Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Expertise in Social Networks: How Do Experts Differ From Other Users?

Benjamin D. Horne, Dorit Nevo, Jesse Freitas, Heng Ji & Sibel Adalı
Rensselaer Polytechnic Institute
110 8th Street, Troy, New York, USA
{horneb, nevod, freitj, jih, adalis}@rpi.edu

Abstract

Expertise location is a difficult task, with expertise often being implied and liable to change. In this paper we propose a heuristic-based approach for automated identification of expertise on Twitter. We collect tweets from experts and non-experts in different domains and compute different types of features based on the heuristics regarding properties of the messages written and rewteeted by the experts. We show that these heuristics provide us with interesting insights regarding how experts differ from other user groups which can help guide future studies in this areas and algorithms for expertise location.

Introduction

Expertise location, especially in computer-mediated environments, is a challenging task, with expertise often being tacit and highly volatile. The location and selection of experts can be greatly aided by the use of heuristics, in particular, meta-representation of what we know about the expert. An important advantage of heuristics is that meta-knowledge tends to be less volatile than knowledge itself, making it easier to keep track of expertise.

In this paper, we propose a heuristic-based approach for automated identification of expertise on Twitter. Our main premise is that experts will use Twitter differently than non-experts, and therefore, by identifying expertise heuristics, we can enable expertise location and tackle the following main research question: *How are experts different in how they use Twitter? Are there heuristics to locate experts on Twitter?*

To address this problem, we chose to first identify Twitter users who are considered experts in a specific topic. We have then used these users to find related groups of users who are interested in the same topics. We also collected friends of our experts and users who are mentioned by experts in messages. Given the messages by these users, we computed features of increasing complexity. We have augmented frequently used behavioral features for Twitter with stylistic features borrowed from the literature on readability of messages and linguistic features that measure the complexity of the sentence structure. We also modeled the complexity of the concepts used by borrowing from the fluency heuristic from psychology (the more common concept is likely the correct choice in a decision). Using features based on these heuristics, we compared experts to all the other user groups.

Our results show that all three groups of heuristics are significant in separating experts from other user groups. In fact, using our heuristics, we arrive at an interesting insight. Experts differ both from friends and mentioned users. Experts tend to receive information from many friends, filter and distill it. They have fewer followers than friends, but more followers than mentioned users. They use slightly less complex language than friends, but more complex language than mentioned users. As a result, mentioned users are consumers of information from experts with smaller networks and stylistically different messages. We also show that heuristics have significant predictive power in distinguishing different groups of users.

Based on our analysis, we argue that use of heuristics provide a viable way to categorize different types of users and characterize experts within a given domain. These heuristics are cheap to compute and do not require a lot of contextual information or network data. In future work, we intend to extend our data set with more topics, experts and heuristics, and consider how our methods can be used to enhance expert search methods.

Related Work

The problem of expertise search, finding an expert on a given topic, has been studied in great detail. Often the underlying need is to get access to high quality non-codified information with little search cost. Expertise location has been studied within large organizations with emphasis on team formation using detailed performance and network data available within the organization (Guy et al. 2003; Yogev et al. 2015). The main problem we study is understanding how expertise is signaled to others based on heuristics (Li and Shan 2013). Work in this field falls into analysis of the complexity of concepts used by experts (Balog, de Rijke, and Weerkamp 2008), the network location of experts (Zhang, Ackerman, and Adamic 2007) and a combination of these two by looking at links between information and individuals who produce or endorse them (Ghosh et al. 2012)(Noll et al. 2009)(Weng et al. 2010). A different approach to searching expertise is through the use of heuristics,

Table 1: Resources used to locate experts

Topic	Resource
Science	`http://www.teachthought.com/learning/100-scientists-on-twitter-by-category/`
Technology	`http://www.hashtags.org/entertainment/celebrities/10-tech-experts-you-should-follow-on-twitter/`
Technology	`http://www.yfncg.com/2009/05/11/100-tech-twitter-accounts/`
Health	`http://diettogo.com/blog/15-health-and-wellness-experts-you-should-be-following`
Health	`http://greatist.com/health/must-follow-health-and-fitness-twitter-accounts-2012`
Health	`http://www.stack.com/a/twitter-fitness-experts`
Business	`http://www.huffingtonpost.com/vala-afshar/twitter-business-tech_b_2355700.html`

rules of thumb used to support decision making, which have been acknowledged as useful in expertise location (Nevo, Benbasat, and Wand 2012). Despite this increased focus on their role and importance, little empirical work exists that identifies which heuristics are relevant in locating expertise. Aiming to close this gap, we identify expertise heuristics on social media that can be used to support the location of expertise.

Methods

We first identified a set of experts by using the the social acclamation approach, relying on the identification of experts by others in their field (Shanteau et al. 2002) and viewing recognition by others in this field "as having the necessary skills and abilities to perform at the highest level". We collected tweets from experts in four topics, *Science*, *Technology*, *Health&Fitness* and *Business* using online sources given in Table 1. We manually chose all the active users from these lists with at least 1000 messages who mainly post in English.

Then, we used a hierarchical sampling method to find users who are interested in the same topics by finding the last (up to 3200) messages by our experts. We then found users *mentioned* by experts in their messages and also users followed by experts (*friend*). We then found the commonly used hashtags by the experts each group (excluding common non-domain tags, about 70 hashtags for each domain) and collected users who used the same hashtags. We grouped these users into low/medium/high hashtag usage based on the statistical distribution of usage. We included the middle range, *mediumtag* in our analysis and excluded the others as outliers. For each of our four sets of users, we collected messages using the Twitter streaming API for a period of 3 months.

We collected three basic types of features for the messages collected for each user, shown in Table 2. We concentrated on heuristics that are relatively simple to compute and would work with limited or incomplete data. These features do not require large scale collection of network or topic based data. The *behavioral* features use very little semantic analysis. The last three features in this set are normalized to the energy of individuals, while the first four features are global and are used in the log scale.

The stylistic features are based on deeper natural language processing of the sentence structure and grammatical elements, borrowing from the work on readability analysis (Feng et al. 2010) with the hypothesis that experts produce messages that are more readable by a large audience. We restricted our features to those that work well for the short messages encountered on Twitter. We also extracted linguistic features to measure the complexity of the sentences and the concept, which are first computed for each message of the user and then averaged over all their messages. All features were computed first for original messages by the users only. We also compute a second set of features for the messages retweeted by the users. This second set is analyzed separately. The final dataset with the features for all different user groups is available at:`https://github.com/rpitrust/expertisedataset`.

Data were analyzed using simple logistic regression with the dependent variable being the group membership of users. To tease out difference between the experts group and all other groups we opted to compute separate pairwise binary models comparing comparing experts to friends, experts to users mentioned by them, and experts to the medium hashtag group. Data were cleaned to eliminate missing values and any highly correlated features. We used the log of followers, friends, and number of tweets to reduce the impact of outliers. We conducted this analysis first for original messages and then for retweeted messages, (Tables 3 (a) and (b) respectively) and also show general descriptive of the distribution of feature values across groups in Figure 1.

Results and Conclusions

We find that the number of years a user has been in Twitter, number of friends and number of followers are always significant in distinguishing experts for original messages. We included them in our analysis of retweeted messages as a control. Overall, all three types of heuristics are useful in identifying experts both original and retweeted messages, however fewer heuristics are effective in distinguishing the types of messages retweeted by experts.

We note that experts tend to be older Twitter users than the other groups, partially due to bias in our data. As our data is based on peer judgments, the likelihood of being known by others' is higher for older Twitter users. Despite this fact, experts tend to follow more users, but have fewer followers than their friends. In essence, one can view friends of experts as news sources that experts collect information from. Experts distill this information and produce their own opinion. This means that experts' friends are not necessarily experts. This trend is reversed for mentioned friends: experts have more followers and follow more users than mentioned users. In other words, mentioned users are consumers of information provided by experts. Mentioned users serve smaller and possibly different communities. Experts sit between the two groups of users: information publishers and information consumers, reducing and interpreting information be-

Table 2: Different features used in our study

Abbr.	Description
friends	# friends (log scale)
followers	# followers (log scale)
total_tweets	# tweets in Twitter (log scale)
years	# years a user has been on Twitter
mentpermsg	# mentions per original msg
tagpermsg	# hashtags per original msg
urlpermsg	# urls per original msg
p_org	% msgs that are original (not retweet)
p_msgwment	% of msgs with at least one mention
p_msgwtag	% of msgs with at least one hashtag
p_msgwurl	% of msgs with at least one url

(a) Behavioral Features

org_med_d	median depth of syntax tree
org_med_d_np	median depth of noun phrase tree
org_med_d_vp	median depth of verb phrase tree
lexco	user's lexical coherence
flu_min	frequency of least common word used
flu_min3	avg. frequency of least common 3 words

(b) Linguistic Features

Abbr.	Description
char	# chars
punc	# punctions
comma	# commas
period	# periods
quesmark	# question marks
exmark	# exclamation mark
semi	# semicolon
colon	# colon
first	# first person terms
vp	# verb phrases
np	# noun phrases
interj	# interjections (e.g. yeah, uh)
stop	# stop words (e.g. a, the, to)
slang	# slang terms (e.g. omg, :))
senti	# sentiment terms
entity	# entities
dt	# definite articles (e.g. the, those, ..)
modal	# modal verbs (e.g. may, should, could)
neg	# negative verbs (e.g. never, cannot)
active	# active verbs (i.e. not passive tense)

(c) Stylistic Features

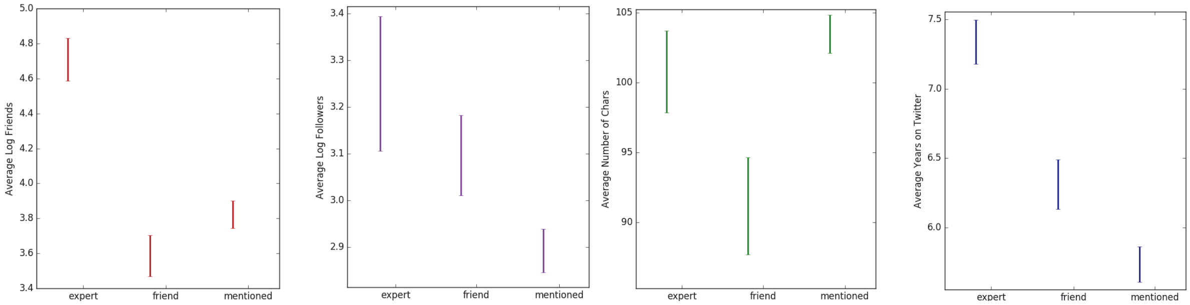

Figure 1: 95% Confidence Interval of the four behavioral features for different user groups.

tween them. The semantic features shows that experts use slightly less specialized language (*flu_min*) and simpler sentences (*org_med_d_vp*) than friends, but significantly more specialized language than mentioned users (*flu_min*). Mentioned users have a great deal of stylistic differences than experts: they use fewer commas and exclamation points, more slang and refer to more entities. The mediumtag group resembles the mentioned user group, but has fewer characteristic stylistic features than mentioned group. Other differences are amplified. We have found that experts in our group are not heavy users of hashtags and neither are the friends. The tag usage increases from experts to mentioned users, and also from mentioned users to the users in the medium tag group.

Retweeted messages of experts, friends and mentioned users are more similar to each other than hashtag users, partially due to our sampling methodology that favored users connected to each other. Despite this similarity, experts retweet longer messages than their friends likely containing many short URLs common to news sources. Similarly, experts tend to retweet messages that contain more complex sentences (*org_med_d_np*) than their friends and mentioned users.

Since we are working with a relatively small group of experts in our data set, we did not partition it for both training and testing. We used the group of users who were mentioned by the experts (n= 1000) and separated 300 random users as test group. As a second group (non-mentioned), we took a random sample of equal size to our test group from the medium hashtags users. We again developed a logistic regression model to estimate the likelihood of falling in the mentioned group, using our training data. We then used this model to make a prediction using the testing data. Our model proved to be very good at predicting the likelihood of a person falling in the mentioned group as opposed to the hashtag group, with an 80.5% accuracy.This prediction was made solely based on the features collected and discussed in this paper.

Our results show that heuristics can be used to help distinguish users who are perceived to be experts by others. Our

Table 3: Significant features that distinguish experts from user groups: B: behavioral, S: stylistic and L: linguistic. A positive sign means a high value increases the likelihood that the user is an expert and a negative sign means a high value decreases the likelihood the user is an expert (significance codes: 0 (***) 0.001 (**) 0.01 (*) 0.05 (.) 0.1).

			User Groups	
Type	Feature	friend	mentioned	mediumtag
B	total_tweets	+**		−.
B	friends	+***	+***	+***
B	followers	−**	+**	−.
B	years	+***	+***	+***
B	p_msgwment	+*		+***
B	p_msgwtag		−**	−***
B	p_org	+**		
S	char			
S	comma	+.	+*	+**
S	exmark		+*	
S	punc		−**	
S	interj			
S	slang		−**	
S	np			+*
S	entity	−*	−**	
S	active	−.		
L	org_med_d_np			+.
L	org_med_d_vp	−.		
L	flu_min	+.	−**	

(a) Features computed from original messages

		User Groups	
Feature	friend	mentioned	mediumtag
total_tweets	+.		+*
friends		+***	+*
followers	+***	+***	+***
years	+***	+***	+***
p_msgwment			+*
p_msgwtag		−*	
p_org			+*
char	+***		
comma			
exmark		−*	−**
punc			+.
interj			+**
slang		+*	+.
np			
entity			+.
active		+*	
org_med_d_np	+**	+**	
org_med_d_vp			+.
flu_min	+.		

(b) Features computed from retweeted messages

heuristics further show that experts consume information from friends and feed it to other users, such as users mentioned by them. Understanding the significant differences between users groups we found and using them in designing more refined expertise search methods is part of our future research.

Acknowledgments

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Balog, K.; de Rijke, M.; and Weerkamp, W. 2008. Bloggers as experts: feed distillation using expert retrieval models. In *SIGIR*, 753–754.

Feng, L.; Jansche, M.; Huenerfauth, M.; and Elhadad, N. 2010. A comparison of features for automatic readability assessment. In *COLING*, 276–284.

Ghosh, S.; Sharma, N.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. P. 2012. Cognos: crowdsourcing search for topic experts in microblogs. In *SIGIR*.

Guy, I.; Avraham, U.; Carmel, D.; Ur, S.; Jacovi, M.; and Ronen, I. 2003. Mining expertise and interests from social media. In *WWW*, 515–526.

Li, C. T., and Shan, M. K. 2013. X2-search: Contextual expert search in social networks. In *TAAI*, 176–181.

Nevo, D.; Benbasat, I.; and Wand, Y. 2012. Understanding technology support for organizational transactive memory: Requirements, application, and customization. *Journal of Management Information Systems* 28(4):69–98.

Noll, M.; Yeung, A.; Gibbins, N.; Meinel, C.; and Shadbolt, N. 2009. Telling experts from spammers: expertise ranking in folksonomies. In *SIGIR*, 612–619.

Shanteau, J.; Weiss, D.; Thomas, R.; and Pounds, J. 2002. Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research* 136(2):253–263.

Weng, J.; Lim, E. P.; Jiang, J.; and He, Q. 2010. Twitterrank: Finding topic-sensitive influential twitterers. In *WSDM*.

Yogev, A.; Guy, I.; Ronen, I.; Zwerdling, N.; and Barnea, M. 2015. Social media-based expertise evidence. In *ECSCW*. Springer International Publishing. 63–82.

Zhang, J.; Ackerman, M.; and Adamic, L. 2007. Expertise networks in online communities: structure and algorithms. In *WWW*, 221–230.