2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing

Finding Credible Information Sources in Social Networks Based on Content and Social Structure

Kevin R. Canini
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
kevin@cs.berkeley.edu

Bongwon Suh†
Adobe Systems, Inc.
Advanced Technology Labs
San Jose, CA 95011
bongwon@adobe.com

Peter L. Pirolli
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304
pirolli@parc.com

***Abstract*—A task of primary importance for social network users is to decide whose updates to subscribe to in order to maximize the relevance, credibility, and quality of the information received. To address this problem, we conducted an experiment designed to measure the extent to which different factors in online social networks affect both explicit and implicit judgments of credibility. The results of the study indicate that both the topical content of information sources and social network structure affect source credibility. Based on these results, we designed a novel method of automatically identifying and ranking social network users according to their relevance and expertise for a given topic. We performed empirical studies to compare a variety of alternative ranking algorithms and a proprietary service provided by a commercial website specifically designed for the same purpose. Our findings show a great potential for automatically identifying and ranking credible users for any given topic.**

I. Introduction

The rising popularity of social networks has made information sharing and discovery easier than ever before, due to the ability to publish content to large, targeted audiences. Such networks enable their participants to simultaneously become both consumers and producers of content, shifting the role of information broker from a few dedicated entities to a diverse and distributed group of individuals. While this fundamental change allows information consumers more flexibility in choosing what content to follow, it makes it necessary for users to discover, evaluate, and select sources of information that are worth their attention from a vast pool of potential choices. If a social network user is interested in receiving information about a particular topic of interest, a task of primary importance is to decide which other users' updates to subscribe to in order to maximize the relevance, credibility, and quality of the information received.

Solving this problem can be a difficult task due to the shear number of accounts to choose from and a lack of helpful tools built into social networking services. For example, Twitter[1] currently has about 200 million registered users and provides only a simple text search mechanism which returns a reverse-chronologically ordered list of the most recent tweets (posted status messages) containing a search term[2]. While this can be helpful for very specific queries that only a handful of experts are expected to mention, for many topics, much of the results can be unhelpful and only tangentially related to the desired information.

Ideally, given a topic of interest, one would hope to find users who provide credible information about that topic. Credibility is often conceived as a combination of expertise and trust [1], and expertise is commonly defined by the support and nomination of other professionals [2]. Additionally, the relevance of a person's discussions can often serve as a cue towards expertise. Applying these definitions to social networks, credibility is associated with people who not only frequently publish topically relevant content but also are trusted by their peers. Unfortunately, social network users are unable to directly observe how well someone is trusted in a particular domain. Therefore, one of the most important aspects of credibility is also of the hardest for a non-expert to gauge. However, links between users in a social network serve the function of a vote of support between them, so it should be possible to estimate expertise from observable link data.

Beyond link analysis, another useful factor in determining topical relevancy is the actual content of a user's messages. Topic modeling has proven to be a useful tool for analyzing natural language data in many problem domains. Our approach combines the analysis of the link structure of social networks with topic models of the content of messages to identify and evaluate topically relevant and credible sources of information in social networks.

The paper is organized as follows. We first discuss relevant prior work. Next, we describe an experimental study designed to measure how much different factors affect both explicit and implicit judgments of credibility in a social network setting. We then introduce our algorithm for identifying and ranking Twitter users and describe two user evaluations we performed to investigate its performance.

II. Related Work

Much research has recently been focused on social networks and microblogs, particularly Twitter. As Twitter grows more

†This work was completed entirely while this author was at the Palo Alto Research Center.

[1]http://www.twitter.com

[2]This capability has recently been augmented with an option for displaying "Top" tweets.

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

popular, it serves as a real-world example for studying the theory of social networks and applying and testing scalable algorithms designed to analyze large social networks. Cha et al. [3] studied the factors that indicate the influence of Twitter users, arguing that despite its widespread use, in-degree alone is not necessarily a good indicator of influence. Duan et al. [4] and Chen et al. [5] explored recommending individual tweets to users based on a variety of cues. Bernstein et al. [6] designed a system for organizing and displaying tweets by topic.

One of the reasons social networks like Twitter are interesting from a research perspective is that they contain information in the form of dynamic social graphs as well as textual content shared along the edges of the graphs. A significant portion of our work focuses on learning representations of the textual content of social networks. A number of recent papers have described ways of applying topic models such as latent Dirichlet allocation (LDA) [7] to microblogs and social networks [8, 9, 10]. Ramage et al. [11] utilized a topic model called labeled LDA to classify individual posts in Twitter into four basic categories. Weng et al. [9] combined topic modeling with webpage ranking techniques to calculate topic-based influence rankings of Twitter users. The 140-character length limit of Twitter posts makes them somewhat unsuitable for analysis with popular topic models. Individual tweets tend to be too short to convey strong information about the precise mixture of latent topics within them. Bernstein et al. [6] overcame this limitation by using web search engines to expand the content of each tweet with words from similar webpages. Other researchers have applied the LDA topic model to Twitter by concatenating all of a user's tweets into a single document [9], which is the approach we follow.

III. Study of Factors Affecting Credibility

The rich variety of data available about social network users enables us to have a better understanding of which factors are most highly correlated with judgments of relevancy and credibility. To this end, we conducted an experiment with users of the Twitter social network to determine to what degree different factors contribute to judgments of other users' credibility and expertise for particular topics.

Our study is modeled after that of Birnbaum and Stegner [12]. In their experiments, participants were asked to judge the fair market value of used cars both before and after observing the Kelley blue book price as well as an appraisal by a third party. The third party varied in both their expertise in the domain of used cars and their bias (they were either affiliated with the car's buyer or seller, or they were neutral). In Birnbaum and Stegner [12], as well as in Birnbaum [13], it is argued that a simple averaging model of source credibility is consistent with a wide variety of experimental results in decision making studies:

$$R = \frac{\sum_i w_i s_i}{\sum_i w_i}, \tag{1}$$

where R is the predicted response (e.g., a participant's estimated value of a car), each s_i variable is the scale value of a source's appraisal, and each w_i variable is a weight determined by the perceived credibility of source i, which may depend on such factors as perceived expertise (e.g., their knowledge and skill in the relevant domain), bias (e.g., general tendency to over- or under-estimate true values), or point of view (e.g., affiliation with either the buyer or the seller). The *a priori* judgment of the subject is represented by the value s_0 with an associated weight w_0.

Our strategy is to perform a simplified version of the used car prices study of Birnbaum and Stegner [12], associating the third-party car price appraisals with profiles of Twitter users. By asking participants to judge the value of used cars both before and after viewing the third-party appraisals, we can measure the effect of each Twitter user's opinion on the participants' judgments, which gives an implicit rating of the perceived credibility and expertise of the Twitter users.

A. Participants

We recruited 98 participants on Amazon Mechanical Turk to participate in our experiment. Participants received $1.00 as compensation for their time and effort. We confirmed that each was a current and active user of the Twitter social network by asking them to provide their Twitter user names (which we verified), the length of time they have been a user, and the frequency with which they check their messages.

B. Materials

1) Twitter Profiles: We first prepared a set of five different domains of expertise: *cars*, *investing*, *wine*, *fantasy football*, and *dating*. For each domain, we manually selected 10 Twitter accounts of a high level of relevance and expertise for that domain. The accounts were collected from a popular Twitter directory service called WeFollow[3], which curates lists of the most influential Twitter accounts for a large number of domains. In addition, we also manually selected 10 Twitter accounts whose tweets we felt did not reflect expertise in any particular domain, but were mostly related to personal issues and day-to-day life. For each of these 60 Twitter accounts, we harvested the timeline of all the tweets posted from that account[4] for use in the experiment.

2) Social Status: To control for the social status indicators within the Twitter profiles, we created two social status levels: *high* and *low*. These two levels were differentiated by the number of followers (users who subscribe to the account), number of followees (users who the account is subscribed to), number of tweets (messages ever published by the account), and number of list memberships (instances where another user added the account to a curated list). For each trial of the experiment, each of these factors was drawn uniformly at random within an interval determined by the social status level.

[3]http://www.wefollow.com

[4]The Twitter API limits the number of tweets that can be accessed to the most recent 3000. In cases where user posted more than 3000 messages, only the most recent 3000 were collected.

For the number of followers, the interval was 10,000–100,000 for the high level 50–200 for the low level. The intervals for the number of followees were 50–5,000 and 5–100 for the high and low levels, respectively. For the number of tweets, they were 1,000–5,000 and 50-1,000, and for the number of list memberships, they were 100-2,500 and 0–5, respectively. These numbers were chosen on the basis of a brief survey of the typical ranges of these statistics for a number of actual Twitter accounts.

3) Word Clouds: Because we were interested in measuring the impact of varying the presentation style of each Twitter profile's textual content, we prepared two different word cloud representations for each account: one based on tf-idf [14] and the other based on latent Dirichlet allocation (LDA) [7]. Tf-idf is a well-known measure of how representative a given word is of a particular document in a corpus of documents. The tf-idf score increases proportionally with the number of times the word appears in the document, but decreases as it appears in more documents in the corpus overall. We calculated the tf-idf score for each word and Twitter profile in our collection. We then combined the top 50 scoring words for each profile to construct word clouds like the examples shown in Figure 1.

LDA is a popular topic model which attempts to discover a set of topics associated with a collection of documents. The LDA model assumes that the documents contain T unobserved topics, where the number T is a parameter of the model. Each topic t is assumed to be a probability distribution over the words in the corpus, so that the probability of word w occurring in topic t is given by $\phi_w^{(t)}$. Each document is modeled as a mixture of these topics, with the mixture weight of topic t within document d given by $\theta_t^{(d)}$. The inference procedure takes as input the number of times each word appears in each document and returns estimates of the values of $\phi_w^{(t)}$ for each topic and word and $\theta_t^{(d)}$ for each document and topic.

In our application, we model each Twitter account as a single document, so the result of inference in LDA is a set of estimates of the strength of association between each Twitter account and each topic, as well as the strength of association between each topic and each vocabulary word. From these quantities, the association between an account d and a specific word w can be calculated as

$$s_{w,d}^{\text{LDA}} = \sum_{t=1}^{T} \theta_t^{(d)} \phi_w^{(t)}, \tag{2}$$

where the summation is taken over the T latent topics. As with the tf-idf scores, we used the LDA scores to create word clouds of the top 50 words for each Twitter profile. Examples of the LDA-based word clouds are shown in Figure 1.

C. Design and Procedure

Each participant completed 33 trials; the first 3 were practice trials and were not used in the analysis (the participants were informed of this). In each trial, the participant was first presented with a list of basic information about a particular used car, including its make and model, year of manufacture, number of miles, and standardized value given by the Kelley Blue Book (KBB) price. After viewing this information, the participant was asked to estimate the fair market value of the car in dollars (the "pre-judgment"). Next, the participant was presented with information about a particular Twitter user who, they were told, had independently appraised the same car and judged it to be worth a particular amount of money. This appraisal value was randomly drawn from a uniform distribution ranging between either 0.6–0.8 of 1.2–1.4 times the KBB price, with each range chosen with probability 50%. After viewing the third party's Twitter profile and appraisal, the participant was asked once again to estimate the car's fair market value. This second judgment was called the "post-judgment". By comparing the participant's pre-judgment and post-judgment, we can calculate an implicit value of the perceived credibility of the Twitter user.

The specific information presented about the third-party appraiser varied according to a factorial experimental design including the three factors. The first factor, "domain of expertise", took on three levels: *on-topic* (cars), *cross-topic* (investing, wine, fantasy football, or dating), or *off-topic* (day-to-day topics with no particular area of expertise). The second factor, "social status", took on two levels: *high* or *low*. Finally, the "visualization style" factor took on five levels: *tweets only*, *LDA word cloud+tweets*, *tf-idf word cloud+tweets*, *LDA word cloud only*, and *tf-idf word cloud only*. The combination of these factors led to an experimental design with $3 \times 2 \times 5 = 30$ trials.

The third-party appraiser for each trial was represented by a Twitter profile containing a randomized user name and icon, a set of social status statistics which varied according to the social status factor, and a visualization of the content of the user's tweets which varied according to the visualization style factor. For the *tweets only* level of the visualization style factor, a list of the user's 40 most recent tweets was presented. For the *LDA word cloud+tweets* and *tf-idf word cloud+tweets* levels, the user's precomputed LDA or tf-idf word cloud was presented above the list of 40 tweets. For the *LDA word cloud only* and *tf-idf word cloud only* levels, the word cloud was presented without the list of 40 tweets.

Using the averaging model of source credibility given in Equation 1, the participant's pre-judgment can be modeled as

$$R_1 = \frac{w_0 s_0 + w_{\text{KBB}} s_{\text{KBB}}}{w_0 + w_{\text{KBB}}}, \tag{3}$$

where s_0 and w_0 are the value and weight of the *a priori* judgment of the car's value (before seeing the KBB price), and s_{KBB} and w_{KBB} are the value and weight of the KBB price. The post-judgment can be modeled as

$$R_2 = \frac{w_0 s_0 + w_{\text{KBB}} s_{\text{KBB}} + w_t s_t}{w_0 + w_{\text{KBB}} + w_t}, \tag{4}$$

where s_t and w_t are the value and weight of the third-party appraisal.

Using these models of the participant's two responses, we can calculate the relative weight attributed to the third party,

(a) tf-idf: cars (b) tf-idf: investing (c) tf-idf: dating (d) tf-idf: off-topic

(e) LDA: cars (f) LDA: investing (g) LDA: dating (h) LDA: off-topic

Fig. 1: Examples word clouds produced to represent the textual content of the Twitter profiles used in the experiment. The size of each word is a function of its tf-idf or LDA score. The top row of word clouds were produced using tf-idf scores, and the bottom row were produced using LDA scores. The two word clouds in each column correspond to the same Twitter account.

which gives us an implicit judgment of the credibility of the Twitter user. The implicit credibility judgment is given by

$$C = \frac{w_t}{w_0 + w_{\text{KBB}} + w_t} = \frac{R_2 - R_1}{s_t - R_1}. \quad (5)$$

This formula calculates how much the participant shifted their judgment towards the third-party appraisal s_t. A value of 0 indicates no shift (the post-judgment R_2 equals the pre-judgment R_1), while a value of 1 indicates a complete shift ($R_2 = s_t$). Values of C less than 0 or greater than 1 are possible, but were rarely observed.

In addition, for each trial we asked participants to answer the question "How much would you trust this person's recommendations about used cars?" using a Likert response scale between 1 and 5. These responses constituted explicit credibility ratings.

D. Results and Discussion

We fit linear models to both the implicit and explicit credibility ratings from every trial. The model included a separate bias term for each participant to account for individual differences in people's general propensities for credibility judgments. Coefficients were included for each of the three experimental factors: social status, domain of expertise, and visualization style. Table I presents the fitted weights and significance levels according to an ANOVA analysis.

The correlation between the explicit and implicit credibility judgments was 0.444. Although this seems somewhat low, is is partially due to the fact that the implicit judgments were continuous-valued while the explicit judgments were discrete-valued. We found that the domain of expertise factor had a strong influence on credibility judgments, and social status had a smaller influence. The visualization factor had the smallest influence on both sets of judgments.

While the results seem to confirm general intuitions about the relative importance of the various factors of a social network user's profile when determining credibility, we found a few surprising results in the data. First, when the five domains of expertise (and the one domain of non-expertise) were modeled separately, there were interesting differences between their individual effects. Not surprisingly, the car domain (the "on-topic" level) led to the highest ratings; however, we also found that the users with domain expertise in wine or dating received significantly *lower* credibility ratings (for both implicit and explicit ratings) than those with no particular domain expertise (the "off-topic" level), suggesting that, for participants, expertise in wine or dating indicates a less-than-average familiarity with used car prices. We also observed a particular effect with respect to the visualization style factor: the combination of tweets with either type of word cloud produced by far the highest credibility ratings (again for both implicit and explicit ratings). This suggests that neither tweets alone nor word clouds alone provide sufficient information for participants to grant a high credibility rating to a Twitter user, but the combination of presenting specific tweets along with a summary word cloud leads to higher judged credibility. These two visualizations apparently provide complementary sources of information about social network profiles; perhaps word clouds convey a user's general tendency to mention particular topics (indicating overall relevance), while individual tweets can provide specific examples of expertise or indicators of

TABLE I: Fitted linear model coefficients from the credibility experiment. Judgments were standardized to have a mean of 0 and standard deviation of 1, so fitted coefficients are in units of standard deviations of the judgments. Coefficients for *low*, *off-topic* and *tweets only* are 0 due to linear dependence between factor levels. ANOVA statistics for the implicit judgments are $F = 21.99, p = 2.94 \times 10^{-6}$ for the social status factor, $F = 104.17, p < 2.2 \times 10^{-16}$ for the domain of expertise factor, and $F = 2.16, p = 0.07$ for the visualization style factor. For the explicit judgments, they are $F = 53.98, p = 2.99 \times 10^{-13}$ for social status, $F = 393.37, p < 2.2 \times 10^{-16}$ for domain of expertise, and $F = 5.43, p = 2.38 \times 10^{-4}$ for visualization style.

	Social status		Domain of expertise			Visualization style				
Judgment	High	Low	On-topic	Cross-topic	Off-topic	LDA	tf-idf	LDA+	tf-idf+	Tweets only
Implicit	0.1779	0	0.5748	-0.0117	0	-0.0325	0.0318	0.1008	0.1103	0
Explicit	0.2388	0	0.9982	0.0657	0	-0.0942	-0.0887	0.1039	0.0388	0

trustworthiness such as word choice and writing style.

IV. Ranking Topically Relevant Users

The results of the experiment described above indicate that the credibility of a Twitter account with respect to a particular domain depends in large part on the strength of association between the textual content of the account and the domain in question, and to a lesser extent, the social status of the account. Taking these factors into account, we designed a novel method of identifying and ranking users in Twitter according to their relevancy to any given topic. Our algorithm first performs a standard Twitter search (which returns a simple reverse-chronological list of results) to identify a small set of users who are associated with a query. It then applies a social filter, identifying users whose followers appear frequently in the search result. Finally, we use topic modeling to analyze the textual content of the highest-scoring users and re-rank them by this criterion. By combining a basic text search with a social ranking technique and topic modeling analysis, the algorithm generates a ranked list of relevant, trusted, and credible Twitter users for any given topic.

A. Identifying Candidates

The first step in our algorithm is to identify a set of candidates who are potentially relevant to the topic of interest. Given a topic expressed as a search term, a standard Twitter search is first executed using the Twitter API[5]. Taken alone, this search procedure is not particularly useful for identifying relevant users because the results are only a chronologically ordered list of the 1,500 most recent tweets containing the search term. However, those who published the tweets in the search result do form a small set of users, which we call *Voters*, who are associated with the topic.

The next step in our algorithm is to measure the opinions of the Voters by observing who they follow. If one user follows another in Twitter, it indicates that the first user values the information published by the second. Taking advantage of this fact, the algorithm next builds a set of users, which we call *Candidates*, by including anyone who is followed by at least one of the Voters. This process not only expands the set of potentially relevant candidates, it also provides a way to compute a relevancy score for each candidate, since a more influential, trustworthy Candidate will presumably be followed by more Voters.

For each user u in the Candidates set, we retrieve the number of Voters who follow user u, called f_u, and the total number of Twitter users who follow user u, called F_u. The number f_u can be explained by a process (depicted in Figure 2) where each of the Voters casts a vote for each of their followees, and f_u is the number of votes received by user u. Using just the two numbers f_u and F_u, we compute a social status score for each member in the Candidates set and rank them accordingly.

B. Social Ranking

Once we have identified a set of Candidates and retrieved the relevant numbers f_u and F_u for each user u in the set, we can compute the relevancy of each user to the query topic. Before describing the formula used by our algorithm, we describe a series of alternative formulas of increasing complexity, building up to our own. For the remainder of this section, we will simplify the notation by writing f and F instead of f_u and F_u, assuming the discussion is specific to a given user. All of the following formulas are summarized in the first two columns of Table II.

The first and most basic relevancy measure one could consider using is just the number f itself. We call this measure *NumVotes*. This measure is appealing because it directly counts how many times a user's followers have recently tweeted about the topic; however, in practice it tends to too heavily favor generally popular Twitter users who are not relevant to the topic of interest. For example, a widely-followed user such as Barack Obama would rank very highly for virtually any search query.

Next, we consider the relevancy measure f/F, called *DivF*. This rationale behind this measure is that it counts the proportion (rather than the actual number) of one's followers who showed up in the search results. Intuitively, the higher the proportion of a user's followers who are associated with a topic, the more trusted that user is. In practice, however, we found that this measure often overpenalizes generally popular users, underpenalizes unpopular users, and is overly sensitive to spuriously large values of f when F is small.

To strike a balance between the NumVotes and DivF measures, we consider the measure $f/\log F$, called *DivLogF* which takes its inspiration from the tf-idf method from the information retrieval literature.

[5]http://dev.twitter.com/doc

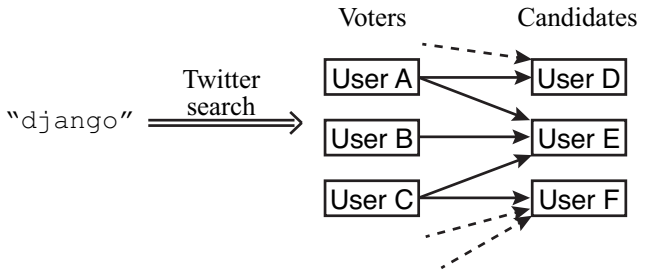

Fig. 2: A depiction of the initial stages of the algorithm which identify the set of Voters and Candidates for a particular query. Arrows between users indicate *following* relationships. Dashed arrows indicate cases where the followers are not in the Voters set. In this example, $f_D = 1$, $F_D = 2$, $f_E = 3$, $F_E = 3$, $f_F = 1$, and $F_F = 3$.

Finally, we introduce our preferred relevancy measure, called BetaBin(α, β). It is motivated from a Bayesian probability perspective. If we assume that each of the user's F followers is randomly included in the Voters set independently and with probability p, then f can be approximated by a Binomial(F, p) probability distribution[6]. We use a Beta(α, β) prior distribution over p, so after observing that f of the user's F followers occur in the Voters set, the posterior probability of p follows a Beta$(f + \alpha, F + \beta)$ distribution. The expected value of this posterior distribution gives us an estimate of how probable each of the user's followers is to show up in the Voters set. The expected value has a simple formula:

$$\mathbb{E}[p|f, F] = \frac{f + \alpha}{F + \alpha + \beta}, \tag{6}$$

which defines our relevancy measure BetaBin(α, β). This measure acts like NumVotes when $F \ll \alpha + \beta$, since

$$\frac{f + \alpha}{F + \alpha + \beta} \approx \frac{f + \alpha}{\alpha + \beta} \sim f, \tag{7}$$

and it acts like DivF when $F \gg \alpha + \beta$, since

$$\frac{f + \alpha}{F + \alpha + \beta} \approx \frac{f}{F}. \tag{8}$$

Thus, it has the benefits of DivF, measuring the proportion of one's followers who are in the search results, while appropriately penalizing unpopular users like NumVotes does.

Since the proportion of a user's followers who show up in the Voters set is expected to be quite low on average, it is generally a good idea to set $\alpha \ll \beta$. As such, in our evaluations, we compare multiple versions of the measure, all with $\alpha = 1$, with β ranging between 10^2 and 10^6.

C. Topic Modeling

The algorithms described above take into account information about the link structure of the social network, restricting attention to sections of the graph highlighted by a simple text search over recent activity. We hypothesized that analyzing the textual content of each account would yield a stronger signal of its topical relevance, so we implemented a method to re-sort the ranked results based on a topic modeling analysis. We compiled a list of the top 28,000 scoring users according to the Beta-Binomial ranking formula for a set of ten queries: "biking", "democrat", "django", "hadoop", "medicine", "photoshop", "republican", "startup", "teaparty", and "wine". The α and β parameters of the Beta-Binomial formula were optimized independently for each query by comparing to the lists collected from WeFollow, although they could also be fit to the search results via maximum likelihood estimation. We collected the entire tweet histories of these users and ran the latent Dirichlet allocation (LDA) topic model on the corpus[7]. The LDA results provide a way of determining the topical similarity of any user to a search query based on the content of the user's tweets. The scoring function we used is the value in Equation 2, where w is the search query. Formally, this quantity gives the probability of the user generating the query w under the learned parameters of the LDA model. We compare the original ranked lists to the re-ranked lists using the LDA analysis.

V. Algorithm Evaluation

To evaluate the various algorithms presented above, we first performed a modest case study on the search query "django" using two human volunteers, which was followed by a more thorough evaluation of five search queries using Amazon Mechanical Turk participants.

A. Case Study: "django"

1) Method: As a preliminary investigation of the feasibility of the algorithms presented above, we first compared the ranked lists they generated for the query "django" (a Python web application framework), along with the list of influential users for the topic "django" provided by WeFollow. Using Twitter's API, we queried the term "django" on July 21, 2010, obtaining 1,500 tweets authored by 980 unique authors, who formed the Voters set. Expanding to those users' followees, we compiled 234,166 users who formed the Candidates set. The Candidates were ranked according to each of the relevancy measures defined above. We also collected the top 200 users for the same query from WeFollow on July 27, 2010. The LDA re-ranking algorithm was not implemented before running this evaluation, so it was not included.

We first measured the precision of each algorithm; that is, how many of each algorithm's top-ranked users actually were relevant to the topic. We prepared the top 20 list for each relevancy measure and merged them all together with the top 20 list from WeFollow, producing a list of 97 candidate experts. We recruited two Twitter users with Django experience and asked them to classify each of the 97 users as either relevant or irrelevant to Django. One participant identified 38

[6]The Binomial approximation is not exact because it has support from 0 to F, while f is actually bounded by the number of Voters returned by the search procedure, typically around 1500. The true distribution is hypergeometric.

[7]We used $T = 500$ topics, with hyperparameters $\alpha = 0.5$ and $\beta = 0.1$, which gave the best perplexity scores out of 24 tested sets of hyperparameters.

TABLE II: Results from the case study for the search term "django". The precision columns show the number of users in each top 20 list who were judged as relevant by two human raters. The recall column shows how many users from a list of 25 known experts were identified by each algorithm.

Measure	Formula	Precision 1	Precision 2	Recall
NumVotes	f	7	6	6
DivF	f/F	0	2	0
DivLogF	$f/\log F$	13	12	8
BetaBin$(1, 10^2)$	$(f+1)/(F+10^2+1)$	15	11	11
BetaBin$(1, 10^3)$	$(f+1)/(F+10^3+1)$	**19**	**17**	**13**
BetaBin$(1, 10^4)$	$(f+1)/(F+10^4+1)$	17	15	11
WeFollow	N/A	**19**	14	10

users as relevant, and the other chose 31 users. They agreed on 27 relevant users and 55 irrelevant users, disagreeing on 15 cases (Cohen's $\kappa = 0.66$, indicating substantial inter-rater agreement).

Next, we measured the recall of each algorithm; that is, given a list of known experts, how many of them were identified by each algorithm. We used a list of 25 recognized Django experts[8] on Twitter compiled by one of the main developers of Django. We then counted how many of these users were present in the top 100 list of each algorithm and the top 100 list from WeFollow. We chose to use the top 100 lists because this is roughly the longest list one can be reasonably expected to look through when searching for relevant users.

2) Results: The results of the evaluation are summarized in Table 1. The *Measure* and *Formula* columns give, respectively, the name of each measure and the formula it uses to calculate relevancy. The *Precision 1* and *Precision 2* columns give the number of users in each measure's top 20 list who were rated as relevant to the topic by the first and second human raters, respectively. The *Recall* column gives the number of the 25 known experts who were found in each algorithm's top 100 list. In the first precision evaluation, the BetaBin$(1, 10^3)$ and WeFollow algorithms had the best performance, and in the second, the BetaBin$(1, 10^3)$ algorithm alone had the best performance. In the recall evaluation, the BetaBin$(1, 10^3)$ algorithm again performed the best. Interestingly, although the BetaBin$(1, 10^3)$ measure is quite similar to the DivF measure, their performances on every evaluation were completely opposite. This suggests that while finding users whose followers are highly associated with the topic of interest is a good strategy, a major obstacle is being able to identify the users with only a few followers who received a relatively large number of votes by chance alone.

B. Mechanical Turk Evaluation

1) Method: Following the "django" case study, we performed a more thorough study of the performance of the various algorithms (as well as the LDA re-ranking algorithm) on five different search queries, using Amazon Mechanical Turk participants to rate the top-ranked Twitter users according to their relevance and expertise and whether they were worth following. The search queries we used are "biking", "medicine", "photoshop", "teaparty", and "wine". For each query, we compiled each algorithm's top-20 list and asked a number of participants on Amazon Mechanical Turk to rate each Twitter by agreeing or disagreeing with each of the following statements: "This Twitter user seems to be a source of relevant information relating to the search term.", "This Twitter user seems to be an expert in an area relating to the search term.", and "If I were interested in learning more things relating to the search term, I would follow this Twitter user." Each Twitter user was evaluated a number of times, and a consensus was found among the participants.

2) Results: The results are summarized in Table III. In general, the WeFollow rankings and the LDA rankings performed very well. The Bayesian Beta-Binomial algorithms without LDA re-ranking also performed well, often producing results competitive with those of LDA and WeFollow. These results suggest that incorporating a content-based topic analysis of users' tweets significantly improves results, producing rankings which are often better than opt-in expert lists.

VI. Discussion

Our algorithm uses a live Twitter search query result as a seed for user expansion. Thus, the result adapts to temporal trends in topic. For example, if the meaning of a search term changes abruptly due to current events, the set of voters and candidates generated by the Twitter search will adapt to these changes and alter the results generated by our algorithm. We have yet to investigate whether our method can maintain stability amidst temporary changes in the search results while adapting to legitimate trends in the way language is used and the set of credible users in social networks.

A variety of other methods for discovering topically relevant users depend on manual curation or input from users. For example, WeFollow requires that a user register their account for a specific keyword before their account can appear in the results for that keyword. Approaches based on Twitter lists, such as MyTwitterCloud, depend entirely on the lists created by Twitter users and are therefore can be misled in cases where few lists exist for a given topic or a topic does not lend itself well to the Twitter list mechanism.

We believe the general method described in this paper can be applied to other social networks where the opinions of the crowd provide a strong signal as to what information within the network is highly relevant. The same techniques

[8]http://twitter.com/simonw/djangonaughts

TABLE III: Results from the Mechanical Turk study. For each of the five search terms, the table lists the number of users from each algorithm's top-20 list who were rated by Turk participants as having tweets relevant to the search term (r), being likely to be an expert in an area related to the search term (e), and being someone who the participant would follow if they were interested in the search term (f).

	"biking"			"medicine"			"photoshop"			"teaparty"			"wine"		
Measure	r	e	f	r	e	f	r	e	f	r	e	f	r	e	f
NumVotes	1	1	1	0	0	0	0	0	0	13	8	12	4	4	4
DivF	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0
DivLogF	1	1	1	1	1	0	1	0	0	**16**	10	**14**	12	9	10
BetaBin$(1, 10^2)$	7	6	6	8	6	6	4	4	4	13	6	12	19	13	11
BetaBin$(1, 10^3)$	18	14	14	16	11	11	9	8	8	**16**	8	13	16	11	11
BetaBin$(1, 10^4)$	18	17	**17**	15	6	8	16	10	10	11	8	7	18	12	11
BetaBin$(1, 10^5)$	17	13	13	15	11	10	15	8	7	15	10	12	18	14	12
BetaBin$(1, 10^6)$	4	3	4	2	2	1	6	2	2	15	10	**14**	18	13	13
LDA	**20**	16	16	14	10	11	**20**	**20**	**20**	11	5	9	**20**	**20**	**20**
WeFollow	19	**18**	**17**	**17**	**14**	**14**	18	16	14	**16**	**11**	11	19	16	14

can also be applied to similar problems such as recommending individual messages or conversation threads rather than users. Combinations of network-based information and topic-based textual analyses will yield powerful tools to discover and evaluate content in social networks.

VII. Conclusion

In this paper, we describe an approach towards solving the problem of identifying reputable, credible sources of relevant information in social networks. We performed an experiment to explore the extent to which various factors affect both explicit and implicit credibility levels between users of a social network. Based on the findings of this study, we designed an algorithm which is sensitive to both the content and social status of social network users. By combining a basic text search with an analysis of the social structure of the network, the algorithm generates a ranked list of relevant users for any given topic. We found that a content-based topic analysis of the social network proved especially useful in identifying relevant and credible users to follow. To investigate the feasibility of the algorithm, we performed a case study and a more thorough evaluation, comparing rankings generated by the algorithm with rankings provided by a commercial website. The algorithm shows great potential to help users identify interesting users to follow in Twitter. We hope that this research will inform the design of recommendation systems for Twitter and other social networks.

Acknowledgements

This research was sponsored by the Army Research Laboratory under Cooperative Agreement W911NF-09-2-0053.

References

[1] S.Y. Rieh and D.R. Danielson. Credibility: A multidisciplinary framework. *Annual review of information science and technology*, 41(1):307–364, 2007.

[2] K. Ericsson, N. Charness, R. Hoffman, and P. Feltovich, editors. *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, 2006.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Int'l Conf. Weblogs and Social Media*, 2010.

[4] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H. Shum. An empirical study on learning to rank of tweets. In *23rd Int'l Conf. on Computational Linguistics*, 2010.

[5] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: Experiments on recommending content from information streams. In *28th Int'l Conf. on Human Factors in Computing Systems*, 2010.

[6] M. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. Chi. Eddi: Interactive topic-based browsing of social status streams. In *23rd Annual ACM Symposium on User Interface Software and Technology*, 2010.

[7] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[8] X. Phan, L. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *17th Int'l Conf. on World Wide Web*, 2008.

[9] J. Weng, E. Lim, J. Jiang, and Q. He. TwitterRank: Finding topic-sensitive influential twitterers. In *3rd ACM Int'l Conf. on Web Search and Data Mining*, 2010.

[10] A. Ritter, C. Cherry, and B. Dolan. Unsupervised Modeling of Twitter Conversations. In *NAACL*, 2010.

[11] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Int'l AAAI Conf. on Weblogs and Social Media*, 2010.

[12] M. Birnbaum and S. Stegner. Source credibility in social judgment: Bias, expertise, and the judge's point of view. *J. Personality and Social Psychology*, 37(1):48–74, 1979.

[13] M. Birnbaum. Base rates in Bayesian inference. In Rdiger F. Pohl, editor, *Cognitive Illusions*. Psychology Press, 2004.

[14] G. Salton and M. McGill, editors. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.