Tweeting Questions in Academic Conferences: Seeking or Promoting Information?

Xidao Wen, University of Pittsburgh
Yu-Ru Lin, University of Pittsburgh

Abstract
The fast growth of social media has reshaped the traditional way of human interaction and information seeking behavior, which draws research attention on characterizing the new information seeking paradigm. However, results from previous studies might not be well grounded under certain social settings. In this paper, we leverage machine learning techniques to identify different types of question tweets within academic communities as an example of one particular social context. By studying over 160 thousands of tweets posted by 30 academic communities, we discovered a different landscape of information-seeking behaviors, where less tweets are regarded as question tweets, and more real information-seeking tweets are observed. We also found that users respond differently to different types of question tweets. We believe our study would be beneficial for understanding the information seeking behaviors in social media.
Keywords: Twitter; Information-Seeking; Classification; Scholars; Data Mining

1 Introduction

Social media, like Twitter and Facebook, have enriched people's everyday social life. The immediacy and reachability of these platforms have facilitated people's information seeking. In particular, Twitter has emerged into a place enabling "social search" where, rather than formulating a query via search engines, users simply "tweet" a question to one or more people in their social network (Liu & Jansen, 2012).

Researchers have begun to understand such new information-seeking paradigm. For examples, Liu and Jansen found that social information seeking exhibits more personalized requirements and more timely needs (Liu & Jansen, 2012). While these studies showed interesting results, there is a lack of coherent understanding about information seeking behavior on Twitter. For example, studies have shown different results in the relationship between social search and direct conversation (Liu & Jansen, 2012; Efron & Winget, 2010). The challenges lie in several aspects. First, most of the research relied on a limited sample of users or tweets and it is unclear how the information-seeking patterns may vary across different social contexts. Second, although social context is central in social search, prior studies did not track a group of users but instead relied on randomly sampled tweets. This resulted in diffuse social contexts in the collected data and hence it is difficult to establish a concrete understanding of information-seeking behavior within certain social contexts.

In this project, we take the first initiate to investigate large-scale information-seeking patterns on Twitter within academic communities. We collected 166,332 tweets posted in 30 academic conferences over five years (2009 to 2013). We leveraged machine learning technique (Li, Si, Lyu, King, & Chang, 2011; Zhao & Mei, 2013) to automatically identify the signals of information seeking from our corpus. We found there is less proportion of tweets framed as questions, of which there are more tweets with real information seeking needs. We also discovered the different responses to the information-seeking question tweets and non-information-seeking question tweets where the latter proportion is more likely to be favored or retweeted, although no preference is observed in replying to either of them. We discuss the implication of our study and future work.

2 Data Collection

We collected data to understand how Twitter is used for information seeking during academic conferences. We utilized the Computer Science Conference rankings provided at CORE[1] to get a list of conferences. We also obtained the acronym, fields and the tier information of each conference. The tier information ranges from A* (flagship conference), A (excellent conference), to B (good conference). A convention widely used in many academic conferences is to post conference-messages on Twitter through official hashtags – they are typically composed by combining its acronym and the year, e.g., #WWW2012 or #WWW12. Hence, we use the TOPSY API[2] to crawl the conference tweets by searching for the conference hashtag as the keyword and limiting the period to be two weeks before conference and two weeks after. We manually examine the tweets retrieved through the list of hashtags were actually posted within the conferences of interest. After removing the noise conference tweets, we obtained 166,332 tweets from 30 conferences between 2009 and 2013.

3 Question Identification

In this study, we applied the approach proposed in (Zhao & Mei, 2013) as they achieved the best accuracy and made their dataset publicly available[3], to our best knowledge. This approach includes two steps: First, we extracted the tweets with question mark as "qweets." Then, we trained binary classifiers based on different sets of features to identify whether a given qweet contains information-seeking needs or not. Our best classifier was built based on a combination of top 2610 most influential lexical features and top 430 most influential POS (Part-of-Speech Taggers) features using naive Bayes, which reached 81.6% in accuracy (*precision=0.70, recal l=0.95, AUC=0.84, 10-fold cross validation*). Table 1 listed several examples in each category identified by the classifier. By revisiting the criteria of information-seeking question in (Zhao & Mei, 2013), where information-seeking qweets are tweets that expect an informational answer, we noticed that Q2 and Q6 are actually mis-classified, while the rest closely align with the definition. Further study is needed to understand how well the classifier works on our dataset and how to improve the classification performance.

Information-Seeking qweets	Non-Information-Seeking qweets
Q1: "Saw the Business think tank this Morning; good speakers. What should we see now? #SIG- GRAPH2011" **Q2**: "Curious what some of #kdd2012 ses- sions will be? Check out the video pitches: http://t.co/X19IXeHS" **Q3**: "@recsys2010 where is the Banquet tonight? Can you post the address in case one want to walk there? #recsys2010"	**Q4**: "Participating in #wise2010 in Qatar but sit- ting in Manchester thanks to technology ,twitter and live stream- a model for learning?" **Q5**: "Interested in "process discovery"? Here is a cool game: http://t.co/WeCx4ZMX (go and beat Ingo Weber's high score!) #bpm2012" **Q6**: "Will there be some sort of video coverage for #ht2012? "

Table 1: Examples of the questions in each category.

4 Question Distribution

Figure 1 shows the distribution of conference tweets. Elite conferences, like *CHI*, *WWW* and *SIGGRAPH*, present the largest amount of tweets, followed by excellent conferences, as *WISE*, *CSCW*, which is rather expected given the amount of attention received from scholars on Twitter and in general. The classificaiton resulted in two type of qweets: information-seeking qweets (real question tweets) and non-information-seeking qweets. We plotted the proportion of information-seeking qweets in the question tweets in addition to the proportion of qweets in all the tweets. Notice that the qweets take less proportion compared to 13% as

[1] http://www.core.edu.au/

[2] http://topsy.com
[3] http://www-personal.umich.edu/ zhezhao/projects/IN/labels

reported in (Efron & Winget, 2010), whereas the conference with the largest proportion of qweets (*ECIS*) from our dataset has only about 10.85%. We suspect the difference can be the result from different question identification processes and can also be due to the different characteristics of the datasets in each study. The proportion of the information-seeking qweets in the qweets varies from one conference to another with a mean at 41.1% and a median at 40%, which is higher than 28.6% reported in (Zhao & Mei, 2013). We believe the difference is largely due to the nature of how our dataset was collected, which was centered around academic communities during conferences. We also noticed that in some of the conferences this proportion is rather low. For example, in *VLDB*, the proportion is only 29.7%. We observed from our data, there has quite a proportion of rhetorical questions from *VLDB*, one example is that "*Data lovers,did u miss the Very Large DataBase event in Riva del Garda?don't worry here there are #VLDB2013 keynotes http://t.co/aHgi8J2dUT*". Rather than expecting any informational answers, the purpose of this tweet was to promote talks in the community. In fact, this type of non-information-seeking qweets appears across the conferences. Simply regarding all types of non-information-seeking qweets as a whole might overlook some important aspects of the communication during conferences. A finer-grained categorization is needed in order to understand the whole picture of scholars communication.
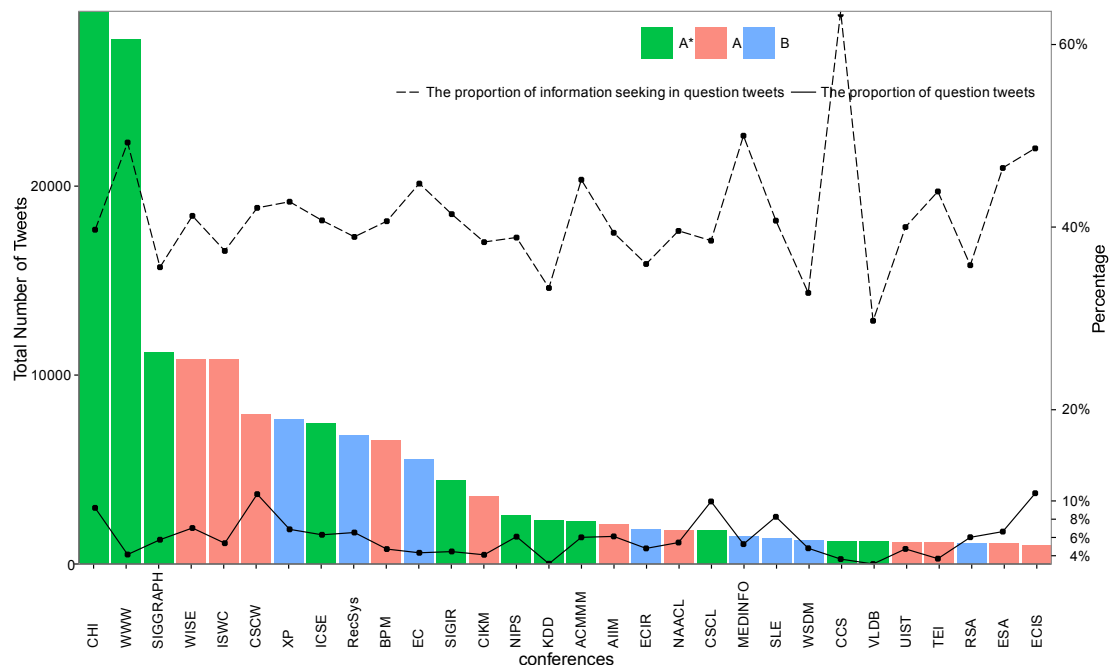

Figure 1: Distributions of total tweets, question tweets and information-seeking question tweets.

5   Question Responses

We seek to examine whether a question tweet that has information needs would be more likely to receive attentions than one does not. In twitter, there are three types of the response that one tweet might receive: replies, retweets, and favorites.

We then compared the number of qweets that have been replied with the number of qweets that have not. The result is shown in Table 2. The odds ratio ($OR_{replied}$) of information-seeking qweets with replies to non-information-seeking qweets with replies is 1.12 with a *p*-value of 0.098. Hence, there is no evidence for an association between the existence of information needs and whether they would have replies.

We then examined the different retweeting behavior towards different types of qweets. Table 3 shows the result. The odds ratio ($OR_{retweeted}$) is 0.83 with a *p*-value of 0.0002. Hence, there is

evidence of a negative association observed between the information-seeking qweets and the retweeting response, indicating that a question tweet without information needs is more likely to be retweeted than one with information

	Replied	¬Replied
Information-Seeking	427	4476
¬Information-Seeking	553	6480

Table 2: There is a positive association between information-seeking qweets and whether they got replied

$(OR_{replied}=1.12$, p-value=0.098). However, the association is not statistically significant.

needs.

	Retweeted	¬Retweeted
Information-Seeking	800	4103
¬Information-Seeking	1335	5698

Table 3: Association between information-seeking and whether they got retweeted is negative and significant
$(OR_{retweeted}=0.83$, p-value=0.0002).

We also studied the association between information-seeking qweets and whether they would be favored. Table 4 shows the result. The odds ratio ($OR_{favored}$) of information-seeking qweets that got favored at least once to non-information-seeking qweets that got favored is 0.82 with a p-value of 0.003. Thus, there is evidence of a negative association between the question type and the favoring response. Together with the negative association between information-seeking qweets and retweeting, these results suggest non-information qweets, which tend to be used in promoting information, are more likely to get response in terms of retweeting and favoring.

	Favored	¬Favored
Information-Seeking	396	4507
¬Information-Seeking	679	6354

Table 4: Association between information-seeking and whether they got favored is negative and significant
$(OR_{favored}=0.82$, p-value=0.003).

6 Discussions and Future Work

In this paper, we leveraged the state-of-the-art classification technique to identify real question tweets posted during academic conferences in Computer Science. We then studied the scholars' information seeking behaviors and how their tweets got responded. We found that more than half of the question tweets are indeed not questions in almost all the conferences. People react differently towards different types of question tweets when they discuss the conferences on Twitter. Our data shows that tweeters tend to favor more the non-information-seeking question tweets by retweeting or favoring them, although they seem to have no preference in replying to a question tweet whether it has information needs. This may be because the non-information-seeking question tweets contains rhetorical questions, humor, etc.(Zhao & Mei, 2013) , and it could be the nature of those types of tweets that makes them easier to be favored or disseminated by other tweeters. Further, it is interesting to investigate how different responses would affect the users' future participation.

One thing to note though is that from figure 1, the tweets activities in our dataset were dominated by a few elite conferences (e.g., *WWW*, *CHI*, etc.). Therefore, the result from our study might be over-represented by these conferences while not being an objective view for the rest of CS

conferences. In future work, we plan to investigate how responses differ across different conferences and over time.

Our current classifiers aim to solve the binary classification task where a tweet is either labeled as an information-seeking qweet or not, while not able to handle multiple-class classification tasks – which is useful when we want to know the purposes of the questions (e.g., request a factual knowledge, ask for recommendation, request an opinion, ect. (Efron & Winget, 2010)). In future work, we plan to develop more

sophisticated classification technique to automatically detect different types of questions on Twitter during a large event and provide them with effective and efficient question-answer pairs, therefore to improve their experience of both using Twitter and participating the event.

References

Efron, M., & Winget, M. (2010). Questions are content: a taxonomy of questions in a microblogging environment. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–10.

Li, B., Si, X., Lyu, M. R., King, I., & Chang, E. Y. (2011). Question identification on twitter. In *Proceedings of the 20th acm international conference on information and knowledge management* (pp. 2477–2480).

Liu, Z., & Jansen, B. J. (2012). Almighty twitter, what are people asking for? *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10.

Zhao, Z., & Mei, Q. (2013). Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the 22nd international conference on world wide web* (pp. 1545–1556).

Table of Figures

Table of Tables