in \mathcal{S}^*$, get its published tweets. Denote the set of all tweets as $\mathcal{T}$.

$\lvert\mathcal{S}\rvert$	996
$\lvert\mathcal{S}^*\rvert$	6748(4050 with more than 10 tweets)
$\lvert\mathcal{T}\rvert$	1,021,039
# following relationships	49,872
Min/Max/Avg #tweets/twitterer	1 / 3200 / 179.57

Reciprocity in the Following Relationships

- Friend count = # twitterers being followed
- Follower count = # twitterers following
- Correlation between friends count and follower count.
- 72.4% of the users follow more than 80% of their followers.
- 80.5% of the users have 80% of their friends follow them back.



Possible Explanations

- Two possible explanations:
  - “Following” relationship is too casual
  - Homophily, implying a stronger notion.
- Does homophily really exist?
  - Are twitterers with “following” relationships more similar than those without according to the topics they are interested in?
  - Are twitterers with reciprocal “following” relationships more similar than those without according to the topics they are interested in?

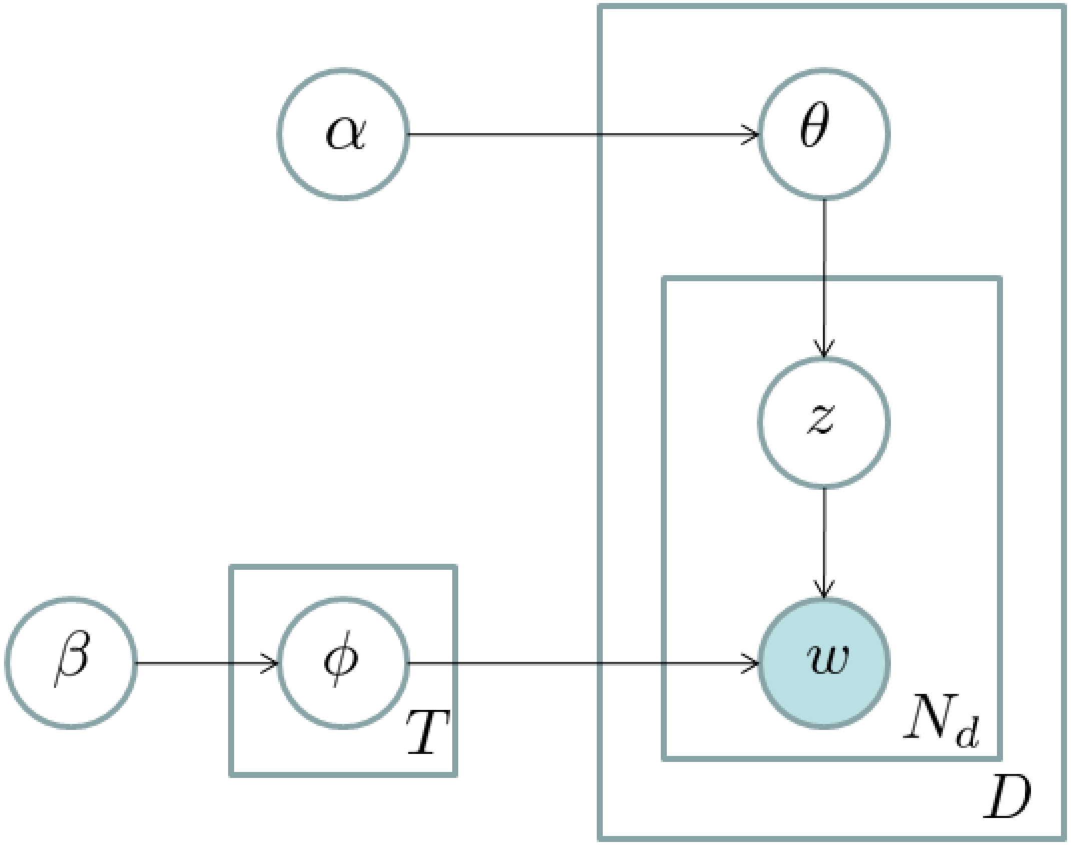
Outline

- Introduction
- Dataset
- Topic Modeling and Homophily among Twitterers
- TwitterRank
- Experiments and Results
- Conclusions and Future Work

Topic Distillation

- Apply LDA to distill topics automatically.
- Find topics in the twitterer's content to represent her interests
  - Twitterer's content = aggregated tweets
- Pre-processing
  - Use only those words without non-English characters
  - Min word length= 3
  - Remove
    - @userid
    - URL
    - All-digit word
    - Stopwords
  - Apply analysis on twitterers with more than 10 tweets. (#twitterer=4050)

LDA



Results of Topic Distillation

- Three matrices:
  - 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 tweets of twitterer s_i has been assigned to topic t_j .
  - 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
  - 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

Hypothesis testing (I)

- Are twitterers with “following” relationships more similar than those without according to the topics they are interested in?
- Topical difference= $\sqrt{2 * D_{JS}(i,j)}$
- μ_{follow} : Mean difference of the pairs with following relationships
- $\mu_{nofollow}$: Mean difference of the pairs without following relationships
- $H_0 : \mu_{follow} = \mu_{nofollow} \quad H_1 : \mu_{follow} < \mu_{nofollow}$
- The null hypothesis is rejected at $\alpha = 0.01$ for both twitterers with more than/less than 30 friends.

Hypothesis testing (II)

- Are twitterers with reciprocal “following” relationships more similar than those without according to the topics they are interested in?
- μ_{sym}: Mean difference of the pairs of users with reciprocal following relationships
- μ_{asym}: Mean difference of the pairs of users with only one-directional following relationships
- $H_0 : \mu_{sym} = \mu_{asym} \quad H_1 : \mu_{sym} < \mu_{asym}$
- The null hypothesis is rejected at $\alpha = 0.01$.

Implication

- Homophily phenomenon does exist.
  - Twitterers with "following" relationships are more similar than those without according to the topics they are interested in.
  - Twitterers with reciprocal "following" relationships are more similar than those without according to the topics they are interested in.
  - There are twitterers who are serious in following others.

Outline

- Introduction
- Dataset
- Topic Modeling and Homophily among Twitterers
- TwitterRank
- Experiments and Results
- Conclusions and Future Work

Topic-specific TwitterRank

- A topic-specific random walk model is applied to calculate each twitterer's influential score.
- The transition matrix for topic *t*, denoted as P_t. The transition probability of the random surfer from *follower* s_i to *friend* s_j *:*

$$P_t(i,j) = \frac{|\mathcal{T}_j|}{\sum_{a:\ s_i\ follows\ s_a} |\mathcal{T}_a|} * sim_t(i,j)$$

$$sim_t(i,j) = 1 - |DT'_{it} - DT'_{jt}|$$

- This captures two notions:
  - The more s_j publishes, the higher portion of tweets s_i reads is from s_j Generally, this leads to a higher influence on s_i.
  - $s_j^{\prime}s$ influence on s_i is also related to the topical similarity between the two as suggested by the *homophily* phenomenon.

Topic-specific TwitterRank (II)

- Topic-specific teleportation
  - $E_t = DT''_{\cdot t}$
- The influence scores of twitterers are calculated iteratively
  - $\overrightarrow{TR_t} = \gamma P_t \times \overrightarrow{TR_t} + (1-\gamma)E_t$

Aggregation of Topic-specific TwitterRank

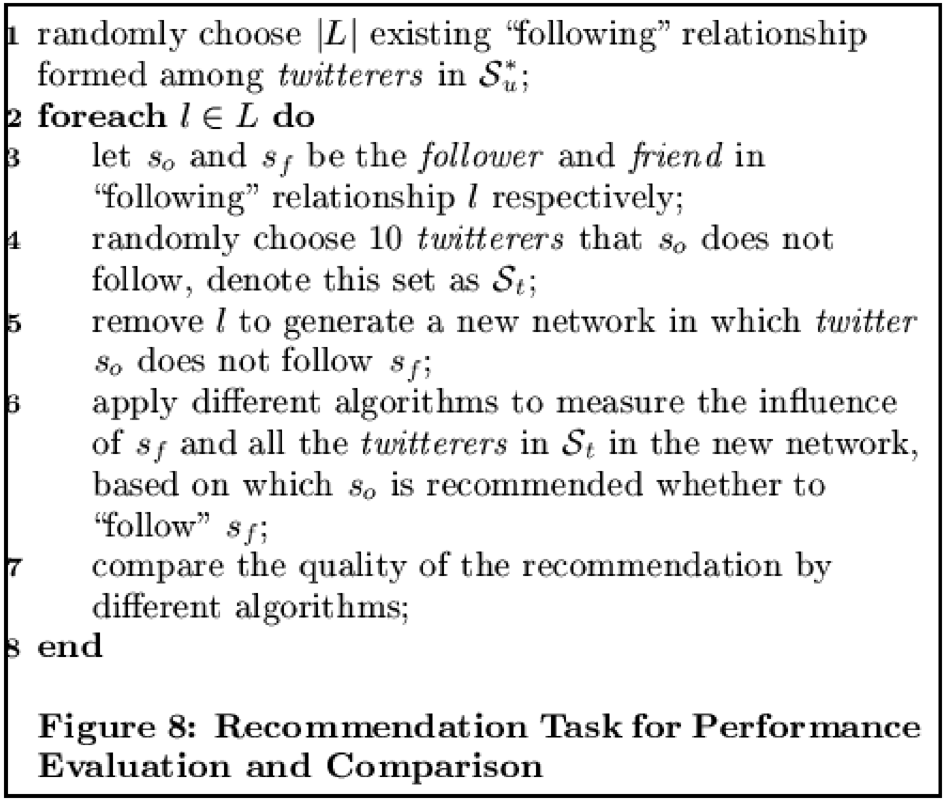
- $\overrightarrow{TR} = \sum_t r_t \cdot \overrightarrow{TR_t}$
- *General influence*: $r_t's$ can be set as the probabilities of different topics' presence
- *Perceived general influence*: $r_t's$ can also be set as the probabilities that a particular twitterer s_i is interested in different topics.

Outline

- Introduction
- Dataset
- Topic Modeling and Homophily among Twitterers
- TwitterRank
- Experiments and Results
- Conclusions and Future Work

Comparison with Other Algorithms

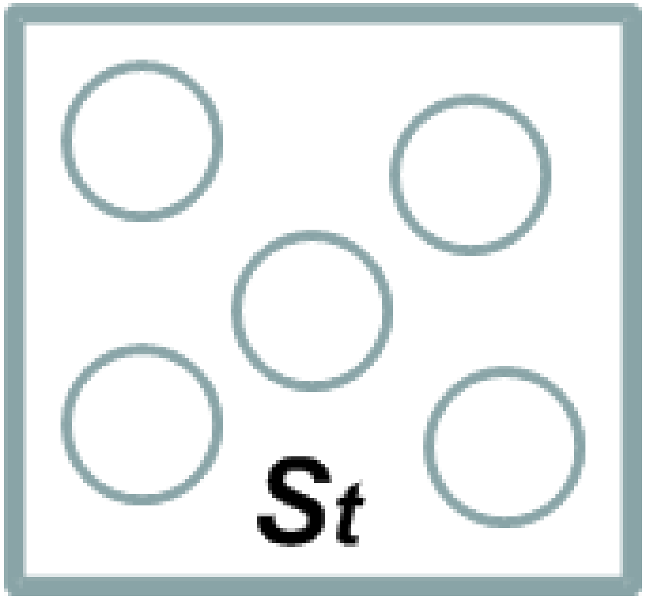
- Comparison of performance in a recommendation task. Set *L* is consider the ground truth.

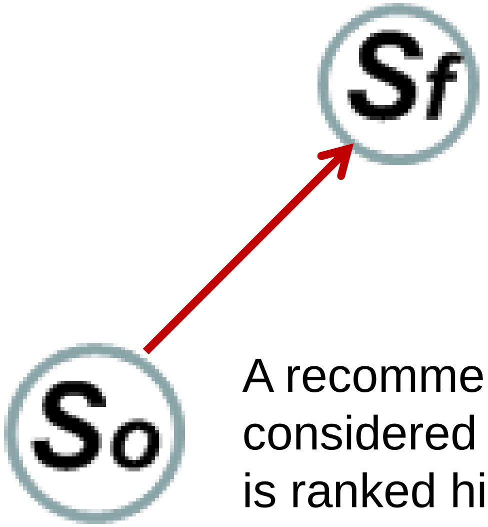


```
randomly choose $|L|$ existing "following" relationship
formed among twitterers in $\mathcal{S}_u^*$;
foreach $l \in L$ do
    let $s_o$ and $s_f$ be the follower and friend in
    "following" relationship $l$ respectively;
    randomly choose 10 twitterers that $s_o$ does not
    follow, denote this set as $\mathcal{S}_t$;
    remove $l$ to generate a new network in which twitter
    $s_o$ does not follow $s_f$;
    apply different algorithms to measure the influence
    of $s_f$ and all the twitterers in $\mathcal{S}_t$ in the new network,
    based on which $s_o$ is recommended whether to
    "follow" $s_f$;
    compare the quality of the recommendation by
    different algorithms;
end
```

Figure 8: Recommendation Task for Performance Evaluation and Comparison

The Recommendation Task



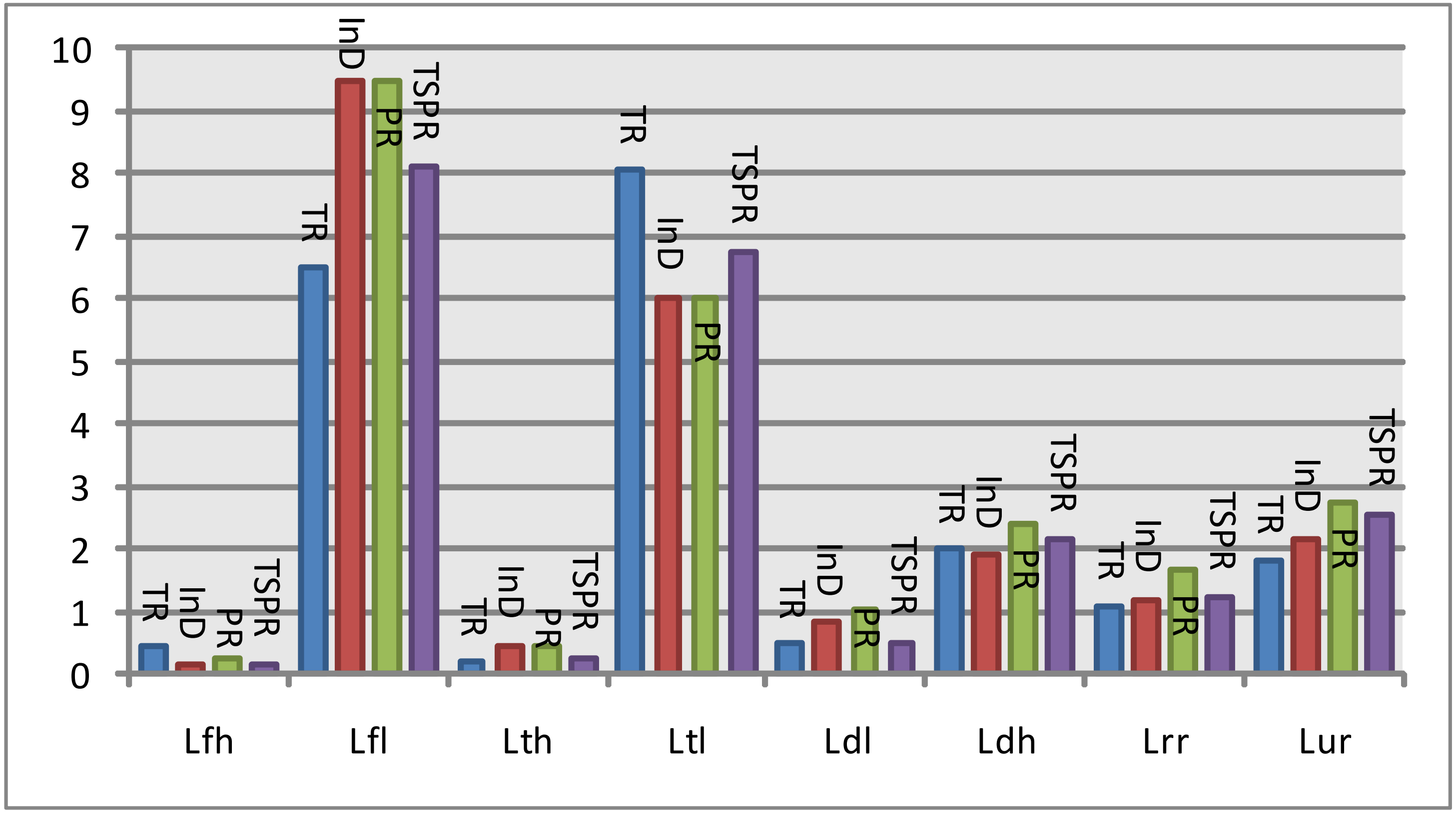


A recommendation is considered “good” if s_f is ranked higher than all the twitterers in S_t

Criteria to generate the *L* Set

- Number of followers that s_f has.
- Number of tweets that s_f published.
- Topical difference between s_f and s_o
- Whether reciprocal relationship exists among s_f and s_o

Comparison with Other Algorithms (III)



Major Observations (I)

- All performs better in L_{dl} than in L_{dh} :
  - There are twitterers who "follow" because of the topical similarity between them and their friends. This supports the phenomenon of *homophily*.
- TR is outperformed in L_{fh}, L_{tl}, and L_{dh}
  - InD performs the best in L_{fh} . This is probably because twitterers' "following" behaviors have already been biased toward those with more followers.

Major Observations (II)

- TR performs the worst in L_{tl}, because LDA-based topic distillation needs more contents to achieve reasonable accuracy
- TR outperforms all the other algorithms except InD in L_{dh}. There still exist some twitterers who do not “follow” based on topical similarity, although homophily is observed.

Outline

- Introduction
- Dataset
- Topic Modeling and Homophily among Twitterers
- TwitterRank
- Experiments and Results
- Conclusions and Future Work

Conclusions and Future Work

- Homophily does exist.
  - Not all users just randomly “follows”.
- Future work:
  - To make the algorithm more robust to manipulation, e.g purposely publish large number of tweets
  - To classify different categories of twitterers by studying their “following” behaviors more closely
  - Incremental topic distillation/event detection

Thank you