Almighty Twitter, What Are People Asking for?

Zhe Liu and Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania 16802
zul112@ist.psu.edu, jjansen@acm.org

ABSTRACT

With the advance of modern technology, social networking sites, such as Twitter, are becoming increasingly important information sources for people to find answers to their questions. Given such trend, in this project, we report results from our analysis of 10,000 English-written question tweets with expectations of helpful answers (which we call "information seeking tweets" in the following paper) collected in one week period. We explore the topical characteristics and patterns demonstrated in people's information seeking behaviors under online social contexts. In particular, through our topical comparisons between social search, traditional search and real time search, we find that social information seekers show more personalized requirements and more timely needs. Technology, healthcare and education related questions appeared extremely frequent among questions asked on Twitter, along with a desire to pursue help from experts in these areas. In addition to our findings on the topical domains, we also observe that social search contains a significantly less proportion of direct communication than general tweets, showing user's relatively higher openness to diversified answers. Our results also indicate the important role that time and location play in social information seeking context. Based on these findings, implications for future design of social search systems or tools are discussed at the end.

Keywords

social search, social networks, Twitter, information seeking

INTRODUCTION

Information seeking is the process or activity of attempting to obtain information in both human and technological contexts (Shih et al., 2011). Long before the invention of computers, the most prevailing information seeking option was the use of other people as valuable information sources. Humans tend to enjoy such a quick way of information seeking due to its ready accessibility and least requirement of effort, even though the people they asked or the technology they employed may not be the best sources available (Johnson, 2004). In the past few years with the fast progress of technologies, more advanced information seeking facilities, such as search engines, online catalogs, question-and-answering sites, etc. allow people to conduct their search in an even more efficient way, which quickly made the Internet another valuable information source for information acquisition. Although technology-based methods may suffer from the problem of being ambiguous and lacking in context, considering the huge benefit that technology could offer and the global trend of computerization in everyday life, people's information seeking options have under gone significant shifts from previous human-centered to technology-centered. Given that both modes have their pros and cons, social search, a combination of the two, might be an effective and a better strategy for information seekers.

Defined as the process of finding information online with the assistance of social resources (Morris et al., 2010), social search lies between the boundaries of technical and human-powered information seeking models. By making use of all possible social interactions online (Evans and Chi, 2008), social search surpasses the traditional information seeking techniques (e.g. search engine and online databases etc.) with more personalized search experience. Among various ways of conducting social searches (reference help online, post questions on Q & A sites or forums etc.), broadcasting questions to one's social network attracts most of our attention due to its popularity, simplicity and convenience. Built on the notion of embodying related people together within one place (Huberman et al., 2008), social networking sites, such as Facebook, Twitter, and Google+, has become increasingly used. Every day, hundreds of millions of new content are posted by web users and shared with their immediate networks as well as the larger Web community (Twitter$_a$, 2011), and this makes these social networks not only a remarkably good place for information broadcasting but also a platform for information seeking. Social search examples include: Does anybody know who won the NBA playoff game last night?, Any recommendation for the cheapest place to get Polaroid 600 film?, etc.

In order to understand this new form of information seeking approach, in this study, we explore behaviors people demonstrated while asking questions on social networks. We choose Twitter as the platform of our study, since

ASIST 2012, October 28-31, 2012, Baltimore, MD, USA.

Twitter is the leading social network, attracting over 100 million of active users every year (Twitter$_b$, 2011). Given such an extraordinary audience as potential information sources, the process of information seeking on Twitter logically becomes meaningful to the users.

According to Jansen et al. (2011), Google and other search engine companies already realized the potential of real time postings and are experimenting with methods to archive this social media content. Online marketers are also working to leverage people's concerns on social networking sites for customer relationship maintaining and potential business opportunities (Jansen et al., 2009). As such, understanding the characteristics of information seeking questions being asked on Twitter, including their contents and patterns, not only would benefit future social search engine development but also would bring potential business opportunities. With that goal in mind, in this research, we collect 10,000 information seeking tweets during a one-week period and conduct our analysis on both word and topical levels. We also propose three hypotheses regarding the interactivity, time, and location dependencies of information tweets compared with randomly crawled general tweets. We conduct crosstab comparisons and Pearson's Chi-square tests to examine these hypotheses.

In the remainder of this paper, we first review literatures with the focus on social search. We then present our three research questions, with hypotheses, of this study followed by our data collection and analysis methods. Next, we report the analysis results and conclude with discussion on our present design and scopes for future work.

RELATED WORK

A number of studies in social search have been conducted to explore the motivations behind this online questioning behavior. Watanabe, Nishimura, Okada, (2008) examine question answering based on mailing lists. Jansen et al (2009) report that 11% of tweets are information seeking. Morris et al. (2010) surveyed 624 social network users upon their reasons of choosing social networks as the platform for Q & A. Their results indicated that people search socially primarily due to their trust in friends over strangers. Other than that, specific audience, weak beliefs on search engine performances, and non-urgent information needs also accounted for the reasons people turn to social networks to seek information. To further examine those factors that influence users' adoption of social search, in their later work, Morris et al. (2010) conducted another in-depth user study and confirmed that seeking information on social networks can provide more personalized answers with higher answer quality.

In addition to literatures focusing on motivations, other researchers have conducted studies on the detection of question tweets. Although almost 15% of everyday tweets contain information needs (Efron and Winget, 2010), due to the high occurrence of irregularities and noise, question detection on Twitter is still a difficult task. Efron and Winget (2010) employed a set of linguistic rules to extract question tweets. Using 100 question tweets, the authors developed the taxonomy of questions asked in microblogs based on their characteristics. More than that, they also suggested an alternative taxonomy consideration from the perspectives of both the question's audience and the questioner's information needs. Through their in-depth analysis, Efron and Winget found that people asking questions in microblog in a way more like their naturalistic interactions, as opposed to their behaviors under traditional information retrieval environments. Li et al. (2011) designed a cascade method by dividing the question detection process into two phases, with one phase identifying interrogative tweets and the other extracting serious questions out of those interrogative ones. Dent and Paul (2011) built a pipeline of tools to work specifically with Twitter input to identify questions in tweets. Though their evaluations, the authors found that their method worked well on finding syntactically formed questions.

To better understand the taxonomy of question tweets, Evans and Chi (2008) conducted their study under the lens of Broder's (2002) proposed taxonomy of traditional search (transactional, navigational, and informational). The authors presented a social search model of user activities before, during, and after search and proved the value of social interactions in information seeking tasks. Utilizing naturally collected tweets from Twitter, Paul et al. (2011) assessed whether Twitter is a good place for asking questions or not. By analyzing question tweets, the authors found that rhetorical questions were the most popular, followed by questions seeking for factual knowledge.

Although attention was given to social search in the past few years, a large proportion of the above mentioned studies were conducted with survey-type data sets, and focused on the intentions behind. Among the few studies characterizing the features of social search through real tweets (Efron and Winget, 2010, Paul et al., 2011), most of the work has been on general questions, including rhetorical ones. Few focused on non-rhetorical ones, since the separation of the two types of questions is algorithmically challenging. Although findings from those studies are informative, given their broad coverage, there still lacks a comprehensive understanding regarding those serious information seeking tweets. In order to address this gap, in this study we limit our scope to only information seeking tweet, which we define as questions seriously looking for information or help. (i.e., we exclude rhetorical questions). We want to investigate information seeking behaviors and patterns people demonstrated in a social environment. As an exploratory study, we hope findings from the present work can offer valuable insights for the design and development of future social search systems or tools and research.

RESEARCH QUESTIONS

With this background and motivations, we propose two overarching research questions in this study:

RQ1. *What are prelevent information seeking topics on Twitter?*

RQ2. *How do Twitter information seeking topics compare to queries on traditional, and real time search engine?*

RQ3. *What are the patterns of people's interaction with Twitter while conducting social search?*

Twitter reflects user's changing interests and focus in real time. In order to better understand those varying interests of information seekers on this social network, we propose our first research question to explore the social searching characteristics of Twitter users on topical level. A number of studies have already been carried out on topical categorization of tweets (O'Connor et al., 2010, Hong and Davison, 2010, Zhao et al., 2011). However, given the different intentions involved specifically with information seeking, we assume that question-asking on Twitter would cover quite different topics as compared to the general tweets discovered in the previous works. In addition, we conjecture that distinctions should also exist between topics looked for on social networks versus other online technologies, given that social search depends more on social resources and requires better collaborations (Efron and Winget, 2010). Based on this, we developed our second hypothesis. We believe that successful exploration of the topical characteristics of information seeking tweets can result in the design and development of more efficient social information seeking systems or tools.

Although knowing people's information seeking topics is of vital value in designing effective search experiences, understanding their search patterns is also of significant importance. For this purpose, we propose our third research question investigating patterns demonstrated in people's information seeking behaviors under social context. The collaborative nature of social questioning and answering intuitively lead us to expect a more interactive and collaborative process as compared with people's participation in an information sharing context. For the same reason, we also assume that questions posted on Twitter tend to contain more real-time-oriented content, revealing more temporal and spatial sensitive characteristics (Jansen et al., 2011). Based on the above rationale and assumptions, we present three hypotheses regarding the interactive, temporal and localized aspect of information seeking on Twitter respectively:

Hypothesis 01: *While on Twitter, people are more likely to use @username in the context of question-asking than in the context of general tweeting.*

Hypothesis 02: *While on Twitter, people are more likely to use temporal expressions in the context of question-asking than in the context of general tweeting.*

Hypothesis 03: *While on Twitter, people are more likely to use location indicators in the context of question-asking than in the context of general tweeting.*

Here the @username is one of the several unique communication strategies developed by the Twitter community. With such mechanism, users are able to direct their updates to certain followers with the usage of the "@" symbol right before the targeted persons' usernames. For instance, the question "@XXX: Anybody knows how to get Spotify to stream from your iPhone or computer to Apple TV?" presents a direct inquiry to user XXX, assuming that he/she could help provide a solution. According to Huberman et al. (2008), this @username feature is widely adopted by Twitter users, with about 25.4% of all daily tweets are directed ones.

There is also the temporal aspect to social search. Twitter, as one of the fastest event broadcasting platform, everyday attracts millions of users to update their status to the world (Twitter$_a$, 2011), and naturally becomes a perfect real time information source. People often require "fresh" content when they are tracking events on Twitter, and this urgent information need makes time an important resource or constrain for question answering on Twitter. In order to better model social information seekers' needs, our second hypothesis tries to understand the usage of temporal expressions in questions asked on Twitter, such as "Does anybody know if that photograph is due tomorrow?"

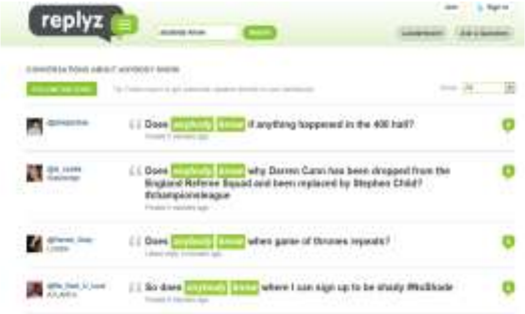
Similar as those temporal indicators, location also plays an important role in information seeking in social context, which may bring significant commercial potential for local businesses. Although recent Twitter studies are focusing on the geographic aspects of tweets, few of them concentrate on the implications of social search. Through previous geographic analysis of search queries (Gan et al., 2008), users not only demonstrated tendencies to conduct searches for local services, but they also indicated their information interests in locations away from their likely home or current location. Assuming such geographical interrogation also exists under social context, we aim to study questions containing location identifiers, such as "Going to California Adventure @Disneyland for the first time tomorrow! What are the can't miss stops for rides and food? #disney", in our third hypothesis.

METHODS

Data Collection

The task of question detection on Twitter seems very easy at the first glance, but after several attempts we found that it is actually a non-trivial process. In our first attempt relying on simple heuristics containing question mark and 5W1H words (who, what, where, why, how), we got relatively poor results with large proportion of non-information seeking contents, which we collectively called rhetorical questions in this study. A rhetorical question is usually defined as any question asked for a purpose other than to obtain the information the question asks. It is the most common interrogation type on Twitter, which accounted about half of the total questions asked on Twitter (Paul et al., 2011). Tweets such as "What are we if we don't help others? We're nothing...nothing at all." and "Why user experience should be your business's top priority? http://bit.ly/HZJszx" are both examples of rhetorical questions. Given the low precision associated with the above heuristics-based approach, we next adopted the rules as introduced by Efron and Winget (2010). Through our experiments, we noticed that although this pattern-based extraction method increased the precision rate from the previous method, this method also suffered from the tradeoff of less recall rate. Moreover, given that information seeking tweets is just a subset of all question tweets, Efron and Winget's method can very well identify interrogative tweets, but its performance on detecting only information seeking questions is still not satisfactory. Taking all these difficulties into consideration, we chose to change our data collection strategy in response to the quality of the received data.

As one popular twitter-based question and answer (Q&A) site, Replyz (replyz.com) extracts and presents its users all questions being asked on Twitter. Users are encouraged to share their knowledge by joining into a conversation according to their expertise. By tapping into the real time stream, Replyz acts as a social search engine, making Q&A on Twitter as simple as it appears in daily life. Unfortunately, unlike Twitter, as a startup company, Replyz currently does not offer any API for developers. So in order to collect questions for this project, the only solution for us is to write a crawler which sent HTTP requests to Relyz every five minutes to return the ten most recent questions. Inspired by Ritter et al.'s work (2010) on Twitter conversation modeling, we choose terms including "anybody/anyone" and "know" as the keywords in this data fetching process.



Does anybody know if anything happened in the 400 hall?

Does anybody know why Darren Cann has been dropped from the England Referee Squad and been replaced by Stephen Child? #championsleague

Does anybody know when game of thrones repeats?

So does anybody know where I can sign up to be study [illegible]

Figure 1: Interface of Replyz

Given the exploratory nature of this study, in this work we collected data for a one-week period, from Oct 13, 2011 to Oct 20, 2011. Given the scope of this study, we have kept only tweets in English, filtering out non-English tweets and duplicated questions from the collection. After that, a total number of 10,000 unique information seeking tweets were randomly selected for our next data analysis process. For comparison purpose as indicated in the hypotheses, we also collected the same amount (10,000) of general tweets during the same period through the usage of Twitter stream API, which returns a random sample of all public statuses.

Data Pre-processing

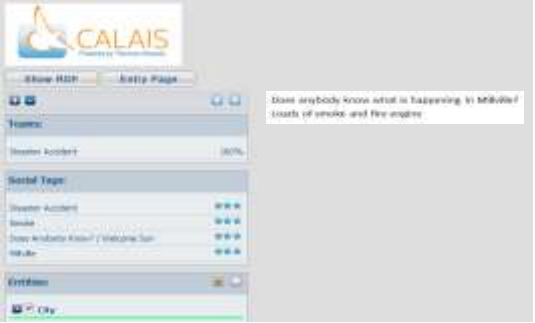
Considering the high percentage of misspellings, slangs, and acronyms contained in tweets, some substantial preprocessing work was first performed with the support of Python NLTK (Natural Language Toolkit) to remove this noise. NLTK basically contains a set of libraries which provide straightforward wrappers that can be used for common NLP tasks (Madnani, 2007).

All 20,000 tweets were first converted into lowercases and being tokenized on spaces. After removing all the stop words, punctuation marks (except @) and smileys, all tokens were then being stemmed into their root forms using the Porter stemmer (Porter, 1980). Following that, we then conducted out-of-vocabulary (OOV) word detection by using NLTK's UnigramTagger. Given that the tagger annotates any OOV word with a "None" stage, we then sent those "None" words to a spelling correction script developed based on the Norvig's algorithm (Norvig, 2007). A pre-defined list of Internet abbreviations was then used to correct those often-used slangs/acronyms into formal-written languages.

Data Analysis

For our first research question, we adopted services provided by OpenCalais (opencalais.com) to conduct our data analysis. As a web service provided by Thomson Reuters, OpenCalais enables the recognition of entities and topics in textual context by leveraging natural language processing and machine learning algorithms. OpenCalais provides both a browser-based viewer and an API for the user to access its services. As demonstrated in Figure 2, with any textual input, OpenCalais will automatically classify topics (including "Human Interest", "Business Finance", "Education", etc.) with their corresponding thresholds. With this topical analysis, we can address research questions one and two. In addition to its established function in topic detection, OpenCalais can also automatically identify all named entities as mentioned in the input. By annotating entities with their corresponding tags including time and location indicator (such as City, Continent, Country, Province/State and Holiday etc.), OpenCalais enables us to address our third research question concerning the interaction patterns as demonstrated in social search. A Perl script was developed to parse those returned document and extract the top-ranked topic (topic with largest threshold in the whole topic distribution) and all holiday and location related entities.

Considering OpenCalais's limitation on temporal annotation, to test our second hypothesis we also developed a Perl script to directly extract all temporal expressions other than holidays based on a predefined list of temporal phrases (such as month, seasons, years, “tomorrow”, “today”, “yesterday, etc.). In a same way, we also extracted all the @username appearance from both the information seeking tweets and the general ones. After the entity extraction process, we next calculated the frequencies for the appearance of all three patterns (@username, temporal indicators and location indictors) involved in both types of tweets. Pearson's Chi-square tests, with the level of significance set to 0.05, were then used to o test whether the above patterns were differently distributed between information seeking tweets and the general ones. All statistical tests in this study were performed using SPSS.



Does anybody know what is happening in Millville? Loads of smoke and fire engine

Figure 2: OpenCalais Annotation Services

RESULTS

RQ1. *What are prelevent information seeking topics on Twitter?*

RQ2. *How do Twitter information seeking topics compare to queries on traditional, and real time search engines?*

Word Level Analysis

Through our analysis on the word level, we noticed that language varied significantly on Twitter under the context of information seeking. We found that most of the words in our question collection occurred only very few times. Of the 12,163 unique terms, 9,080 (74.65%) of them were used only once, 1,476 (12.14%) twice, and 537(4.42%) three times. Around 97% of the total terms occurred less than 10 times in the whole data set, which formed the long "tail" as shown in Figure 3.

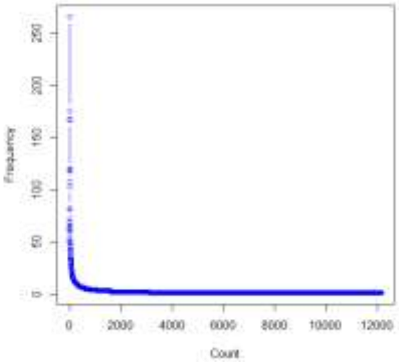


Figure 3. Term Distribution within Information Seeking Tweets

As shown by previous studies on query log analysis, Zipf’s law holds remarkably well for such kind of highly skewed distribution (Jansen et al., 2011). In this study, we also plotted a log-log graph to validate the accordance of our tweet collection with the Zipfian distribution. Although slightly deviate from the central line, the overall log-log slope fit leads us to claim that the term frequencies in information seeking tweets follows the Zipfian distribution, which means that a small proportion of the terms counted for a considerable fraction of people’s information needs on social networks.

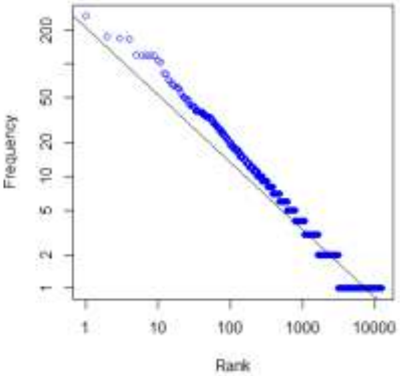


Figure 4. Rank-frequency (log-log) Distribution of Terms Used within Information Seeking Tweets

Table 1 displays the top 10 noun words used in all collected information seeking tweets. We found that technology related terms, such as “twitter”, “iphone”, “app”, etc. dominated that usage list and forms a major topic for questions asked on Twitter. Although also popular in traditional search, the occurrence of technology terms in search engine query logs is still far less frequent than our findings on information seeking tweets. Another interesting finding that we detected in our term usage list is that the frequency of celebrity names in information seeking tweets is not as high as that found in traditional search logs.

Noun Words	Frequency
Twitter*	2.26% (226)
iphone*	1.57% (157)
Christmas	1.19% (119)
app*	1.19% (119)
facebook*	0.76% (76)
website*	0.71% (71)
Channel	0.54% (54)
TV	0.53% (53)
Album	0.51% (51)
google*	0.50% (50)

Table 1. Top 10 Noun Words Used in Information Seeking Tweets

Topical Level Analysis

As for the results of our topic analysis, a comparison between the topic frequencies of information seeking question tweets and general tweets is listed as below (Table 2). We can see from this table that the majority of questions asked on Twitter fall into the category of personal interest and entertainment, accounting respectively for 17.31% and 11.80% of the whole collection. This is very consistent with the top two topics we found in the general tweets. Examples of information needs in these two categories include: “Does anybody have red heels I can wear tomorrow?” and “Where is Selena and Justin right now does anybody know?” From this, we can conclude that Twitter is more of a personal sharing platform motivated by individual interests and needs regarding its social aspect.

From Table 2, “Technology Internet” and “Health Medical” related topics are doubled in frequency for information seeking tweets as compared to those general tweets. This indicates that people treat Twitter as a platform for expert-seeking purpose in the domain of technology and healthcare. Examples of those expert-seeking related tweets include: “Anyone got any ideas as to why my iPhone backing up takes several hours and never completes?” and “Anybody here a doctor? How do I get rid of this cold? Its making me nearly puke.”

Besides “Technology Internet” related topic, in our comparison results we also observed a notable increase in questions about “Education” topics. Questions such as “Does anybody know if we have homework in Ms Caylor's honors biology class...I wasn’t paying attention?” showed the possibility of using Twitter for quick questions about school assignments, curriculum, and problems encountered. Compared with real-world tutoring, questions asked on Twitter enables students to have 24/7 access to their peer’s opinions and solutions. Students can collaborate with each other to accomplish their problem-solving tasks across school, or even across country.

Question Topics	Frequency	General Topics	Frequency
Human Interest	17.31% (1,731)	Human Interest	16.42% (1,642)
Entertainment Culture	11.80% (11,80)	Entertainment Culture	13.28% (1,328)
Technology Internet	7.06% (706)	Hospitality Recreation	8.28% (828)
Hospitality Recreation	7.00% (700)	Sports	6.83% (683)
Sports	6.78% (678)	Health Medical	4.51% (451)
Health Medical	4.71% (471)	Technology Internet	3.02% (302)
Education	3.66% (366)	Religion Belief	2.50% (250)
Social Issues	2.61% (261)	Social Issues	2.31% (231)
Business Finance	2.54% (254)	Education	2.24% (224)
Law Crime	2.20% (220)	Environment	2.16% (216)

Table 2. Top 10 Topic Frequencies of Information Seeking Tweets verses General Tweets

Traditional Search Topics	Frequency	Real Time Search Topics	Frequency	Information Seeking Tweet Topics	Frequency
Commerce, Travel	30.4%	Society	25.9%	Human Interest	17.31% (1731)
People, Places, Things	16.0%	Arts	17.6%	Entertainment Culture	11.80% (1180)
Unknown, Others	13.2%	Computers	16.4%	Technology Internet	7.06% (706)
Health, Science	8.9%	Business	13.7%	Hospitality Recreation	7.00% (700)
Entertainment, Recreation	7.0%	News	4.9%	Sports	6.78% (678)
Computers, Internet	5.7%	Sports	4.9%	Health Medical	4.71% (471)
Education, Humanities	5.6%	Recreation	4.3%	Education	3.66% (366)
Society, Culture	4.7%	Science	3.2%	Social Issues	2.61% (261)
Sex, Pornography	3.8%	Shopping	2.7%	Business Finance	2.54% (254)
Government, Legal	3.6%	Health	2.6%	Law Crime	2.20% (220)

Table3.Topic Frequencies of Traditional Search versus Real Time Search versus Information Seeking Tweets

In addition to comparing the topic distribution for two different types of tweets, we also investigated the topical differences exist between the information seeking tweets and traditional search engine queries as reported in the previous work (Jansen et al. 2007). As indicated by Table 3, topics shifted considerably. Categories such as "Commerce, Economy", "Health, Science", "Society, Culture", and "Sex, Pornography" have a significant drop from traditional search to information seeking on Twitter. In contrast, notable increases are observed for categories including "Human Interest", "Entertainment, Culture", "Technology Internet", and "Sports". From these significant topic shifts, we can infer that, as for more private inquiries posted on social networks than traditional search engines, Twitter can be a better medium for more personalized questions and answers. Besides, from this table, we observe that information seekers on Twitter show little interests to public events on economy, policy, politics, and society themes. However, they maintain huge interests on entertainment related topics, such as entertaining news and celebrity gossips. Based on all these changes, we conclude that if one defines the evolvement of human information seeking behavior from the 19th century to the 20th as from e-sex to e-commerce (Spink, 2002), then from the topical distinctions between the traditional search engine and the information seeking tweets, we can define such change as from e-public to e-private.

Besides comparing it with the traditional search as detailed above, in this work we also related information seeking on Twitter with real time search, trying to characterize the similarities and the differences in terms of their topic coverage. Through our comparison, we identify the same topic shift from public to private. From Table 3, we can see that one fourth of the real time queries are of societal intentions. Queries such as "separate but equal", "divorce rates" and "abortion" are of high occurrence under the context of real time search (Jansen et al., 2010). However, social information seekers care far less about such public topics. Only 2.61% of the total questions posted on Twitter are regarding social issues. Although distinctions exist, compared with the traditional information retrieval, real time search more closely resembles social search. Like what we found in information seeking tweets, topic distribution of real time search also contains a high proportion of entertainment and technology related queries, but a low occurrence of sexually-related ones. As those similarities are noticed, we can infer the characteristics of those information seeking behaviors on Twitter have some relatively real time requirements.

RQ3. *What are the patterns of people's interaction with Twitter while conducting social search?*

Regarding the three hypotheses under research question three, we conducted Pearson's Chi-square tests in our analysis with a pre-set significance level of 0.05. To better display the joint distribution of the tests variables, crosstab views of frequencies are also presented.

Hypothesis 01: *While on Twitter, people are more likely to use @username in the context of question-asking than in the context of general tweeting.*

Originally, we assume that for quicker response and more personalized answers, people may tend to adopt more @username to direct their questions to a specific audience or group for more addressivity. However, to our surprise, as can be seen in Table 4, we found that people use far less @username under the context of information seeking than sharing. The majority (91.48%) of our information seeking tweets contain no @username. Only 8.52% of them are questions to certain addressees. To understand if the usage of @username is related to tweet types, we then conduct the Pearson's Chi-square test. We found significant association between the frequency of @usernames appeared and the tweet type (information tweet vs general tweet) (χ^2=8434.944, df =1, p = 0.000 < 0.05). People are actually more likely to use @username under the context of information sharing rather than information seeking. So, although with a significant p-value, we reject our hypothesis 01, people are LESS likely to use @username in the content of question asking.

Appearance of @username	Tweet Type		Total
	Information Seeking	General	
Yes	852 (8.52%)	7,225 (72.25%)	8,077 (40.39%)
No	9,148 (91.48%)	2,775 (27.75%)	11,923 (59.62)
Total	10,000 (100.00%)	10,000 (100.00%)	20,000 (100.00%)

Table4. Cross-Tabular Distribution of the Frequency of @username and Tweet Type

Through our further investigation on the unexpected results, we noticed that @username is used mainly in the following three conditions in information seeking tweets: (1) seeking for information about certain people, for instance "Does anybody know if @XXX got a new phone? Cuz the number I have apparently doesnt work". (2) looking for credible answers regarding certain people or events by addressing direct attention from related experts, celebrities, and organizations. Example tweet of this category are "Does anybody know if @XXX will have her album out for Christmas?" "can someone….anyone actually, riddle me how Lance Berkman did not get a silver slugger? @XXX_ESPN @XXX @XXX_ESPN" and "Does anyone know if you can use multiple @groupon discount codes at the same time?" (3) searching for solutions for personal issues. Examples are "Does anybody know if we are getting paid today? @XXX @XXX @XXX" " and "Anybody know what happened to Riley? @XXX @XXX".

We also detected that opinion and recommendation-seeking tweets tend to use less or no @username, for instance "Can anyone recommend any good PS3 games? Getting bored with Fifa…" and "Does anybody have both a Kindle Fire AND a nook Colore Tablet? Which one is better in your opinion?" We try to explain this pattern of behavior from the standpoint of Granovetter's (1973) weak tie theory, attributing it to people's intention to maximize their information by opening their questions to larger audiences. According to Granovetter's theory, strong ties may provide more credible and readily available information sources for people. However, due to their constraints on the breadth and the non-redundancy of information, strong ties also limit people's opportunities to access more diverse solutions. Noticing such disadvantages of strong ties, people on Twitter tend to use less @username for more divergent and non-redundant information or resources.

Hypothesis 02: *While on Twitter, people are more likely to use temporal expressions in the context of question-asking than in the context of general tweeting.*

As for hypothesis 02, a crosstab is constructed, as shown in Table 5. Compared with the previous tests on @usernames, this time both types of tweets showed less differences on the occurrence of temporal adverbs. Nearly 10% (8.04%) of our information seeking tweets contain temporal expressions. Similarly, about 10.74% of general tweets also demonstrate temporal characteristics.

Appearance of Temporal Expressions	**Tweet Type**		**Total**
	Information Seeking	**General**	
Yes	804 (8.04%)	1,074 (10.74%)	1,878 (9.39%)
No	9,196 (91.96%)	8,953 (89.53%)	18,149 (90.75%)
Total	10,000 (100.00%)	10,000 (100.00%)	20,000 (100.00%)

Table5.Cross-Tabular Distribution of the Frequency of Temporal Adverbs and Tweet Type

From the Pearson's Chi-Square test, again we detected significant association between the frequency of temporal expressions appeared and the tweet type (information tweet or general tweet) (χ^2 = 42.035, df =1, p = .000 < 0.05). However, given that people are actually more likely to broadcast real time status than asking time-related questions. We have to reject our hypothesis 02, as people are LESS likely to use temporal expressions in the context of question-asking.

Although there are fewer temporal expressions observed under the information seeking context, still around 10% of those tweets are associated with temporal-sensitive inquiries. Through our analysis, we found that temporal related tweets can be roughly divided into two groups: time as the question focus and time as the question qualifier. Regarding the former condition, we can say that a majority of the questions are also spatial-sensitive. To answer questions, such as "What time does the apple store shut? Anyone…" and "Does anyone know what day or time we can sign up for classes next semester?", respondents need to know where the searching is performed in order to provide credible answers. While for questions in the latter case, even though less location-sensitive than the previous condition, they require more immediacy answers. Questions such as "Does anybody know what JoePa's movements were today?" and "Does anybody know if Manchester United won today?" represent people's real time information needs on following the progress of certain events.

Hypothesis 03: *While on Twitter, people are more likely to use location indicators in the context of question-asking than in the context of general tweeting.*

As can be seen from Table 6, only a small proportion (2.89%) of the information seeking tweets contains location identifiers specifically in the tweet and the frequency of that is even less under the context of information sharing (1.10%).

Pearson's Chi-Square test again demonstrated significant association between the frequency of location identifiers appeared and the tweet type (information tweet or general tweet) (χ^2 = 81.938, df =1, p = .000 < 0.05). Also given the almost doubled frequency of location identifier used under the information seeking context, hypothesis 03 is supported.

Appearance of Location Identifiers	**Tweet Type**		**Total**
	Informatio n Seeking	**General**	
Yes	289 (2.89%)	110 (1.10%)	399 (2.00%)
No	9,711 (97.11%)	9,890 (98.90%)	19,601 (98.01%)
Total	10,000 (100.00%)	10,000 (100.00%)	20,000 (100.00%)

Table6.Cross-Tabular Distribution of the Frequency of Location Identifier and Tweet Type

In further analysis, we found that information seeking tweets containing location identifiers are generally concentrated in two categories: (1) travel advice seeking and (2) location business search. The first category involves inquiries traditionally pursued by travelers. Those kinds of questions usually cover topics such as "attractions", "activities'' "flight", and "hotel" et al., Typical examples of those tweets are "Does anybody know a seriously cheap way to go to Liverpool?" and "Does anybody have any Melbourne hotel recommendations? Close the St Kilda". Different from the former type, the second category mainly covers tweets searching for local business information, including "restaurant", "club" and "gym". Recommendation-seeking tweets such as "Does anybody have any restaurant recommendations in Toronto? I feel like I am always eating at the same places.. #needchange" and "Does anyone have a recommendation for a dentist in central Iowa who does mercury-free fillings?" tend to dominate this whole category.

DISCUSSION

We believe results from this study provide valuable insights into people's information seeking behaviors on social platforms. First of all, we saw clear topical shift from public concerns to individual inquires while comparing our information seeking tweets with queries from traditional search engines. Accompanied with this change, information seeking on social networks requires more understanding of one's personal profile or context, so that it can provide more credible and reliable solutions. Regarding this aspect, we believe our study could provide a basis for future studies on the design of more personalized social search tools.

In addition to the personalization requirements, we also observed people's strong needs in real-time solutions. Consistent with the topical distribution of real time search, a large proportion of questions asked on Twitter pertain to the latest events. This raises an increasing essential to readjust Twitter's current displaying mechanism. Instead of mixing all the updates together, there could be some visually salient features that enable the highlighting display of those time-sensitive questions. Or there could be certain techniques, through which those real-time questions can be delivered to their potential answerers to address their special attentions.

Another interesting finding of this study was the topical differences found between information seeking tweets and those general ones. Consistent with previous work (e.g., Paul et al., 2011), we noticed people's significant needs on professional solutions and recommendations on both technology and healthcare related topics. Based on this observation, we infer that any way of matching someone's professional knowledge to another person's questions, such as the techniques of personal profiling and expert assessment may bring significant benefits. Besides above mentioned two cases, we also detected an increasing trend of using twitter for education purpose. We believe this finding should be considered in future work on developing Twitter into a useful education platform for users to collaborate and share knowledge.

In regards to behavior changes in social search, we found several interesting patterns in this study. First, social information seekers want not only personalized answers, but also diverse and nonredundant ones. This requires them to develop a diffusion network with a large number of structural holes that can maximize their information benefits. Future research in recommendation of relevant people to follow on Twitter should take this point into consideration. Second, although not explicitly appeared as frequently as expected, temporal and spatial qualifiers still play a very important role in finding the relevant answers. Many times, people can only answer one's question on Twitter by first leveraging its temporal and spatial context. Given their importance and the lack of focus within traditional search techniques, findings summarized in our work suggest the necessity for developing a multidimensional search method by taking both the relevant temporal and spatial context into consideration. In addition to its technical implications, we also think that our analysis on location-sensitive search on Twitter provides valuable marketing information to those social network advertisers and local businesses.

As with any research, there are limitations to this study. Given its exploratory nature, one limitation is that the small scale dataset collected. Given that Twitter has more than 100 million (Twittera, 2011) tweets posted daily, 10,000 tweets may not represent the universal information seeking patterns on Twitter. Further research with larger dataset is needed to ensure that the results are generalizable. However, we believe that the research reported here is an important step in this direction and will assist in directing avenues for future research. Second, due to the complexity of question extraction on Twitter, this work constraint the information seeking tweets with keywords. This probably discarded a number of other questions given the free style of writing for Twitter users. Third, with only explicit temporal and spatial expressions considered in this study, we may not completely present the value of both features, even though our current results are already very persuasive. It would be useful for future studies to include implicit temporal and spatial information.

CONCLUSION

In this paper we presented an analysis on information seeking behaviors of Twitter users', including the topics of the questions they ask, as well as the interaction patterns they demonstrate. Our results showed that, compared with traditional search, information seeking on social networks demonstrates an obvious topical shift towards a more personal aspect. Consistent with previous studies and our discoveries on the general tweets, we found that the most popular categories among questions asked on Twitter were "Human Interest" and "Entertainment Culture", containing nearly 30% of the sample. We also noticed that an increasing number of questions within the categories of "Technology Internet", "Health Medical" and "Education" were being asked on Twitter as compared to their frequencies in those general tweets. We compared information seeking tweets with queries from one real time search engine. Although there were differences, the large overlap highlights the real time characteristics of information seeking on Twitter.

We conducted hypothesis tests regarding the interactivity, the temporal and spatial sensitivity aspect of question-asking on Twitter. While compared with the general tweets, our results showed a less frequent usage of direct communication in information seeking scenarios, indicating a more diversified information interests from the inquirers.

Besides, we also detected a significant usage of temporal and location identifiers in questions asked on Twitter. We believe that our study offers valuable insights into the future development of social search systems or tools which can make good use of those temporal and spatial context cues.

For future work, a more comprehensive analysis with a more complete dataset could be beneficial. Besides, with all these findings at hand, we next plan to develop practical tools that can translate all the implications as we discussed in this study into real world search tools for social network users.

REFERENCES

Broder, A. (2002). "A taxonomy of web search" *SIGIR Forum*, 36(2):3–10.

Dent, K., & Paul, S. (2011). "Through the twitter glass: Detecting questions in micro-text". *Proceedings of AAAI-11 Workshop on Analyzing Microtext*.

Efron, M., & Winget, M. (2010). "Questions are content: A Taxonomy of Questions in a Microblogging Environment". *Proceedings of the 2010 Annual Meeting of the American Society for Information Science and Technology*.

Evans, B. M., & Chi, E. H. (2008). "Towards a model of understanding social search". *Proceeding CSCW '08. ACM Press*, New York, 485-494.

Gan, Q., Attenberg, J., Markowetz, A., & Suel, T. (2008). "Analysis of geographic queries in a search engine log". *Proceeding the WWW'08 Workshop on Location and the Web*, pages 49–56. ACM Press.

Granovetter, M. S. (1973). "The strength of weak ties". *American journal of Sociology*, 78: 1360-1380.

Hong, L. & Davison, B. D. (2010) "Empirical Study of Topic Modeling in Twitter". *Proceedings of the First Workshop on Social Media Analytics (SOMA) at KDD 2010*, pages 80-88, Washington, DC, July.

Huberman,B. A., Romero, D. M., & Wu, F. (2008). "Social networks that matter: Twitter under the microscope". *First Monday*, 14(1).

Jansen, B. J., Spink, A., Blakely, C. & Koshman, S. (2007). "Web Searcher Interaction with the Dogpile.com Meta-Search Engine". *Journal of the American Society for Information Science and Technology*, vol. 58, no. 5, pp. 744–744.

Jansen, B. J., Liu, Z., Weaver, C., Campbell, G., & Gregg, M. (2011). "Real time search on the web: Queries, topics, and economic value", *Information Processing & Management*, vol. 47, pp. 491–506.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009) "Twitter power: Tweets as electronic word of mouth". *Journal of the American Society for Information Science and Technology*, 2009.

Johnson, C. A. (2004). "Choosing people: the role of social capital in information seeking behavior". *Information Research* 10(1).

Li, B., Si, X., Lyu, M. R., King, I., & Chang, E. Y. (2011) "Question Identification on Twitter". *Proceedings of the 20th ACM international conference on Information and knowledge management*.

Madnani, N. (2007). "Getting started on natural language processing with Python". *ACM Crossroads*, 13(4).

Morris, M. R., Teevan, J., & Panovich, K. (2010). "What do people ask their social networks, and why? A Survey study of status message Q&A behavior". *Proceedings of CHI, 2010*.

Morris, M. R., Teevan, J., & Panovich, K. (2010). "A Comparison of Information Seeking Using Search Engines and Social Networks". *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Norvig. P. (2007). "How to Write a Spelling Corrector". http://norvig.com/spell-correct.html. Retrieved on Oct 2011.

O'Connor, Krieger, M., & Ahn. D. (2010). "TweetMotif: Exploratory search and topic summarization for twitter". *Proceedings of ICWSM. 2010*.

Paul, S. A., Hong, L., & Chi, E. H. (2011). "Is Twitter a good place for asking questions? A characterization study". *Proceedings of ICWSM 2011*, AAAI Press, 1-4.

Porter. M.F. (1980). "An algorithm for suffix stripping". *Program*, vol 14, no 3, pp 130-130.

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised Modeling of Twitter Conversations. *NAACL 2010*.

Shih, C., Chen, M., Chu, H., & Chen, Y. (2011). Enhancement of information seeking using information needs radar model. *Information Processing & Management*, vol. 48, pp. 524–536.

Spink, A., Jansen, B.J., Wolfram, D., and Saracevic, T. (2002) "From E-Sex to E-Commerce: Web Search Changes". *IEEE Computer*, 35(3), 107-109, 2002.

Twitter$_a$. (2011). "#numbers". http://blog.twitter.com/2011/03/numbers.html. Retrieved on April, 2011.

Twitter$_b$. (2011). "One hundred million voice". http://blog.twitter.com/2011/09/one-hundred-million-voices.html. Retrieved on April, 2011.

Watanabe, Y., Nishimura, R., and Okada, Y. (2008). "A Question Answer System based on Confirmed Knowledge Acquired from a Mailing List." *Journal of Internet Research*, 18 (2), pp. 165-176.

Zhao, X., Jiang, J., Weng, J., He, J., Peng, L., Yan, H., and Li. X. (2011). "Comparing Twitter and traditional media using topic models". *Proceedings of the 33rd European Conference on Information Retrieval*.