Tweet Ranking Based on Heterogeneous Networks

Hongzhao Huang, Arkaitz Zubiaga, Heng Ji
Computer Science Department and Linguistics Department
Queens College and Graduate Center
City University of New York
New York, NY, USA
hongzhaohuang@gmail.com

Hongbo Deng, Dong Wang, Hieu Le, Tarek Abdelzaher, Jiawei Han
Computer Science Department
University of Illinois at Urbana-Champaign
Urbana-Champaign, IL, USA

Alice Leung
Raytheon BBN Technologies
Cambridge, MA, USA

John Hancock
Artistech Inc.
Fairfax, VA, USA

Clare Voss
Multilingual Computing Branch
Army Research Lab
Adelphi, MD, USA

ABSTRACT

Ranking tweets is a fundamental task to make it easier to distill the vast amounts of information shared by users. In this paper, we explore the novel idea of ranking tweets on a topic using heterogeneous networks. We construct heterogeneous networks by harnessing cross-genre linkages between tweets and semantically-related web documents from formal genres, and inferring implicit links between tweets and users. To rank tweets effectively by capturing the semantics and importance of different linkages, we introduce Tri-HITS, a model to iteratively propagate ranking scores across heterogeneous networks. We show that integrating both formal genre and inferred social networks with tweet networks produces a higher-quality ranking than the tweet networks alone. [1]

Author Keywords

tweet ranking, heterogeneous networks, iterative propagation model

INTRODUCTION

Twitter has become a popular service for online communication through short messages of up to 140 characters, known as tweets. Its users produce millions of tweets a day, enabling both individuals and organizations to disseminate information about current affairs and breaking news in a timely fashion. This information is sometimes posted by users on-site or in the vicinity of events, providing first-hand accounts from a wide variety of sources. However, the sheer volume of tweets sent during events of general interest is overwhelming and hence difficult to distill for the most relevant information, while also filtering out non-informative tweets.

To facilitate finding informative and trustworthy content in tweets, it is crucial to develop an effective ranking method. This is particularly useful in emerging situations. Eyewitnesses might be live-tweeting about anything happening at ongoing events [7] such as natural disasters. To assist in these situations, we aim to develop a ranking system that organizes tweets by informativeness, so that informative tweets are readily identified, while pointless and speculative observations are filtered out. However, the definition of informativeness might vary for different points of view. Twitter users can produce diverse content ranging from news and events, to conversations and personal status updates. While personal updates and conversations might be relevant to a specific group of people, we aim to find tweets on topics that are informative to a general audience, such as breaking news and real-time coverage of on-going events. For example, during Hurricane Irene in 2011, updates from a user living in New York City about her own safety might be very informative to her friends and relatives, but not so informative to others. To produce rankings that are as relevant to as many people as possible, we define informativeness as the extent to which a tweet meets the general interest of people involved with or tracking the event.

While previous research has relied on either the text of tweets or explicit features of social network such as retweets, replies, and follower-followee relationships, we believe that such networks can be enhanced by integrating information from a formal genre. On one hand, tweets from different sources tend to contain non-informative noise such as subjective comments and conversations. Therefore it is challenging to identify salient information from tweet content alone. On the other hand, events of general interest such as natural disasters or political elections are the topics of tweets sent by many users from multiple communities which are not connected to each

[1] Related resources and software are freely available for research purposes at http://nlp.cs.qc.cuny.edu/tweetranking.zip; the system demo is at http://nlp.cs.qc.cuny.edu/tweet_summary/ground-truth-demo.xhtml.

other. In these situations, users are likely to be unaware of each other. As a result, they fail to connect with many others on topics of mutual interest. This lack of social interaction produces networks with few explicit linkages between users, and therefore between tweets and users. The sparsity of linkages would limit the effectiveness of features extracted from social network.

In this work, we introduce Tri-HITS, a novel propagation model that leverages global information iteratively computed across heterogeneous networks constructed from web documents, tweets, and users, to rank tweets on a topic by informativeness. The model addresses the two issues mentioned above (noisy tweets, limited social connections). Using Tri-HITS, we establish cross-genre linkages between tweets and web documents, filter informal writing and noise contained in tweets, and infer implicit tweet-user relations beyond the explicit ones, so that networks are enriched by connecting users that are sharing similar contents. We propose three high-level hypotheses that motivate the presented methods of constructing heterogeneous networks of tweets, users, and web documents. The proposed model, Tri-HITS, operates iteratively over all networks incorporating the semantics and importance of different linkages. By ranking tweets about the Hurricane Irene, we demonstrate that incorporating a formal genre such as web documents, inferring implicit social networks and performing effective ranking score propagation with the proposed model can significantly improve the ranking quality.

BACKGROUND

In this section, we describe the basic techniques used in the paper: information networks, the ranking approach TextRank, and a widely used method for redundancy removal.

Information Networks

We define an information network as a graph $G = (V, E)$ on $X = \{X_1, X_2, ..., X_Z\}$ for Z types of vertices, where $V(G) = X_1 \cup X_2 \cup ... \cup X_Z$ and $E(G) = \langle x_i, x_j \rangle$, for $x \in X$. An edge $\langle x_i, x_j \rangle$ is a binary relation between two vertices x_i and x_j. An information network is **heterogeneous** when the vertices are from multiple distinct types of sources ($Z \geq 2$). [5] defined a text-rich heterogeneous information network as an information network that integrates a set of text documents $D = \{d_1, d_2, ..., d_n\}$ with other types of vertices, so that $V(G) = D \cup X_1 \cup ... \cup X_{Z-1}$. In this work, we construct heterogeneous networks that include web documents, tweets, and users, as shown in Figure 1.

TextRank: Baseline Approach

Graph-based ranking algorithms have been widely used to generate rankings for vertices in graphs.. Adapted from PageRank [22] to weighted graphs, TextRank [20] is a well-known ranking algorithm for homogeneous networks, which is defined as follows:

$$s(v_i) = (1-d) + d * \sum_{v_j \in In(v_i)} \frac{w_{ji} s(v_j)}{\sum_{v_k \in Out(v_j)} w_{jk}} \quad (1)$$

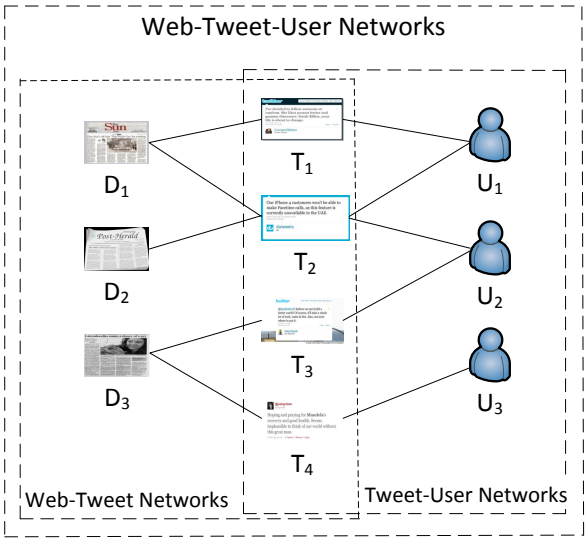

Figure 1. Web-Tweet-User heterogeneous networks

where v_i is a vertex with $s(v_i)$ as the ranking score, $In(v_i)$ as the set of incoming edges, and $Out(v_i)$ as the set of outgoing edges; w_{ij} is the weight for the edge between two vertices v_i and v_j. An edge exists between two vertices that represent text units when their computed shared content (cosine similarity) exceeds or equals a predefined threshold δ_{tt}.

Given its success when applied to sentence ranking for the task of extractive document summarization [19], we choose TextRank as the **baseline** method to compute ranking scores in tweet-only networks where edges between tweets are determined by their cosine similarity.

Redundancy Removal

Since users on Twitter can be tweeting similar information obliviously, and retweet and reply others' tweets, redundancy has been shown to be a pervasive phenomenon [32]. This issue has not been considered in previous works on tweet ranking [9, 14]. In this work, we perform a redundancy removal step to diversify top ranked tweets. To do so, we adopt the widely used greedy procedure [2, 18] to apply redundancy removal after the completion of each ranking method, as follows: tweet t_i in position i is removed when its cosine similarity with tweets $t_j \in [t_1, t_{i-1}]$ in more highly-ranked positions exceeds or equals a predefined threshold δ_{red}[2]

MOTIVATIONS AND HYPOTHESES

Next, we describe the motivational aspects and hypotheses in this work, which we aim to prove.

Hypothesis 1: *Informative tweets are more likely to be posted by credible users; and vice versa (credible users are more likely to post informative tweets).* [9, 14] consider that users who have more followers, mentions, and retweets, and are listed more, are more likely to be authoritative. They used retweet, reply, user mention and follower counts to to compute the degree of authoritativeness of users; and showed that

[2]We choose $\delta_{red} = 0.6$ as a threshold, obtained from our empirical studies with values from 0.1 to 1.0 in the development set.

user account authority is a helpful feature for tweet ranking. However, for events of general interest involving multiple communities, users are more likely to be unaware of each other, and rarely interact. This makes it insufficient to rely on user-user networks constructed from retweet and reply interactions to compute user credibility scores. To overcome this problem, we apply a Bayesian approach to compute the credibility of users by incorporating the contents shared by them.

Hypothesis 2: *Tweets involving many users are more likely to be informative.* Having many users share similar tweets at the same time helps identify informative tweets. For example, in the context of Hurricane Irene, users were likely to share information about the Evacuation Zone when they found relevant news or events. The synchronization of information within groups has been successfully harnessed in other fields like financial trading, autonomous swarms of exploratory robots, and flocks of communicating software agents [4, 23]. This idea has also been successfully exploited for event summarization from tweets [33].

Hypothesis 3: *Tweets aligned with contents of web documents are more likely to be informative.* Tweets come from diverse sources, and can diverse content ranging from news and events, to conversations and personal status updates. Therefore, informative tweets tend to be interspersed with noisy and non-informative tweets. This differs from formal genres such as web documents, which tend to be cleaner. In the case of current events such as natural disasters or political elections, there are tight correlations between social media and web documents. Important information shared in social media tends to be posted in web documents. For example, the following informative tweets would rank highly because they are linked to informative web documents: " *New Yorkers, find your exact evacuation zone by your address here: http://t.co/9NhiGKG /via @user #Irene #hurricane #NY*" and "*Details of Aer Lingus flights affected by Hurricane Irene can be found at http://t.co/PCqE74V201d*". As far as we know, this is the first work to integrate information from a formal genre such as web documents to enhance tweet ranking.

ENHANCED APPROACH: TRI-HITS

Based on the formulated hypotheses, we describe how Tri-HITS works.

Overview

Figure 2 depicts how Tri-HITS works. For a set of tweets on a specific topic, a rule-based filtering component is first applied to filter out a subset of non-informative tweets. For the remaining tweets, we define queries based on top terms in tweets, and use Bing Search API[3] to retrieve the titles[4] of the top m web documents for those queries ($m = 2$ for these experiments). Then we apply TextRank and a Bayesian approach that initialize ranking scores for tweets, web documents, and users. Finally, we iteratively propagate ranking scores for web documents, tweets, and users across the networks to refine the tweet ranking.

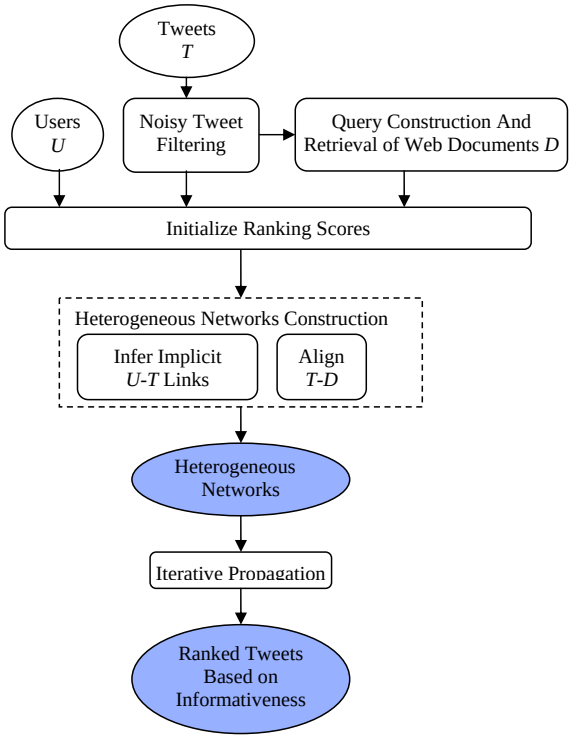

Figure 2. Overview of Tri-HITS

Filtering non-informative Tweets

Tweets are more likely to be shortened or informally written than texts from a formal genre such as web documents. Thus, a prior filtering step would clean up the set of tweets and improve the ranking quality. We observed that numerous non-informative tweets have some common characteristics, which help infer patterns to clean up the set of tweets. In our filtering method, we define several patterns to capture the characteristics of a non-informative tweet, i.e., very short tweets without a complementary URL, tweets with first personal pronouns, or informal tweets containing slang words[5]. These features have been shown to be effective in previous work on tweet ranking and information credibility [9, 3, 27]. Our filtering component accurately filters out non-informative tweets, achieving 96.59% at precision.

Initializing Ranking Scores

Initializing scores for tweets and web documents. For a set of tweets T, we first construct an undirected and weighted graph $G = (V, E)$. After removing stopwords and punctuations, the bag-of-words of each tweet t_i is represented as a vertex $v_i \in V$, and the weight for the edge between tweets is the cosine similarity using TF-IDF representations. Then, we use TextRank to compute initial scores. The same approach is used to initialize ranking scores for web documents.

[3] http://www.bing.com/toolbox/bingdeveloper/

[4] We rely on page titles, but it could be extended to the whole content of web documents straightforwardly.

[5] http://www.mltcreative.com/blog/bid/54272/Social-Media-Minute-Big-A-List-of-Twitter-Slang-and-Definitions

Initializing user credibility scores. Based on Hypothesis 1, we define two approaches to compute initial user credibility scores. First, we construct a user network based on retweets, replies and user mentions as in [9]. This results in a directed and weighted graph $G_d = (V, E)$, where V is the set of users and E is the set of directed edges. A directed edge exists from u_i to u_j if user u_i interacts with u_j (i.e., mentions, retweets, or replies to u_j). The weight of the edge is defined as N_{ij}, according to the number of interactions. In this case, we use TextRank to compute initial user credibility scores.

In addition, we also use the *Bayesian ranking* approach [28, 29] that considers the credibility scores of tweets and users simultaneously based on Tweet-User networks. Given a set of users $U = \{u_1, u_2, ..., u_m\}$, and a set of claims $C = \{c_1, c_2, ..., c_n\}$ the users make (each claim corresponds to a cluster of tweets in this paper). We also define matrix W^{cu} where $w_{ji}^{cu} = 1$ if user u_i makes claim c_j, and is zero otherwise. Let u_i^t denote the proposition that *'user u_i speaks the truth'*. Let c_j^t denote the proposition that *'claim c_j is true'*. Also, let $P(u_i^t)$ and $P(u_i^t|W^{cu})$ be the prior and posterior probability that user u_i speaks the truth. Similarly, $P(c_i^t)$ and $P(c_i^t|W^{cu})$ are the prior and posterior probability that claim c_i is true. We define the credibility rank of a claim $Rank(c_j)$ as the increase in the posterior probability that a claim is true, normalized by prior probability $P(c_i^t)$. Similarly, the credibility rank of a user $Rank(u_i)$ is defined as the increase in the posterior probability that a user is credible, normalized by prior probability $P(u_i^t)$. In other words, we can get:

$$Rank(c_j) = \frac{P(c_j^t|W^{cu}) - P(c_j^t)}{P(c_j^t)} \quad (2)$$

$$Rank(u_i) = \frac{P(u_i^t|W^{cu}) - P(u_i^t)}{P(u_i^t)} \quad (3)$$

In our previous work, we showed that the following relations hold true regarding the credibility rank of a claim $Rank(c_j)$ and a user $Rank(u_i)$:

$$Rank(c_j) = \sum_{k \in Users_j} Rank(u_k) \quad (4)$$

$$Rank(u_i) = \sum_{k \in Claims_i} Rank(c_k) \quad (5)$$

where $Users_j$ is the set of users makes claim c_j, and $Claims_i$ is the set of claims the user u_i makes. From the above, the credibility of sources and claims can be derived as:

$$P(c_j^t|W^{cu}) = p_a^t(Rank(c_j) + 1) \quad (6)$$

$$P(u_i^t|W^{cu}) = p_s^t(Rank(u_i) + 1) \quad (7)$$

where p_a^t and p_s^t are initialization constants, which are the ratio of true claims to the total claims, and the ratio of credible users to the total users.

Then, Equation 7 is used to compute initial user credibility scores as our second approach.

Constructing Heterogeneous Networks

Next, we describe the two types of networks we build as constituent parts of heterogeneous networks:

Tweet-User networks. Based on Hypothesis 2, we expand the Tweet-User networks by inferring implicit tweet-user relations. If a user u_i posted a set of tweets T_i during a period of time, we say an implicit relation exists between u_i and a tweet t_j if the maximum cosine similarity between t_j and $t_i \in T_i$ exceeds or equals a threshold δ_{tu}.

Web-Tweet networks. Given a set of tweets T and a set of associated web documents D, we build a bipartite graph $G = T \cup D, E$, where an undirected edge with weight w_{ij}^{td} is added when the cosine similarity between $t_i \in T$ and $d_j \in D$ exceeds or equals δ_{td}. This approach creates cross-genre linkages between tweets and web documents on similar events (e.g., evacuation events).

In subsection *Effect of Parameters*, we will discuss the effects of parameters δ_{td} and δ_{tu}.

Iterative Propagation

We introduce a novel algorithm to incorporate both initial ranking scores and global evidence from heterogeneous networks. It propagates ranking scores across heterogeneous networks iteratively. Our algorithm is an extension of Co-HITS [6], which is limited to bipartite graphs. Co-HITS was designed to incorporate links of a bipartite graph with content from two types of objects. The intuition behind the score propagation is the mutual reinforcement to boost co-linked objects.

Let $G = (U \cup V, E)$ be a bipartite graph, in which the vertices are divided into two disjoint sets U and V, and each edge in E connects one vertex in U to another in V. We use w_{ij}^{uv} (or w_{ji}^{vu}) to denote the weight for the edge between u_i and v_j. To put all the weights between sets U and V together, we can use $W^{uv} \in \mathbb{R}^{|U|\times|V|}$ (or $W^{vu} \in \mathbb{R}^{|V|\times|U|}$) to denote the weight matrix between U and V. Note that $W^{uv} \in \mathbb{R}^{|U|\times|V|}$ is the transpose of $W^{vu} \in \mathbb{R}^{|V|\times|U|}$ as we have $w_{ij}^{uv} = w_{ji}^{vu}$. For each $u_i \in U$, a transition probability p_{ij}^{uv} is defined as the probability that vertex u_i in U reaches vertex v_j in V at the next step. Formally, it is defined as a normalized weight $p_{ij}^{uv} = \frac{w_{ij}^{uv}}{\sum_k w_{ik}^{uv}}$, such that $\sum_{j \in V} p_{ij}^{uv} = 1$. Similarly, we obtain the transition probability $p_{ji}^{vu} = \frac{w_{ji}^{vu}}{\sum_k w_{jk}^{vu}}$ and $\sum_{i \in U} p_{ji}^{vu} = 1$ for each $v_j \in V$. The Co-HITS algorithm is defined as follows:

$$s(u_i) = (1 - \lambda_u)s^0(u_i) + \lambda_u \sum_{j \in V} p_{ji}^{vu} s(v_j), \quad (8)$$

$$s(v_j) = (1 - \lambda_v)s^0(v_j) + \lambda_v \sum_{i \in U} p_{ij}^{uv} s(u_i), \quad (9)$$

where $\lambda_u \in [0, 1]$ and $\lambda_v \in [0, 1]$ are personalized parameters, $s^0(u_i)$ and $s^0(v_j)$ are initial ranking scores for u_i and v_j, and $s(u_i)$ and $s(v_j)$ denote updated ranking scores of vertices u_i and v_j. In this algorithm, the initial scores are normalized

to $\sum_{i\in U} s^0(u_i) = 1$ and $\sum_{j\in V} s^0(v_i) = 1$, and the sum of updated $s(u_i)$ and $s(v_j)$ will be 1 as well.

The problem with Co-HITS in our experimental settings is the transition probability. As mentioned before, we choose cosine similarity as the weight for the edge between two vertices, and a similarity matrix W is obtained to denote the weight matrix where each entry w_{ij} is the similarity between vertex u_i and vertex v_j. Although the transition probability is a natural normalization for the weight between two vertices, it may not be suitable for similarity matrix. The reason is that the original similarity between different objects has already been normalized, so a further normalization from the similarity matrix to transition matrix may weaken or damage inherent meanings of the original similarity. For example, if a tweet u_i is aligned with one and only one document v_j with relatively low similarity weight, the transition probability w_{ij}^{uv} will be increased to 1 after normalization. Similarly, some higher similarity weights may be normalized to small transition probabilities.

By extending and adapting Co-HITS, we develop Tri-HITS to handle heterogeneous networks with three types of objects: users, tweets and web documents. Given the similarity matrices W^{dt} (between documents and tweets) and W^{tu} (between tweets and users), and initial ranking scores of $s^0(d)$, $s^0(t)$ and $s^0(u)$, we aim to refine the initial ranking scores and obtain the final ranking scores $s(d)$, $s(t)$ and $s(u)$. Starting from document $s(d)$, the update process considers both the initial score $s^0(d)$ and the propagation from connected tweets $s(t)$, which can be expressed as:

$$\begin{aligned} \hat{s}(d_i) &= \sum_{j\in T} w_{ji}^{td} s(t_j), \\ s(d_i) &= (1-\lambda_{td}) s^0(d_i) + \lambda_{td} \frac{\hat{s}(d_i)}{\sum_i \hat{s}(d_i)}, \end{aligned} \quad (10)$$

where W^{td} is the transpose of W^{dt}, and $\lambda_{td} \in [0,1]$ is the parameter to balance between initial and propagated ranking scores. Tri-HITS normalizes the propagated ranking scores $\hat{s}(d_i)$, while Co-HITS propagates normalized ranking scores by using the transition matrix instead of the original similarity matrix, potentially weakening or damaging the inherent meanings of the original similarity. Similarly, we define the propagation from tweets to users as:

$$\begin{aligned} \hat{s}(u_k) &= \sum_{j\in T} w_{jk}^{tu} s(t_j), \\ s(u_k) &= (1-\lambda_{tu}) s^0(u_k) + \lambda_{tu} \frac{\hat{s}(u_k)}{\sum_k \hat{s}(u_k)}, \end{aligned} \quad (11)$$

Each tweet $s(t_j)$ may be influenced by the propagation from both documents and users:

$$\begin{aligned} \hat{s}_d(t_j) &= \sum_{i\in D} w_{ij}^{dt} s(d_i), \\ \hat{s}_u(t_j) &= \sum_{k\in U} w_{kj}^{ut} s(u_k), \\ s(t_j) &= (1-\lambda_{dt}-\lambda_{ut}) s^0(t_j) \\ &\quad + \lambda_{dt} \frac{\hat{s}_d(t_j)}{\sum_j \hat{s}_d(t_j)} + \lambda_{ut} \frac{\hat{s}_u(t_j)}{\sum_j \hat{s}_u(t_j)}. \end{aligned} \quad (12)$$

where W^{ut} is the transpose of W^{tu}, λ_{dt} and λ_{ut} are parameters to balance between initial and propagated ranking scores. The λ variables define the networks being considered: (i) when λ_{dt} is set to 0, only Tweet-User networks are considered (Method 3 in Table 1); (ii) when λ_{ut} is set to 0, only Web-Tweet networks are considered (Method 4); (iii) when both λ_{dt} and λ_{ut} are different from 0, the entire heterogeneous Web-Tweet-User network is considered (Method 5). For methods relying on bipartite graphs, we define as *one-step propagation* when the propagation is performed in a single direction, while we call it *two-step propagation* when it is performed in both directions. The selection of one-step propagation and two-step propagation is controlled by λ parameters.

Model Convergence Proof: From Equation (10), and assuming $\lambda_{td} > 0$ (the ranking scores $s(d)$ for web documents would not change if $\lambda_{td} = 0$), we get:

$$\bar{s}(d_i) = \frac{1}{\lambda_{td}} [s(d_i) - (1-\lambda_{td}) s^0(d_i)] = \frac{\hat{s}(d_i)}{\sum_i \hat{s}(d_i)}. \quad (13)$$

$\bar{s}(d)$, the normalized score of $\hat{s}(d)$, is similar to the normalized authority or hub scores defined in HITS [17], the difference being only the function to select vector norms. Kleinberg proved that $\bar{s}(d_i)$ converges as the iterative procedure continues, from which the convergence of the ranking scores $s(d)$ for web documents is guaranteed. The same assumption proves the convergence of ranking scores for tweets and users.

Algorithm 1 summarizes Tri-HITS.

EXPERIMENTS

Next, we present the experiment settings and analyze the methods shown in Table 1.

Data

We use tweets on the Hurricane Irene from August 26 to September 2, 2011 for our experiments. Using the query terms *hurricane* or *irene* to monitor tweets, we collected 176,014 tweets posted by 139,136 users within that timeframe. For evaluation purposes, we segment the tweets into 153 hours with an average of 1,150 tweets in each hour.

We randomly chose tweets from three hours to be manually annotated as our reference. This subset contains 3,460 tweets posted on different days: August 27, 2011, August 28, 2011 and September 1, 2011. Following the annotation guidelines defined by [14], two annotators parallelly assigned each

Methods	Descriptions	Hypotheses
1. Baseline	TextRank based on tweet-tweet networks.	
2. 1+Filtering	Baseline with filtering included.	
3. 2+Tweet-User*	Propagation on explicit and implicit Tweet-User networks.	1 and 2
4. 2+Web-Tweet	Propagation on Web-Tweet networks.	3
5. 3+4 Web-Tweet-User*	Propagation on Web-Tweet-User networks.	all

Table 1. Description of methods (method with * make use of the Bayesian Approach to initialize user credibility scores.

Input: A set of tweets (T), and users (U) on a given topic.
Output: Ranking scores (S_t) for T.

1: Use rule-based method to filter out noisy tweets (remaining $\hat{T}$ posted by users $\hat{U}$);
2: Retrieve relevant web documents D for $\hat{T}$;
3: Use TextRank and Bayesian Ranking to compute initial ranking scores S_t^0 for $\hat{T}$, S_d^0 for D and initial credibility scores S_u^0 for $\hat{U}$;
4: Construct heterogeneous networks across $\hat{T}$, $\hat{U}$ and D;
5: $k \leftarrow 0$, $diff \leftarrow 10e6$;
6: **while** $k <$ MaxIteration and $diff >$ MinThreshold **do**
7: Use Eq. (12) to compute S_t^{k+1};
8: Use Eq. (11) to compute S_u^{k+1};
9: Use Eq. (10) to compute S_d^{k+1};
10: Normalize S_t^{k+1},S_d^{k+1}, and S_u^{k+1};
11: $diff \leftarrow \sum(|S_t^{k+1} - S_t^k|)$;
12: $k \leftarrow k + 1$
13: **end while**

Algorithm 1: Tri-HITS: Tweet ranking using heterogeneous networks

Grade	5	4	3	2	1
Hour 1	65	48	93	119	847
Hour 2	135	159	255	164	458
Hour 3	129	102	162	123	602

Table 2. Tweet distribution by grade

tweet a grade in a 5-star likert scale. Tweets with grade 5 are the most informative, while tweets with label 1 are the least informative. When the label difference between annotators was 1, the lower grade was selected. When the label difference was greater than 1, those tweets were re-annotated until the label difference did not exceed 1. Table 2 shows the distributions of all grades for each of the three hours of tweets.

Evaluation Metric

To evaluate tweet ranking, we rely on three-fold cross validation using *nDCG* as a measure [16], which considers both the informativeness, and the position of a tweet:

$$nDCG(\Phi, k) = \frac{1}{|\Phi|} \sum_{i=1}^{|\Phi|} \frac{DCG_{ik}}{IDCG_{ik}},$$

$$DCG_{ik} = \sum_{j=1}^{k} \frac{2^{rel_{ij}} - 1}{\log(1+j)},$$

where Φ is the set of documents in the test set, each document corresponding to an hour of tweets in our case, rel_{ij} is the human-annotated label for the tweet j in the document i, and $IDCG_{ik}$ is the DCG score for the ideal ranking. The average *nDCG* score for the top k tweets is: $Avg@k = \sum_{i=1}^{k} nDCG(\Phi, i)/k$. To favor diversity of top ranked tweets, redundant tweets are penalized to lower down the final score.

Effect of Parameters

We study the impact of different parameters on the training set. We present the most representative figures to show the effect, due to the lack of space. For TextRank, we explore δ_{tt} values from 0 to 1. For the enhanced approaches, we firstly perform one-step propagation of ranking scores from web documents to tweets by considering all pairs of δ_{td} and λ_{dt} from 0 to 1 with a step of 0.1. For each δ_{td}, the corresponding λ_{dt} and the best average nDCG scores for top 10 and 100 tweets are shown in Figure 3(a). We notice that when both initial tweet ranking scores and propagated ranking scores from web documents are considered (i.e., δ_{td} is set from 0 to 0.9 and $\lambda_{dt} > 0$), the ranking quality outperforms that by simply considering initial ranking scores of tweets (i.e. $\delta_{td} = 1$). Secondly, for the ranking performance of double-step ranking scores propagation, we choose to set $\delta_{td} = 0.1$, $\lambda_{dt} = 0.4$ and test λ_{td} from 0 to 1. Figure 3(b) shows an encouraging improvement in the ranking quality, and more stable over the baseline and one-step propagation. This suggests that two-step propagation provides mutual improvement in the ranking quality. The reason is that the ranking of web documents may also be refined using tweet and user evidence thanks to the large volume and synchrony of tweeting [32]. Here, $\lambda_{td} = 0.2$ yields the best performance. The aforementioned process is followed for Tweet-User networks, finding the best performance for $\delta_{tu} = 0.1$, $\lambda_{ut} = 0.2$, and $\lambda_{tu} = 0.6$.

When validating on the test set, Method 4 based on Web-Tweet networks outperforms Method 3 relying on Tweet-User networks. Therefore, for Web-Tweet-User networks, we keep the above values, and explore λ_{ut} values from 0 to 0.6 (e.g., $1 - \lambda_{dt}$). Figure 3(c) shows that integrating web documents, tweets and users, the ranking quality improves over both Web-Tweet networks and Tweet-User networks.

Performance and Analysis

Figure 4 shows the performance of ranking methods. The performance gain from Method 1 to Method 2 shows the need of filtering short and informal tweets. In this case, filtering

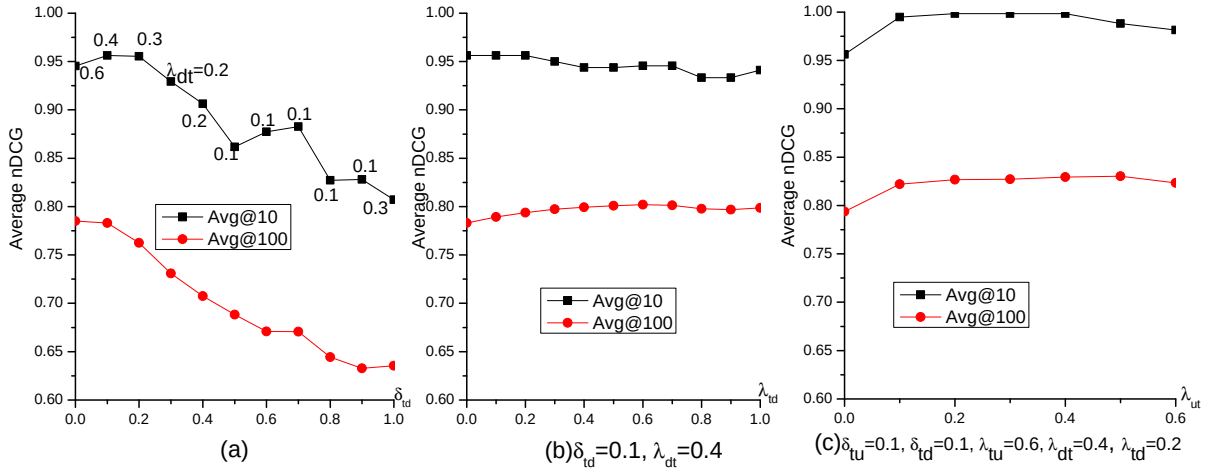

Figure 3. Effect of parameters: (a) δ_{td} and λ_{dt} for Web-Tweet networks, (b) λ_{td} for Web-Tweet networks, (c) λ_{dt} for Web-Tweet-User networks.

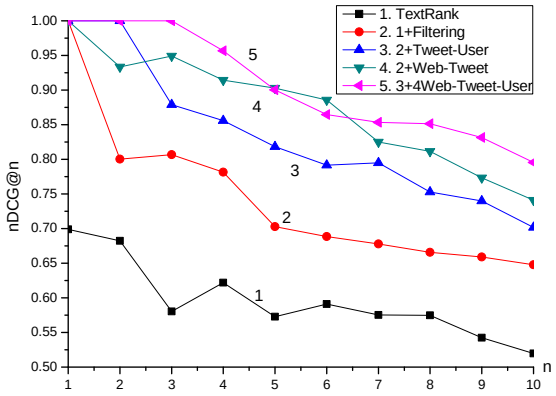

Figure 4. Performance comparison of ranking methods

Grade	5	4	3	2	1
Percentage	0.11%	0.17%	3.13%	4.84%	91.75%

Table 3. Grade distributions for filtered tweets.

reduced from 3,460 to 1,765 tweets ($\sim$ 49% reduction). Table 3 shows the distribution of labels for filtered tweets: a great majority of 91.75% had been annotated as 1, while only 0.11% had been annotated as 5.

Methods 3, 4 and 5, which integrate heterogeneous networks after filtering, outperform the baseline TextRank. When tweets are aligned with web documents (Method 4), the ranking quality improves significantly, proving that web documents can help infer informative tweets adding support from a formal genre. The fact that tweets with low initial ranking scores are aligned with web documents helps improve their ranking positions (Hypothesis 3). For example, the ranking of the tweet "*Hurricane Irene: City by City Forecasts http://t.co/x1t122A*" is improved compared to TextRank, helped by the fact that 10 retrieved web documents are about this topic.

Integrating users (Method 5) further improves performance. This indicates that Web-Tweet and Tweet-User networks may complement each other in improving ranking. For example, the tweet "*A social-media guide to dealing with Hurricane Irene http://t.co/0XBEnEJ*" is not top-ranked when only using Web-Tweet networks, since none of the retrieved web documents is related to it. However, similar tweets appear with high frequency in the tweet set. Hence, inferring implicit tweet-user relations and propagating information through the tweet-user network also improves the ranking.

Figure 5(a) shows that inferring implicit tweet-user relationships outperforms the only use of explicit tweet-user relations, especially for top positions. Looking into lower positions, we find that the redundancy removal performs better for the only use of explicit relations. However, both approaches can still perform similarly in positions $5 \sim 10$. This corroborates the synchronous behavior of users as an indicator of informative contents (Hypothesis 2). Since it is likely that a large set of users only tweet once within a short timeframe, limiting to explicit tweet-user relations results in sparse links, and ranking quality cannot be bootstrapped. Interestingly, inferring implicit tweet-user relations can capture synchronous behavior of users, which indicates subjects that users are concerned about.

Figure 5(b) shows that initializing user credibility scores with the Bayesian approach and performing one-step ranking score propagation from users to tweets based on the explicit tweet-user networks also outperforms TextRank. This corroborates our hypothesis that credible users are more likely to post informative tweets (Hypothesis 1). In addition, using only retweets, replies, and user mentions to compute initial user ranking scores, the performance does not improve over TextRank. The reason is that for an event of general interest like the Hurricane Irene, users from different communities rarely interact with each other.

Finally, Figure 6 shows that Tri-HITS significantly outperforms Co-HITS over bipartite graphs, with the only exception of position $n = 2$ for the Web-Tweet network. This corroborates that normalizing the similarity matrix weakens

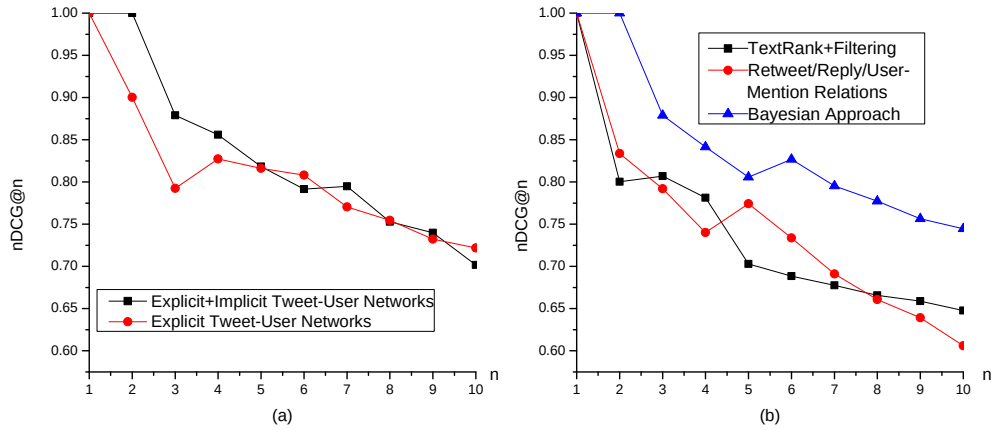

Figure 5. (a) Explicit vs Inferred Implicit Tweet-User Relations to Construct Tweet-User Networks; (b) TextRank vs One-step Propagation on Explicit Tweet-User Networks Using Bayesian Approach and Retweet/Reply/User Mention Relations.

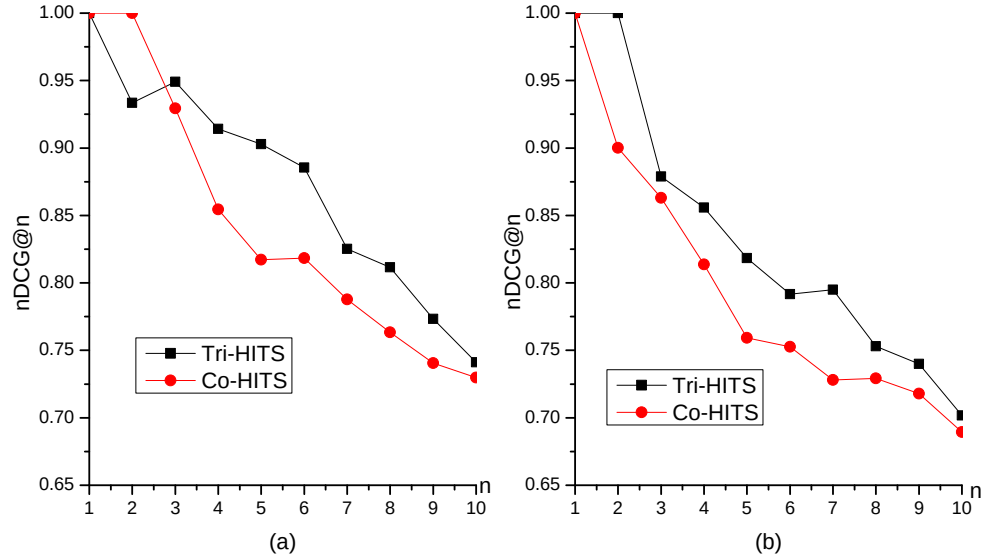

Figure 6. Co-HITS vs Tri-HITS on (a) Web-Tweet Networks, (b) Tweet-User Networks

semantic relations between different objects, and that capturing inherent meanings of cross-genre linkages is crucial for information propagation.

Remaining Error Analysis

Tri-HITS shows encouraging improvements in ranking quality with respect to a state-of-the-art model like TextRank. However, there are still some issues to be addressed for further improvements.

(i) Topically-relevant tweet identification. We tracked tweets containing the keywords "Hurricane" and "Irene". Using such a query to follow tweets might also return tweets that are not related to the event being followed. This may occur either because the terms are ambiguous, or because of spam being injected into trending conversations to make it visible. For example, the tweet "*Hurricane Kitty: http://t.co/cdIexE3*" is an advertisement, which is not topically related to Irene.

(ii) Non-informative tweet identification. Our rule-based filtering component achieves high precision (96.59%) on the identification of non-informative tweets, while there are still a number of false positives with a 70.7% recall. Performing deeper linguistic analysis, such as exploring subjectivity, might help clean up the tweet set by identifying additional non-informative tweets. For example, an analysis of writing styles would help identify the tweet "*Hurricane names hurricane names http://t.co/iisc7UY ;)*" as informal because it contains repeated phrases. And the tweet "*My favorite parts of Hurricane coverage is when the weathercasters stand in those 100 MPH winds right on the beach. Good stuff.*" is clearly subjective commentary that may entertain but will not meet the general interest of people involved with or tracking the event.

(iii) Deep semantic analysis of the content. Users may rely on distinct terms to refer to the same concept. More extensive semantic analyses of text could help identify those terms, possibly enhancing the propagation process. For example,

information extraction tools can be used to extract entities and events, and their coreferential relations, such as "NYC" and "New York City", or "MTA closed" and "subway shutting down". Likewise, existing dictionaries such as WordNet [21] can be utilized to mine synonym/hypernym/hyponym relations, and Brown clusters [1] can be explored to mine other types of relations.

RELATED WORK

We discuss relevant research on tweet ranking, information credibility for tweets, and the use of graphical models.

Previous research on tweet ranking has relied on the analysis of content [24], user credibility [10, 30, 31, 13, 27] and URL availability, or combinations of them [9, 14]. In addition, [14] also exploited content similarity to propagate evidence within the tweet genre. Most work has been based on supervised learning models such as RankSVM, Naive-Bayes classifier, and Linear Regression. [15] compared various unsupervised methods to rank tweets for summarization purposes, but only used lexical-level content analysis features.

In analyzing the information credibility of tweets, [3] relied on various levels of features (i.e., message-based, user-based, topic-based and propagation-based features) and supervised learning models for information credibility assessment in Twitter, which [12] extended by capturing relations among events, tweets, and users. [28, 29] proposed a Bayesian interpretation to assess tweet credibility. However, it remains as a preliminary approach due to the linear assumption made in the iterative algorithm of the basic fact-finding scheme. Intensive research has also been conducted on information credibility analysis (cf. [11]).

Graphical models have been effectively used in document summarization [19, 26, 25] demonstrating their power of propagating information across linked instances. However, most of these models, such as TextRank [20], as originally developed apply only to homogeneous networks. In contrast to existing research, we introduce Tri-HITS, a novel method that incorporates evidence from multiple genres, by exploiting semantically-related links to external web documents and inferring the implicit tweet-user relations. Following a different method for linking tweets and web documents, [8] used outgoing links from tweets to improve recency ranking for a search engine.

CONCLUSIONS AND FUTURE WORK

We have introduced Tri-HITS, a novel propagation model that makes use of heterogeneous networks composed of tweets, users, and web documents to rank tweets. To the best of our knowledge, this is the first approach to integrating tweets with formal genres that improves tweet ranking quality. Using propagation models to define ranking scores, we have shown that information from the formal genre of web documents can help improve the ranking quality. By introducing this new propagation model, studying the integration of different genres, presenting a way of inferring implicit tweet-user relations, and exploring the impact of parameters, this work sheds light on the challenging task of ranking tweets that are written informally by a diverse community of users.

Our next step is to develop metrics to predict ranking confidence so that we can remove low-confidence results and outliers from the evidence propagation. In addition, ranking tweets (and later, news) by their informativeness within a given time frame, will help in identifying elements of information for inclusion in a summary. More ambitiously, in future work, we plan to generate automatic summaries from the information jointly provided by tweets and web documents.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053, the U.S. NSF Grants IIS-0953149, IIS-1144111, IIS-0905215, CNS-0931975 and the U.S. DARPA BOLT program, the U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

1. Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. Class-based n-gram models of natural language. *Computational Linguistics 18* (1992), 467–479.

2. Carterette, B., and Chandar, P. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, ACM (New York, NY, USA, 2009), 1287–1296.

3. Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, ACM (2011), 675–684.

4. Couzin, I. Collective minds. *Nature 445* (2007).

5. Deng, H., Han, J., Zhao, B., Yu, Y., and Lin, C. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proc. ACM SIGKDD2011*, ACM (2011), 1271–1279.

6. Deng, H., Lyu, M. R., and King, I. A generalized co-hits algorithm and its application to bipartite graphs. In *Proc. ACM SIGKDD2009* (2009).

7. Diakopoulos, N., De Choudhury, M., and Naaman, M. Finding and assessing social media information sources in the context of journalism. In *Proc. Conference on Human Factors in Computing Systems (CHI)* (2012).

8. Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, ACM (New York, NY, USA, 2010), 331–340.

9. Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, Association for Computational Linguistics (Stroudsburg, PA, USA, 2010), 295–303.
10. Golder, S. A., Marwick, A., and Yardi, S. A structural approach to contact recommendations in online social networks. In *Proc. SIGIR2009 Workshop on Search in Social Media* (2009).
11. Gupta, M., and Han, J. Heterogeneous network-based trust analysis: a survey. *SIGKDD Explor. Newsl. 13*, 1 (Aug. 2011), 54–71.
12. Gupta, M., Zhao, P., and Han, J. Evaluating event credibility on twitter. In *SDM* (2012), 153–164.
13. Hannon, J., Bennett, M., and Smyth, B. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender Systems* (2010).
14. Huang, M., Yang, Y., and Zhu, X. Quality-biased ranking of short texts in microblogging services. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, Asian Federation of Natural Language Processing (Chiang Mai, Thailand, November 2011), 373–382.
15. Inouye, D., and Kalita, J. K. Comparing twitter summarization algorithms. In *IEEE SocialCom 2011* (2011).
16. Jarvelin, K., and Kekalainen, J. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems 20* (2002).
17. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. ACM 46*, 5 (Sept. 1999), 604–632.
18. McDonald, R. A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European conference on IR research*, ECIR'07, Springer-Verlag (Berlin, Heidelberg, 2007), 557–564.
19. Mihalcea, R. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proc. ACL2004* (2004).
20. Mihalcea, R., and Tarau, P. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, vol. 4, Barcelona: ACL (2004).
21. Miller, G. A. Wordnet: A lexical database for english. *Communications of the ACM 38* (1995), 39–41.
22. Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. In *Proc. the 7th International World Wide Web Conference* (1998).
23. Saavedra, S., Hagerty, K., and Uzzi, B. Synchronicity, Instant Messaging and Performance among Financial Traders. *PNAS* (2011).
24. Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. Twitterstand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2009).
25. Sharifi, B., Hutton, M.-A., and Kalita, J. K. Experiments in microblog summarization. In *IEEE Second International Conference on Social Computing (SocialCom)* (2010).
26. Sornil, O., and Greeu, K. An automatic text summarization approach using content-based and graph-based characteristics. In *IEEE Conference on Cybernetics and Intelligent Systems* (2006).
27. Uysal, I., and Croft, W. B. User oriented tweet ranking: A filtering approach to microblogs. In *Proc. CIKM2011 (Poster)* (2011).
28. Wang, D., Abdelzaher, T., Ahmadi, H., Pasternack, J., Roth, D., Gupta, M., Han, J., Fatemieh, O., Le, H., and Aggrawal, C. On bayesian interpretation of fact-finding in information networks. In *Proc 14th International Conference on Information Fusion (Fusion '11)* (2011).
29. Wang, D., Le, H., Kaplan, L., and Abdelzaher, T. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proc. 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN)* (2012).
30. Weng, J., Lim, E.-P., Jiang, J., and He, Q. Twitterrank: Finding topic-sensitive influential twitterers. In *Proc. WSDM2010* (2010).
31. Yamaguchi, Y., Takahashi, T., Amagasa, T., and Kitagawa, H. Turank: Twitter user ranking based on user-tweet graph analysis. In *Proc. WISE2010* (2010).
32. Zanzotto, F. M., Pennacchiotti, M., and Tsioutsiouliklis, K. Linguistic redundancy in twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, Association for Computational Linguistics (Stroudsburg, PA, USA, 2011), 659–669.
33. Zubiaga, A., Spina, D., Amigó, E., and Gonzalo, J. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, ACM (2012), 319–320.