2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust

It's not in their tweets: Modeling topical expertise of Twitter users

Claudia Wagner*, Vera Liao†, Peter Pirolli‡, Les Nelson‡ and Markus Strohmaier§
*Institute of Information and Communication Technologies, JOANNEUM RESEARCH, Graz, Austria
Email: claudia.wagner@joanneum.at
†Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois
Email: liao28@illinois.edu
‡Palo Alto Research Center, Palo Alto, California
Email: pirolli@parc.com, lnelson@parc.com
§ Knowledge Management Institute and Know-Center, Graz University of Technology,Graz, Austria
Email: markus.strohmaier@tugraz.at

***Abstract*—One of the key challenges for users of social media is judging the topical expertise of other users in order to select trustful information sources about specific topics and to judge credibility of content produced by others. In this paper, we explore the usefulness of different types of user-related data for making sense about the topical expertise of Twitter users. Types of user-related data include messages a user authored or re-published, biographical information a user published on his/her profile page and information about user lists to which a user belongs. We conducted a user study that explores how useful different types of data are for informing human's expertise judgements. We then used topic modeling based on different types of data to build and assess computational expertise models of Twitter users. We use Wefollow directories as a proxy measurement for perceived expertise in this assessment.**

Our findings show that different types of user-related data indeed differ substantially in their ability to inform computational expertise models and humans's expertise judgements. Tweets and retweets — which are often used in literature for gauging the expertise area of users — are surprisingly useless for inferring the expertise topics of their authors and are outperformed by other types of user-related data such as information about users' list memberships. Our results have implications for algorithms, user interfaces and methods that focus on capturing expertise of social media users.

***Index Terms*—expertise, user profiling, microblogs, Twitter**

I. Introduction

On social media applications such as Twitter, information consumption is mainly driven by social networks. Therefore, judging topical expertise of other users is a key challenge in maximizing the credibility and quality of information received. Recent research on users' perception of tweet credibility indicates that information about the authors is most important for informing credibility judgments of tweets [1]. This highlights that judging the credibility and expertise of Twitter users is crucial for maximizing the credibility and quality of information received. However, the plethora of information on a Twitter page makes it challenging to assess users' expertise accurately. In addition to the messages a user authored (short *tweets*) and re-published (short *retweets*), there is additional information on the Twitter interface that could potentially inform expertise judgements. For example, with fewer than 160 characters, the biographical section (short *bio*) may contain important information that indicates users' expertise level, such as his/her self summarized interests, career information, and links to his/her personal web page. Another feature of Twitter that could potentially be useful for assessing users' level of expertise is the support of user lists (short *lists*). User lists allow users to organize people they are following into labeled groups and aggregate their tweets by groups. If a user is added to a list, the list label and short description of the list will appear on his/her Twitter page. Unlike bio information, which may contain self-reported expertise indication, users' list memberships can reflect external expertise indications, i.e., followers' judgements about one's expertise. However, little is known about the motivations of users for adding other users to lists and the type of information which is revealed by users' list memberships, their bio section and tweet and retweets published by them.

This paper aims to shed some light on the usefulness of different types of user-related data (concretely we use *tweets*, *retweets*, *bio* and *list* data) for making sense of the domain expertise of Twitter users. We use Wefollow[1] directories as a proxy measurement for perceived expertise in this assessment. Wefollow is an application that allows Twitter users to register themselves in a maximum of 3 topical directories. Although Wefollow directories may not provide perfect ground truth for perceived expertise, canonical ranking and social judgments by peers are commonplace for identifying expertise [2]. We assume the way that Wefollow functions, by ranking users according to the number of followers in the same field, is a reflection of such social judgment. Our assumption is supported by previous research which has shown that the majority of the top 20 Wefollow users for selected directories were perceived as experts for the corresponding topic [3] and that experts tend to agree that users with high Wefollow rank are more knowledgeable than users with low or no Wefollow rank [4]. We leverage these findings for our study which aims to address the following research questions:

[1]http://wefollow.com

978-0-7695-4848-7/12 $26.00 © 2012 IEEE
DOI 10.1109/SocialCom-PASSAT.2012.30

1) *What type of user-related social media data is most useful for informing human's expertise judgements about Twitter users?*
2) *Do different types of user-related social media data lead to similar topical expertise profiles of Twitter users or are these profiles substantially different?*
3) *What type of user-related social media data is most useful for creating topical expertise profiles of Twitter users?*

We approached this question from two complementary perspectives. First, we conducted a user study to explore how useful different types of data are for *informing participants' expertise judgements*. Second, we investigated how useful different types of user-related data are for informing *computational expertise models of users*, which represent each user as a set of topic-weight pairs where a user is most knowledgeable in the topic with the highest weight. We used standard topic modeling algorithms to learn topics and annotate users with topics inferred from their *tweets*, *retweet*, *bio* and *list* memberships, and compared those topic annotations via information theoretic measures and a classification task.

Our findings reveal significant differences between various types of user-related data from an expertise perspective. The results provide implications that are not only relevant for expert recommender algorithms in the context of social media applications, but also for user interface designer of such applications. Although our experiments are solely based on Twitter, we believe that our results may also apply to other micro-blogging applications, and more broadly, to applications that allow users to create and organize their social network and share content with them.

The remainder of this paper is structured as follows: In Section 2 we discuss related work on modeling expertise of social media users. Section 3 describes our user study on how humans perceive and judge domain expertise of Twitter users. In Section 4 we present our experiments on modeling perceived expertise of Twitter users. We discuss our results in Section 5 and highlight implications of our work in Section 6. Section 7 describes limitations of our work and discusses ideas for future work. Finally, we conclude our work in Section 8.

II. Related Work

A widely used approach for identifying domain experts is peer-reviews [2]. Many state of the art expertise retrieval algorithms rely on this idea and often use content of documents people create, the relations between people, or a combination of both. For example, in [5] the authors use generative language models to identify experts among authors of documents. In [6] the authors explore topic-based models for finding experts in academic fields. The work presented in [7] uses network analysis tools to identify experts based on the documents or email messages they create within their organizations. In [8] the authors propose a probabilistic algorithm to find experts on a given topic by using local information about a person (e.g., profile info and publications) and co-authorship relationships between people. While previous research often neglects the variety of different types of data that can be observed for social media users, we focus on comparing different types of user-related data from an expertise perspective.

One of the key challenges of expert search algorithms is to accurately identify domains or topics related with users. Topic models are a state of art method for learning latent topics from document collections and allow annotating single documents with topics. Standard topic models such as LDA [9] can also be used to annotate users with topics e.g. by representing each user as an aggregation of all documents he/she authored. More sophisticated topic models, such as the Author Topic (AT) model [10] assume that each document is generated by a mixture of its authors' topic distributions. The Author Persona Topic (APT) model [11] introduces several personas per author because authors often have expertise in several domains and therefore also publish papers about different topics. The Author Interest Model [12] is similar to the APT model except that the personas are not local (i.e. not every user has an individual local set of personas) but global (i.e. all users share a common set of personas). In our work we do not introduce a new topic model, but empirically study how, and to what extent, existing topic modeling algorithms can be used to model the perceived expertise of Twitter users. Although our work is not the first work which applies topic models to Twitter (see e.g. [13] or [14]), previous topic modeling research on Twitter only took tweets into account, while we systematically compare different types of data that can be observed for Twitter users.

Recently, researchers started exploring different approaches for identifying experts on Twitter. For example, in [15] the authors present TwitterRank, an adapted version of the topic sensitive PageRank, which allows identifying topical influential Twitter users based on follow relations and content similarities. In [16] the authors compare different network-based features and content/topical features to find authoritative users. To evaluate their approach they conducted a user study and asked participants to rate how interesting and authoritative they found the author and his/her tweets. The work of [3] presents an approach to find topical relevant Twitter users by combining standard Twitter text search mechanism with information about the social relationships in the network and evaluate their approach via Amazon Mechanical Turk. Previous research agrees on the fact that one needs both, content and structural network features, for creating a powerful expert retrieval algorithm. However, to our best knowledge most existing expert retrieval work on Twitter limits their content features to tweets, while our results suggest that tweets are inferior for making sense of the expertise of Twitter users compared to other types of user-related data.

The issue of how users perceive the credibility of microblog updates is only just beginning to receive attention. In [1] the authors present results from two controlled experiments which were designed to measure the impact of three features (user image, user name and message content) on users' assessment of tweet credibility. Unlike our work, Morris et al. explore which factors influence users' perception of the

credibility of a tweet, while we focus on users' perception of other users expertise. Further, our study sets out to gain empirical insight into the usefulness of different types of data (such as tweets, retweets, user lists and bio information) for informing expertise or credibility judgements of users, while their experiments aim to identify the factors which influence such judgments. That means, while Morris et al. manipulate data (i.e., tweets, user images and user names) within their experiment to measure the impact of their manipulation on users' judgments, we do not manipulate any user-related data, but manipulate the type and amount of data we show. Similar to our results their results indicate that users have difficulty discerning trustfulness based on content alone. In [17] the authors do not examine expertise or credibility per se. In their study they asked users to rate how "interesting" a tweet was and how "authoritative" its author was, manipulating whether or not they showed the author's user name. In our work we decided not to show user names at all amongst others for the following reasons: first, showing user names may add uncontrolled noise to our experiment since participants may recognize some of the users to judge. Therefore their expertise judgments would be based on their background knowledge rather than on the information which is shown to them during the user study. Second, algorithms and automated methods can not exploit user names but will require further information related with those names to gauge users' potential expertise. Since our aim was to create expertise models of users, our experiment set out to evaluate only information which can be accessed and exploited by humans and automated methods.

III. User Study

We conducted a user study to explore how useful different types of user-related social media data are for *informing humans' expertise judgements* about Twitter users. To that end, we compare the ability of participants to correctly judge the expertise of Twitter users when the judgement is based on the contents they published (*tweets and retweets*), self-reported and externally-reported contextual information (*bio* and *user lists*), or both contents and contextual information.

A. Participants

We chose "semantic web" to be the topic in the experiment. We recruited a group of 16 participants consisting of users with rather basic and high knowledge about the topic semantic web. We recruited 8 participants by contacting the faculties and students of the International Summer School on Semantic Computing 2011 held at UC Berkeley and 8 participants from a university town in the United States. Participants' age ranged from 20 to 34.

B. Design and Procedure

We used Wefollow[2] to select candidate Twitter users to be judged. Wefollow is a user powered Twitter directory where users can sign up for at most 3 directories. Wefollow ranks all users based on a proprietary algorithm which takes amongst others into account how many users in a certain directory follow a user. Users who are followed by more users signed up for a topic directory get a higher rank of the particular topic. At the time we crawled Wefollow (July 2011), the Wefollow directory of the topic "semantic web" suggested 276 Twitter users relevant to the topic. For candidates to represent high level of expertise, we randomly selected six users from rank 1–20 and six users from rank 93–113. For candidates of low expertise, we randomly selected six users from rank 185–205 and six users from the public Twitter timeline who did not show any relation to the topic. To validate the manipulation, we also conducted a pilot study by asking 3 raters to compare the expertise of 50 pairs of candidates randomly selected from the high and low expertise group. The results showed that all of them had 95% or higher agreement with our manipulation, and the inter-rater agreement was 0.94. This result proved that our expertise level manipulation was successful.

Our experiment tested three conditions: 1) participants saw the latest 30 messages published by a user (i.e., the user's most recent tweets and retweets) and contextual information including the user's bio information and his/her latest 30 user list memberships ; 2) participants saw only the latest 30 tweets and retweets of a user; 3) participants saw only the bio and the latest 30 list memberships (or all list memberships if fewer than 30 were available). Each of the 24 pages which we selected in step one was randomly assigned to one of the three conditions. In other words, for each condition, we had four Twitter user candidates of high expertise and four Twitter user candidates of low expertise. To tease out the influence of the Twitter interface and further uncontrolled variables such as user images or user names, we presented only the plain textual information in a table. The users' names, profile pictures and list creators' names were removed to avoid the influence of uncontrolled variables. For condition 1 the table had two randomly ordered columns to present tweets and contextual information separately. For condition 2 and 3 the table only had one column to present everything.

Before the task, participants were asked to answer demographical questions and complete a knowledge test. Then they were presented with 24 evaluation tasks (three conditions, eight pages for each condition) in sequence. They were told that the information in the table was derived from a real Twitter user, and asked to rate how much this person knew about the topic, semantic web, on a one (least) to five (most) scale. The tasks took about 30-40 minutes.

C. Results

We analyzed participants' expertise ratings by performing two-way repeated measure ANOVA with Twitter user expertise (high/low) and conditions (content and contextual information/only content/only contextual information) as within subjects variables.

Interestingly, there is an interaction between conditions and Twitter user expertise ($F(2, 30) = 8.326$, $p < 0.01$). It means there exists significant differences in users' ability of differentiating high and low expertise across these three

[2]www.wefollow.com

Fig. 1. Example of the experimental task under condition 1. Randomly ordered tables and plain text without pictures and usernames were used to present different types of user-related data to participants.

conditions. To understand the difference, we compared each pair of conditions by performing the same ANOVA test. When comparing between condition 1, where participants saw both content and contextual information, and condition 2, where participants' expertise judgments were only informed by content, participants were significantly more able to make the correct judgment in condition 1 ($F(1,15) = 23.39$, $p < 0.01$). When comparing condition 3, where participants' judgments were informed by contextual information, to condition 2, where participant's expertise judgments were only informed by content, participants made significantly better judgments in condition 3 ($F(1,15) = 5.91$, $p = 0.03$). There was no significant difference observed between condition 1 and condition 3 ($F(1,15) = 2.19$, $p = 0.16$). These results indicated that participants made the worst expertise judgments when the judgments were based on tweets and retweets only. Interestingly, participants' expertise judgments, when only based on contextual information (i.e., information about users' bio and list memberships), were almost as good as judgments based on both content and contextual information. To illustrate the interaction, we plot participants' average ratings in different conditions in Figure 2. The slopes in Figure 2 reflect the ability of participants to differentiate between Twitter users of high and low expertise in different conditions.

Our findings highlight the low quality of topical expertise judgement based solely on tweets' and retweets' contents. It implies that there is a large variance of information in what people tweet and retweet about. Experts of a particular topic do not necessarily publish or re-published content about the topic all the time, if any. In contrast, contextual information such as bio and user list memberships provides salient and straightforward cues for expertise judgements since these cues often provide descriptive information about the person himself, such as personal interests, professional experience, community the person belongs to, etc.

IV. Experiments

Since our user study supported our hypothesis that different types of user-related data differ in their ability to inform

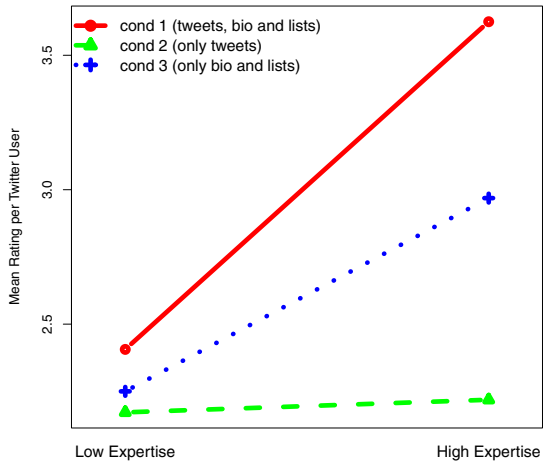

Fig. 2. Average expertise ratings given to Twitter users with high/low expertise by participants in each condition. The slope of each line indicates the ability of participants to differentiate between experts and novices.

humans' expertise judgments we further aim to compare how useful different types of data are for learning *computational expertise models* of Twitter users by using topic modeling. Therefore, we first compare topic distributions of users inferred from different types of user-related data, namely *tweets*, *retweets*, *bio* and *user list* data and study if those topic distributions differ substantially on average. Second, we explore to what extent different topic distributions reflect users' perceived expertise categories by using information theoretic measures and by casting our problem as a user classification task.

A. Dataset

For our experiments we selected the following 10 topics (including rather general and rather specific topics and topics with high and low polarity): semanticweb, biking, wine,

democrat, republican, medicine, surfing, dogs, nutrition and diabetes. For each topic we selected the top 150 users from the corresponding Wefollow directory (i.e., the 150 user with the highest rank). We excluded users whose required information (i.e. tweets, retweets, lists memberships and biographical information) were not available to crawl. We also excluded users who appeared in more than one of the 10 Wefollow directories and users who mainly do not tweet in English. For all remaining 1145 users we crawled at maximum their last 1000 tweets and retweets, the last 300 user lists to which they were added and their bio info. Tweets, retweets and bio information often contain URLs. Since information on Twitter is sparse, we enriched all URLs with additional information (title and keywords) obtained from the meta-tags in the headers of webpages they are pointing to. User list names and descriptions usually do not contain URLs, but list names can be used as search query terms to find web documents which reveal further information about the potential meaning of list labels. We used the top 5 search query result snippets obtained from Yahoo Boss[3] to enrich list information. After enriching our dataset, we removed standard English stopwords and performed stemming using Porter's algorithm [18].

B. Topic Models

Topic models are a powerful suite of algorithms which allow discovering the hidden semantic structure in large collection of documents. The idea behind topic models is to model documents as arising from multiple topics, where each document has to favor few topics. Therefore, each document exhibits different topic proportions and each topic is defined as a distribution over a fixed vocabulary of terms and has as well to favor few words.

Topic models treat our data as arising from a generative process that includes hidden variables. This generative process defines a joint probability distribution over both the observed and hidden random variables. Given this joint distribution one can compute the conditional distribution of the hidden variables given the observed variables. This conditional distribution is also called the posterior distribution.

The most basic topic modeling algorithm, Latent Dirichlet Allocation (LDA) [9], encodes the following generative process: First, for each document d a distribution over topics θ is sampled from a Dirichlet distribution α. Second, for each word w in the document d, a single topic z is chosen according to its document specific topic distribution θ. Finally, each word w is sampled from a multinomial distribution over words ϕ which is specific for the sampled topic z.

Fitting an LDA model to a collection of training documents requires finding the parameters which maximize the posterior distribution $P(\phi, \theta, z|\alpha, \beta, w,)$ which specifies a number of dependencies that are encoded in the statistical assumptions behind the generative process. In our experiments we used MALLET's [19] LDA implementation and aggregated all user-related data into artificial user-documents which we used to train the model. We chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$ and the number of topics T= 10, 30, 50, 100, 200, 300, 400, 500, 600 and 700) and optimized them during training by using Wallach's fixed point iteration method [20]. Based on the empirical findings of [21], we decided to place an asymmetric Dirichlet prior over the topic distributions and a symmetric prior over the distribution of words. Given enough iterations (we used 1500) the Markov chain (which consists of topic assignments z for each token in the training corpus) has potentially converged and we can get estimates of the word distribution of topics ($\hat{\phi}$) and the topic distribution of documents ($\hat{\theta}$) by drawing samples from the chain. The estimated distributions $\hat{\phi}$ and $\hat{\theta}$ are predictive distributions and are later used to infer the topics of users via different types of user-related data. Figure 3 shows some randomly selected sample topics learned via LDA when the number of topics was 50 ($T = 50$).

C. Evaluation Metrics

To answer whether different types of data related to a single user lead to substantially different topic annotations, we compare the average Jensen-Shannon (JS) divergence between pairs of topic annotations inferred from different types of data related with a single user. We always use the average topic distributions inferred via 10 independent runs of a Markov chain as topic annotations. The JS divergence is a symmetric measure of the similarity between two distributions. The JS divergence is 0 if the two distributions are identical and approaches infinity as they differ more and more. The JS divergence is defined as follows:

$$D_{JS} = \frac{1}{2} D_{KL}(A||B) + \frac{1}{2} D_{KL}(B||A) \tag{1}$$

where $D_{KL}(A||B)$ represents the KL divergence between random variable A and B. The KL divergence is calculated as follows:

$$D_{KL}(A||B) = \sum_i A(i) \log \frac{A(i)}{B(i)} \tag{2}$$

To address the question which user-related data are more suitable for creating topical expertise profiles of users, we aim to estimate the degree to which different types of users' topic annotations reflect their perceived expertise. Since we know the ground truth label of all 1145 users in our dataset, we can compare the quality of different topic annotations by measuring how likely the topics agree with our true expertise category labels. Here, we use Normalized Mutual Information (NMI) between users' topic distribution (θ_{user}) and the topic distribution of users' Wefollow directories (θ_{label}) which is defined as the average topic distribution of all users in that directory.

$$NMI(\theta_{label}, \theta_{user}) = \frac{I(\theta_{label}, \theta_{user})}{[H(\theta_{label}) + H(\theta_{user})]/2} \tag{3}$$

[3]http://boss.yahoo.com

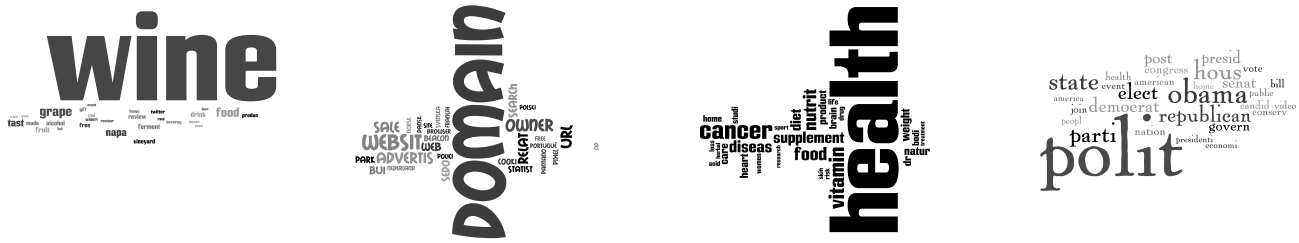

Fig. 3. Top 20 stemmed words of 4 randomly selected topics learned via LDA with number of topics $T = 50$.

$I(\theta_{label}, \theta_{user})$ refers to the Mutual Information (MI), $H(\theta_{label})$ refers to the entropy of the Wefollow-directory-specific topic distribution and $H(\theta_{user})$ refers to a user-specific topic distribution which is inferred based on each of the four different types of user-related data.

$$I(\theta_{label}, \theta_{user}) = H(\theta_{user}) - H(\theta_{user}|\theta_{label}) \quad (4)$$

NMI is always between 0 and 1. A higher NMI value implies that a topic distribution more likely matches the underlying category information. Consequently, NMI is 1 if the two distributions are equal and 0 if the distributions are independent.

Finally, we aim to compare different types of topic annotations within a task-based evaluation. We consider the task of classifying users into topical categories (in our case Wefollow directories) and use tweet-, bio-, list-and retweet-based topic annotations as features to train a Partial Least Square (PLS) classifier[4]. We decided to use PLS, since our features are highly correlated and the number of features can be relative large (up to 700) compared to the number of observations for each trainings split (consisting of 916 users). PLS regression is particularly suited in such situations. Within a 5-fold-cross evaluation we compare the classification performance by standard evaluation measures such as Precision, Recall, F-Measure and Accuracy.

D. Results

In this section, we present our empirical evaluation of perceived expertise models of users based on different types of user activities and their outcomes. Firstly we investigate how similar topic distributions of an individual user inferred from different types of user-related data are on average. Secondly we explore how well different types of topic distributions capture the perceived expertise of users.

First, we aimed to explore whether the topic distributions of a single user inferred from different types of user-related data are differ substantially. Therefore, we compared the average JS divergence of different topic distributions inferred via different types of user-related social media data. Figure 4 shows that different types of user-related data lead to different topic annotations. Not surprisingly, we find that tweet- and retweet-based topic annotations are very similar. Further, bio- and tweet- and bio- and retweet-based topic distributions show

[4]http://cran.r-project.org/web/packages/pls/

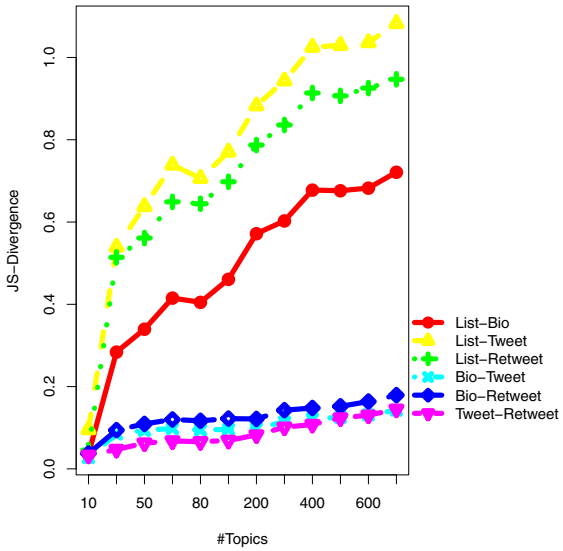

Fig. 4. Average JS-Divergence of 1145 Wefollow users' topic annotations inferred via their tweets, retweets, bio and list information.

high similarity, while list- and bio- and list- and tweet- and list- and retweet-based topic distributions are more distinct. This suggests that users with high Wefollow rank tend to tweet and retweet about similar topics and that they also mention these topics in their bio (or the other way around). Users' list memberships however do not necessarily reflect what users tweet or retweet about or the topics they mention in their bio, amongst others for the following three reasons: First, sometimes user lists describe how people feel about the list members (e.g., "great people", "geeks", "interesting twitterers") or how they relate with them (e.g., "my family", "colleagues", "close friends"). Consequently, these list labels and descriptions do not reveal any information about the topics a user might be knowledgable about. Second, some user lists are topical lists and may therefore reveal information about the topics other users associate with a given user. However, these topical associations can also be informed by exogenous factors, meaning a given user does not necessarily need to use Twitter to share information about a topic in order to be

associated with that topic by other users. Third, since everyone can create user lists and add users to these list, spam can obviously be a problem, especially for popular users.

To get an initial impression of the nature of user list labels and descriptions, we randomly selected 455 lists memberships of 10 randomly selected users (out of our 1145 users) and we asked 3 human raters to judge whether a list label and its corresponding descriptions may reveal information about expertise domains or topics in general. To give an example: list labels such as "my friends" or "great people" do not reveal any information about the expertise of users in that list, while list labels such as "healthcare professionals" or "semanticweb" may help to gauge the expertise of users who are members of that lists. Our results suggest that 77.67% of user lists reveal indeed information about potential expertise topics of users with a fairly good inter-rater agreement ($\kappa = 0.615$).

Second, we explored how useful different types of user-related data are for inferring the perceived expertise of users by estimating how likely the topics agree with the true expertise category labels of users. So far we only know that it makes a difference which type of user-related data we use for inferring topic annotations of users. However, we don't know which types of data lead to "better" topic annotations of users, where better means that a topic distribution captures the perceived expertise of a user more accurately. Since we have a ground truth label of all users in our dataset (their Wefollow directories), we can estimate the quality of different topic annotations by measuring how likely the topics agree with the true category labels. Here, we used the Normalized Mutual Information (NMI) between users' topic distribution based on different types of data and the topic distribution of a users' Wefollow directory which is defined as the average topic distribution of all users in that directory. A higher NMI value implies that a topic distribution might more likely match the underlying category information. Figures 5 shows that list-based topic annotations tend to have higher NMI values than retweet-, tweet- and bio-based topic annotations. It suggests that list based topic annotations reflect the underlying category information best. In other words, users in a given Wefollow directory tend to be in topical similar lists, while the topics they tweet or retweet about or mention in their bio are more distinct. Firstly, this suggests that users assign other users to lists about topics which tend to reflect their self-view, because users have to register themselves for certain topics in Wefollow. Secondly, it indicates that users make these list assignments not only based on the content authored by the users they assign. They also seem to use background knowledge or other types of external information sources to inform their list assignments. As expected, the NMI values become lower with increasing number of topics.

1) User Classification Experiment: To further quantify the ability of different types of topic annotations to reflect the underlying ground truth category information of users, we performed a task-based evaluation and considered the task of classifying users into topical categories such as Wefollow directories. We used tweet-, bio-, list- and retweet-based

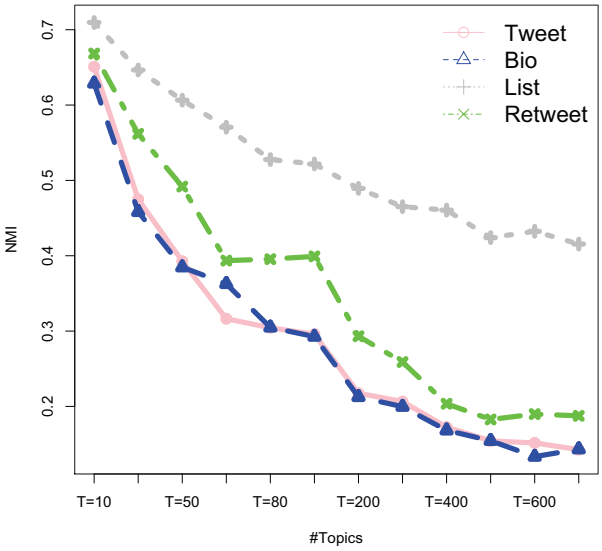

Fig. 5. Average Normalized Mutual Information (NMI) between 1145 users' tweet-, retweet-, list- and bio-based topic annotations and users' Wefollow directory

topic annotations as features, trained a Partial Least Square classifier and performed a 5-fold-cross validation to compare the performance of different trainings schemes. Note that since we used topic distributions as features rather than term vectors the number of features corresponds to the number of topics and does not depend on the length of different user-related data such as bio, tweet, retweet and list data. In other words, the number of features used for tweet-, bio-, list- and retweet-based classifiers were equal although different types of user-related data may differ in their content length.

Figure 6 shows the average F-measures and Accuracy of the classifier trained with different number of topics (T= 10, 30, 50, 70, 80, 100, 200, 300, 400, 500, 600 and 700) inferred via different types of user-related data. One can see from these figures that no matter how fine-grained topics are (i.e., how one chooses the number of topics), list-based topic annotations always outperform topic-annotations based on other types of user-related data.

We also compared the average classifier performance for individual Wefollow directories. Figure 6 shows the average F-measures and Accuracy of the classifier for each Wefollow directory. We averaged the classifier performance for each Wefollow directory over the results we got from the 5-fold cross validations of classifiers trained with different number topics inferred via different types of user-related data for each class. One can see from these figures that for all classes a classifier trained with list-based topic annotations performs best, i.e. yields to higher F-measures and Accuracy values than classifiers trained with other types of user-related data.

However, for certain classes such as democrats or republicans the F-measures of all classifiers are very low, also if trained with list-based topic annotations. It suggests that although list-based topic annotations are best for classifying users into mutual exclusive topical expertise directories, for very similar topics information about users' list memberships might not be detailed enough. For example users which seem to have high knowledge about democrats or republicans, are all likely to be members of similar lists such as "politicians" or "politics".

To explore the classifiers' performance in more detail we also inspected their confusion matrices. Figure 7 shows the confusion matrices of a classifier trained with topic distributions over 30 topics (first row) and 300 topics (second row) inferred via different types of user-related data as features. Note that a perfect classifier would lead to a red map with a white diagonal. The confusion matrix for a classifier trained with list-based topic annotations (Figure 7) shows the closest match to the ideal and hence indicates least confusion. Again, on can see that confusion mainly happens for very similar classes such as democrats and republicans, since those users are likely to be members of similar lists.

V. Discussion

Judging expertise of social media users will continue to represent a relevant and challenging research problem and also an important task for social media users since judging topical expertise of other users is a key challenge in maximizing the credibility and quality of information received.

Through our experiments and our user study, we showed that different types of user-related data differ substantially in their ability to inform computational expertise models of Twitter users and expertise judgements of humans. We argue that these findings represent an important contribution to our research community since in past research topical user profiles are often learned based on an aggregation of all documents a user has authored or is related with, without taking the differences between various types of user activities and related outcomes into account.

Our experiments demonstrate that the aggregation of tweets authored or retweeted by a given user is less suitable for inferring the expertise topics of a user than information about users' list memberships. In addition, our user study clearly confirms that it is as well difficult for humans to identify experts based on their tweets and retweets. Further, our results show that topic annotations based on users' list memberships are most distinct from topic annotations based on other types of user-related data. Topic annotations based on bio information are however surprisingly similar to topic annotations based on the aggregation of tweets and retweets, which indicates that users tend to tweet and retweet messages about topics they mention in their bio or the other way around. This is interesting from a practical point of view, since it suggests that computational expertise models of users which just rely on their bio information achieve similar accuracy as models which are based on the aggregation of their tweets or retweets.

VI. Implications

Our experimental findings suggest that users' have difficulties in judging users' expertise based on their tweets and retweets only. Therefore, we suggest that user interface designer should take this into account when designing users' profile pages. We suspect that Twitter users' profile pages are amongst others used to inform users about the expertise, interests, authoritativeness or interestingness of a Twitter user. Therefore those type of information which facilitates these judgements should be most prominent.

Further, our results suggest that computational expertise models benefit from taking users' list memberships into account. Therefore, we argue that also expert-recommender systems and user-search systems should heavily rely on user list information. Further we argue that also social media provider and user interface designer might want to think of promoting and elaborating list features (or similar features which allow to tag or label other users) more, since user list information seems to be very useful for various tasks.

VII. Limitations and Future Work

The result of our user study is limited to a small subject population and one specific topic, semantic web. Readers who try to generalize our results beyond Twitter should also note that the motivation of users for using a system like Twitter in general and their motivation for creating user lists in specific, may impact how useful information about list memberships are for the expertise modeling task. On Twitter we found that indeed a large percentage of lists may potentially reveal information about the expertise of users assigned to the list. However, this can be different on other social media systems. Nevertheless, our results highlight the potential of user lists and if lists are used for different purpose automated methods can be applied in order to group lists by its purpose.

Our work highlights that different types of social media data reveal different types of information about users and therefore enable different implications. We will explore this avenue of work by investigating which implications different types of activities and related outcomes may enable and how they can be combined for creating probabilistic user models.

VIII. Conclusions

Information consumption on social media is mainly driven by social networks and credibility judgements of content are mainly informed by credibility judgements of authors [1]. Therefore, judging topical expertise of other users is a key challenge in maximizing the credibility and quality of information received. In this work we examined the usefulness of different types of user-related data (concretely we used *tweets*, *retweets*, *bio* and *user list memberships*) for making sense of the domain expertise of Twitter users. Our results suggests that different types user-related social media data are useful for different computational and cognitive tasks, and the task of expertise modeling benefits most from information contained in user lists as opposed to tweet, retweet or bio information. We hope our findings will inform the design

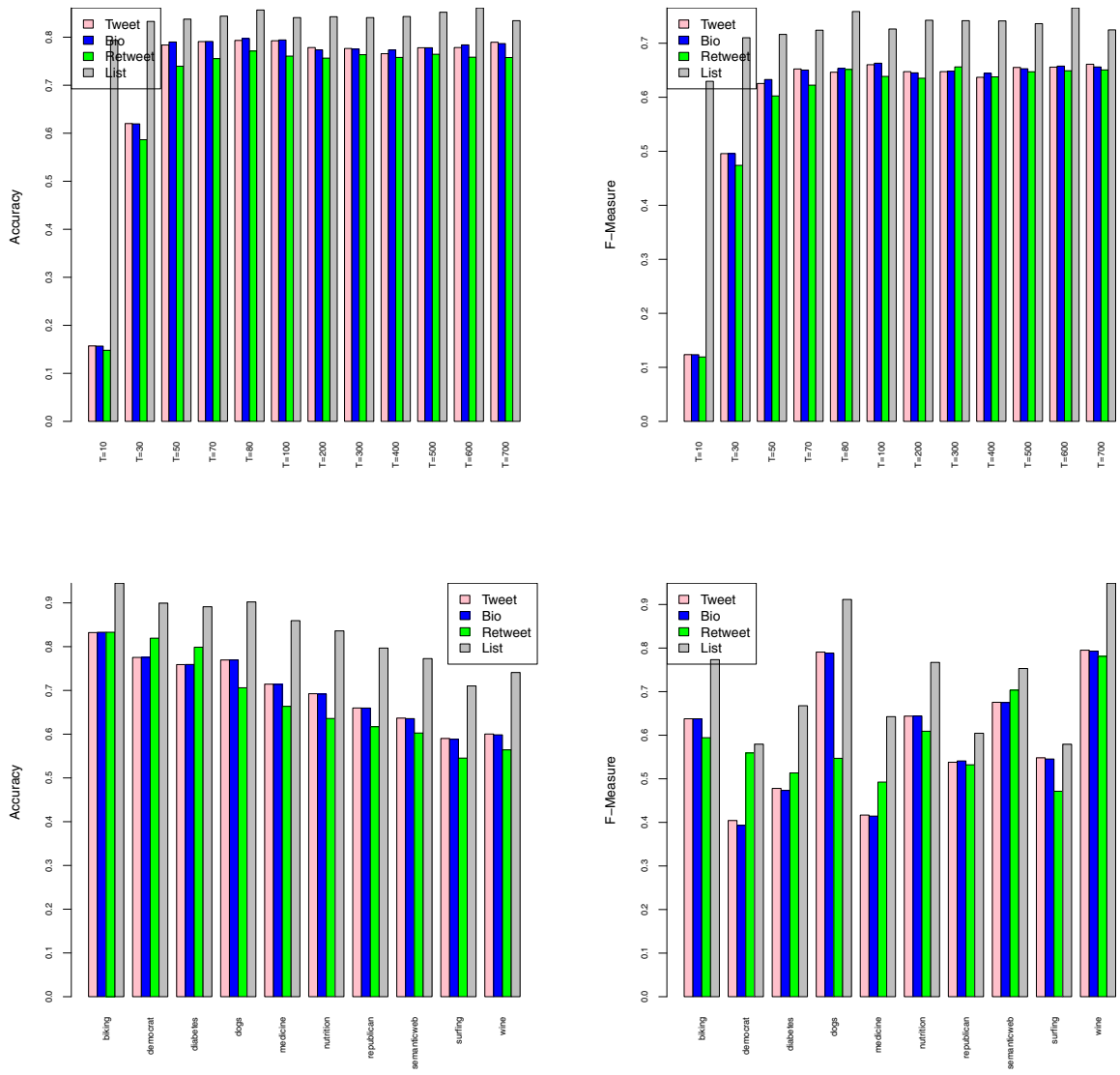

Fig. 6. Average Accuracy and F-measure of PLS classifier trained with bio-, list-, retweet-, and tweet-based topic distributions. The x-axes of the figures in the first row show the number of topics per distributions. The x-axes of the figures in the second row show the 10 Wefollow directories (biking, democrat, diabetes, dogs, medicine, nutrition, republican, semanticweb, surfing, and wine). The y-axes show the accuracy or F-measure of the classifier averaged over 5 folds and different numbers of topics (T=10, 30, 50, 70, 80, 100, 200, 300, 400, 500, 600, 700) or the 10 Wefollow directories.

of future algorithms, user interfaces and methods that focus on capturing expertise of social media users and stimulate research on making sense of different types of user-activities and related outcomes.

ACKNOWLEDGMENT

This work was supported in part by Office of Naval Research Contract No. N00014-08-C-0029 to Peter Pirolli and by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner.

REFERENCES

[1] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ser. CSCW '12. New York, NY, USA: ACM, Feb. 2012, pp. 441–450.

[2] K. A. Ericsson, *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press, 2006, ch. Protocol Analysis and Expert Thought: Concurrent Verbalizations of Thinking during Experts' Performance on Representative Tasks, pp. 223–241.

[3] K. R. Canini, B. Suh, and P. Pirolli, "Finding relevant sources in twitter based on content and social structure," in *NIPS Workshop*, 2010.

[4] Q. Liao, C. Wagner, P. Pirolli, and W.-T. Fu, "Understanding experts'

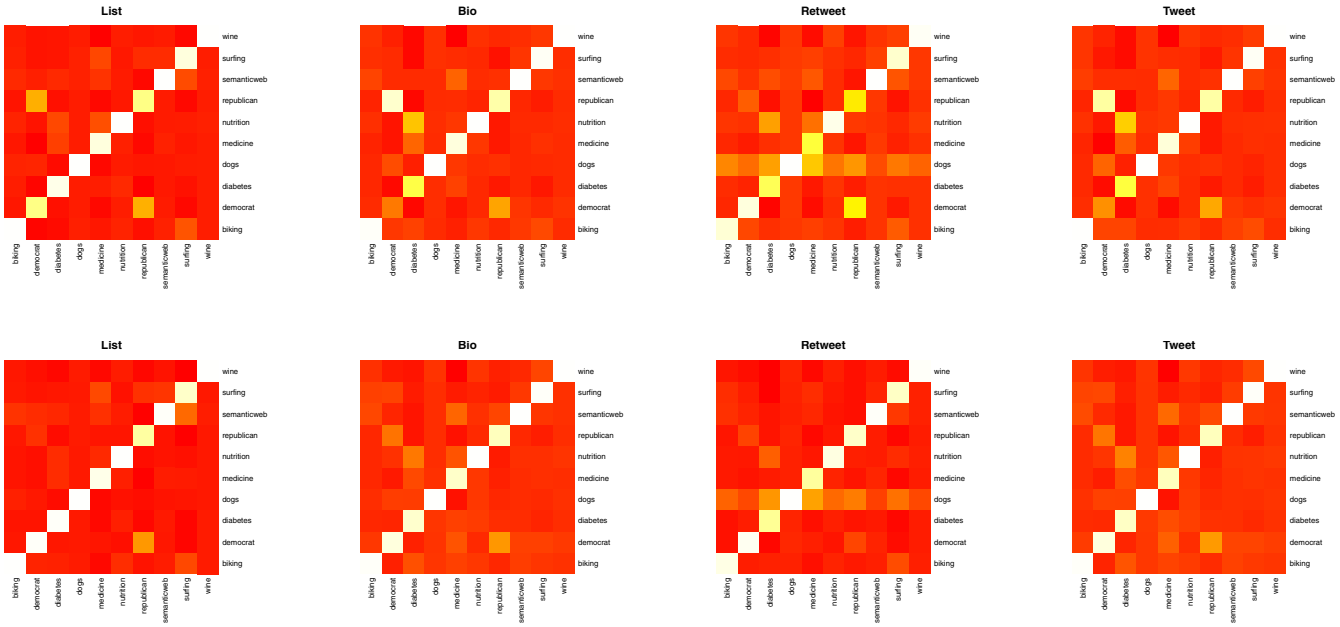

Fig. 7. Average confusion matrices (across 5 folds) of a classifier trained with bio-, list-, retweet-, and tweet-based topic annotations with 30 topics (first row) and 300 topics (second row). The x-axis of each confusion matrix shows the reference values and the y-axis shows the predictions for the10 Wefollow directories (biking, democrat, diabetes, dogs, medicine, nutrition, republican, semanticweb, surfing, and wine). The lighter the color the higher the value.

and novices' expertise judgment of twitter users," in *Proceedings of the 30th ACM conference on Computer-Human Interaction(CHI)*, 2011.

[5] K. Balog, L. Azzopardi, and M. de Rijke, "Formal models for expert finding in enterprise corpora," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '06. New York, NY, USA: ACM, 2006, pp. 43–50.

[6] H. Deng, I. King, and M. R. Lyu, "Formal models for expert finding on DBLP bibliography data," in *Proceedings of International Conference on Data Mining*, 2008.

[7] J. Davitz, J. Yu, S. Basu, D. Gutelius, and A. Harris, "ilink: Search and routing in social networks," in *Proceedings of The Thirteenth International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, August 2007.

[8] J. Li, J. Tang, J. Zhang, Q. Luo, Y. Liu, and M. Hong, "Eos: expertise oriented search using social networks," in *Proceedings of the 16th international conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 1271–1272.

[9] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[10] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, ser. UAI '04. Arlington, Virginia, United States: AUAI Press, 2004, pp. 487–494.

[11] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *KDD*, 2007.

[12] N. T. T. Comware, "Author interest topic model," *Notes*, pp. 887–888, 2010.

[13] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[14] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 80–88.

[15] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 261–270.

[16] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 45–54.

[17] A. Paland and S. Counts, "What's in a @name? how name value biases judgment of microblog authors." in *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press, 2011.

[18] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[19] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[20] H. M. Wallach, "Structured topic models for language," Ph.D. dissertation, University of Cambridge, 2008.

[21] H. M. Wallach, D. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in *Proceedings of NIPS*, 2009.