Questions about Questions: An Empirical Analysis of Information Needs on Twitter

Zhe Zhao
Department of EECS
University of Michigan
zhezhao@umich.edu

Qiaozhu Mei
School of Information
University of Michigan
qmei@umich.edu

ABSTRACT

Conventional studies of online information seeking behavior usually focus on the use of search engines or question answering (Q&A) websites. Recently, the fast growth of online social platforms such as Twitter and Facebook has made it possible for people to utilize them for information seeking by asking questions to their friends or followers. We anticipate a better understanding of Web users' information needs by investigating research questions about these questions. How are they distinctive from daily tweeted conversations? How are they related to search queries? Can users' information needs on one platform predict those on the other?

In this study, we take the initiative to extract and analyze information needs from billions of online conversations collected from Twitter. With an automatic text classifier, we can accurately detect real questions in tweets (i.e., tweets conveying real information needs). We then present a comprehensive analysis of the large-scale collection of information needs we extracted. We found that questions being asked on Twitter are substantially different from the topics being tweeted in general. Information needs detected on Twitter have a considerable power of predicting the trends of Google queries. Many interesting signals emerge through longitudinal analysis of the volume, spikes, and entropy of questions on Twitter, which provide insights to the understanding of the impact of real world events and user behavioral patterns in social platforms.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining

General Terms

Experimentation, Empirical Studies

Keywords

Information Need, Time Series Analysis, Twitter

1. INTRODUCTION

Recent years have witnessed an explosion of user-generated content in social media. Online social platforms such as Facebook, Twitter, Google+, and YouTube have been complementing and replacing traditional platforms in many daily tasks of Web users, including the creation, seeking, diffusion, and consumption of information. Indeed, a very recent research interest has been shown in understanding how people seek for information through online social networks [26, 8, 29, 28, 21], how this "social information seeking" behavior differs from that through traditional channels such as search engines or online question answering (Q&A) sites, and how the social channel complements these channels [27, 16, 25]. Based on a survey conducted by Morris et al. in 2010, 50.6% of the respondents[1] reported having asked questions through their status updates on social networking sites [26]. The questions they ask involve various needs of recommendations, opinions, factual knowledge, invitations and favor, social connections, and offers. They covered many topics such as technology, entertainment, shopping, and professional affairs [26]. An analysis later by Efron and Winget suggested that 13% of a random sample of tweets (microblogs posted on Twitter.com) were questions [8].

Why is it compelling to understand the questions asked on social platforms? This emerging research interest largely attributes to the importance of understanding the information needs of Web users. Indeed, as the core problem in information retrieval, a correct interpretation of the information needs of the users is the premise of any automatic system that delivers and disseminates relevant information to the users. It is the common belief that the analysis of users' information needs has played a crucial role behind the success of all major Web search engines and other modern information retrieval systems. Better understanding and prediction of users' information needs also provides great opportunities to business providers and advertisers, leading to effective recommender systems and online advertising systems.

Long have Web search engines been the dominating channel of information seeking on the Web. According to recent statistics [2], 4 billion of search queries are submitted to Google every day. The rest of the territory is shared by other channels such as online question answering (Q&A) sites such as Yahoo! Answers. A statistic in 2010 reported a daily volume of 823,966 *questions and answers* [3], in which each question on average earned five to six answers according to [30]. This is much smaller than the number of information needs asked through search engines.

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

[1]Note for the selection bias as all were Microsoft employees.
[2]http://www.comscore.com/Insights/Press_Releases/2012/4/comScore_Releases_March_2012_U.S._Search_Engine_Rankings
[3]http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served/

The emergence of social platforms seems to be a game-changer. If the ratio reported by Efron and Winget [8] still holds today, there will be over 50 million questions asked through Twitter according to a recent statistic of 400 million tweets posted per day [4]. This number, although still far behind the number of search queries, has already overwhelmed the number of questions in traditional Q&A sites. Moreover, it has been found that people tend to ask different questions to their friends rather than to search engines or to strangers on Q&A sites. In Figure 1, we can see people asking questions in their tweets by either broadcasting so that any of their followers can respond to them, or by targeting the question to particular friends. The results of the survey by Morris et al. suggested that respondents especially prefer social sites over search engines when asking for opinions and recommendations, and they tend to trust the opinions of their friends rather than strangers on Q&A sites [26]. It is reported in [27] that users enjoy the benefits of asking their social networks when they need personalized answers and relevant information that unlikely exists publicly on the Web. It is also reported that information needs through social platforms present a higher coverage of topics related to human interest, entertainment, and technology, compared to search engine queries [29].

Tweets Conveying Information Need	Tweets not Conveying Information Need
Do you know whether there is a roadwork on I94	Man so everybody a frank ocean fan now? Idc I was an original…
Which restaurant nearby has a discount?	Why do I always do this? #hesatool #fml
@someuser u work today???	@someuser how are you?
Can anyone suggest some local restaurants in Beijing?	They're still together, why haven't they broken up yet?!?!
@someuser, do you what I am doing is good?	Umm what? It's already August? Hey Summer, #wheredygo?
What's your favorite summer album to throw on a car stereo?	Im still gone smile! What are you thanking?! Em not
Is my avi cute?	Why won't people understand that?!

Figure 1: Instances of tweets conveying an information need, and those which don't.

All evidence suggests that the questions being asked through social networks present a completely new perspective of online information seeking behaviors. By analyzing this emerging type of behavior, one anticipates to help users effectively fulfill their information needs, to develop a new paradigm of search service that bridges search engines and social networks (e.g., social search [9, 25]), and to predict what the users need in order to strategize information service provision, persuasion campaign, and Internet monetization.

The availability of large scale user-generated content in social network sites has provided a decent platform for this kind of analysis. This revives our memory about the early explorations of analyzing search engine query logs (e.g., [31, 32, 23]). Indeed, the analysis of information needs with large-scale query logs has provided tremendous insights and features to researchers and practitioners, and it has led to a large number of novel and improved tasks including search result ranking [1], query recommendation [2], personalization [33], advertising [4], and various prediction tasks [14, 17]. We believe that a large-scale analysis of information needs on online social platforms will reproduce and complement the success of query log analysis, the results of which will provide valuable insights to the design of novel and better social search and other online information systems.

In this paper, we take the initiative and present the first very large scale and longitudinal study of information needs in Twitter, the leading microblogging site. Questions that convey information needs are extracted from a collection of **billions** of microblogs (i.e., tweets). This is achieved by an automatic text classifier that distinguishes real questions (i.e., tweets conveying real information needs) from tweets with question marks. With this dataset, we are able to present a comprehensive description of the information needs with both the perspectives of content analysis and trend analysis. We find that questions being asked on Twitter are substantially different from the content being tweeted in general. We prove that information needs detected on Twitter have a considerable power of predicting the trends of search engine queries. Through the in-depth analysis of various types of time series, we find many interesting patterns related to the entropy of language and bursts of information needs. These patterns provide valuable insights to the understanding of the impact of bursting events and behavioral patterns in social information seeking.

The rest of the paper is organized as follows. We start by introducing the related work. The setup and dataset of our experiments is presented in Section 3, followed by the description of an automatic classifier of information needs in Section 4. In Section 5, we describe the detailed results and insights drawn from the analysis of the large collection of information needs. We then conclude in Section 6.

2. RELATED WORK

To the best of our knowledge, this is the first work to detect and analyze information needs from billion level, longitudinal collection of tweets. Our work is generally related to the qualitative and quantitative analysis of information seeking through social platforms (e.g., [29, 5, 27]) and temporal analysis of online user behaviors (e.g., [3, 14]).

2.1 Questions on Social Platforms

As described in Section 1, there is a very recent interest in understanding how people ask questions in social networks such as Facebook and Twitter [5, 26, 8, 29, 28, 21]. This body of work, although generally based on surveys or small scale data analysis, provides insights to our large-scale analysis of information needs in Twitter. For example, In [29], the authors labeled 4,140 tweets using Mechanical Turks and analyzed 1,351 of them which were labeled as real questions. They presented a rich characterization of the types and topics in these questions, the responses to these questions, and the effects of the underlying social network. In [26, 27, 36], Morris et al. surveyed whether and how people ask questions through social networks, the differences between these questions and questions asked through search engines, and how different cultures influence the behaviors. Efron and Winget further confirmed this difference with a study of 375,509 tweets. Using a few simple rules, they identified 13% of these tweets as questions. They also provided preliminary findings on how people react to questions.

[4]http://news.cnet.com/8301-1023 3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/

More sophisticated methods have been proposed to detect questions in online forums and Q&A sites [6, 35]. A recent work [19] studied the same problem in the context of Twitter, which presented a classifier that achieved 77.5% of accuracy in detecting questions from tweets. A much more accurate classifier is needed, however, to analyze information needs at a very large scale.

It is interesting to see the effort of making use of the understandings of social information seeking. In [16], Morris et al. proposed SearchBuddies, an automatic content recommendation for information seeking behavior. The proposed work finds relevant content based on the content and social context of Facebook status asking for information. Such effort can also be found in work like [9, 27, 34], where a new paradigm of search service, social search, is discussed. These explorations provided good motivations to our effort of large-scale analysis of information needs on social platforms.

2.2 Temporal Analysis of User Activities

The techniques of analysis used in our work is related to the existing work of analyzing user behaviors in general. For example, in [3], the authors proved that sentiment trends in Twitter has a power of predicting the Daw Jones Industrial Average. In their approach, the Granger Causality Test is used to test this predictive power. In [14], the authors used the Google trend related to influenza spread worldwide to detect which stage the flu was at and to predict the trend of the flu. Our analysis provides another important application of these methods. Note that our analysis is also related to the analysis of large scale search engine logs (e.g., [31, 32, 23]). Indeed, we do anticipate the analysis of information needs in social platforms to complement the analysis of information needs through search engines, and provide a totally different perspective and insights to search engine practitioners.

3. EXPERIMENT SETUP

We analyze a longitudinal collection of microblogs (tweets) collected through the Twitter stream API with Gardenhose access, which collects roughly 10% of all public statuses on Twitter. The collection covers a period of 358 days, from July 10th, 2011 to June 31st 2012. A total number of 4,580,153,001 (12.8 million tweets per day) tweets are included in this collection, all of which are self-reported as tweets in English. Every tweet contains a short textual message constrained by 140 characters, based on which we determine whether it conveys an information need. For every tweet, we keep the complete metadata such as the user who posted the tweet, the time stamp at which it was posted, and geographical locations of the user if provided. In the analysis in this paper, we adopt only the time and user information but leave the richer metadata for future analysis.

Note that a tweet may be a retweet of an existing tweet, may mention one or more users by "@" their usernames, and may contain one or more hashtags (user-defined keywords starting with an "#"). In our analysis, we keep the original form of all hashtags, but de-identify all usernames mentioned in the tweets (e.g., substituting all of them with a token "@someuser").

To analyze information needs in these tweets, we focus on tweets that appear to be questions. Specifically, we focus on tweets that contain at least one question mark. Note that this treatment could potentially miss information needs that are presented as statements. According to statistics in [26], 81.5% of information needs asked through social platforms were explicitly phrased as questions and included a question mark. Questions phrased as statements were often preceded by inquisitive phrases like "I wonder," or "I need" [26]. Because there is little foreseeable selection bias, we choose to focus on questions with explicit question marks instead of enumerating these arbitrary patterns in an ad hoc manner. In our collection of tweets, 10.45% of tweets contain explicit appearance of question mark(s).

4. DETECTING INFORMATION NEEDS

Not all tweets with question marks are real questions. In order to detect information needs from tweets collected in Section 3, we need to distinguish tweets that convey a real information need from many false positives such as rhetorical questions, expressions of sentiments/mood, and many other instances. Figure 1 presents examples of tweets that convey real information needs and those which don't.

In this section, we present the task of detecting information needs from tweets which is casted as a text classification problem. Given a tweet that contains one or two question marks, the task is to determine whether it expects an *informational* answer or not. In this section, we first give a formal definition of this problem and rubrics based on which human annotators can accurately classify a tweet. A qualitative content analysis is conducted in order to develop a codebook of classification and generate a set of labeled tweets as training/testing examples. We then introduce a classifier trained with these examples, using the state-of-the-art machine learning algorithms and a comprehensive collection of features. The performance of the text classifier is evaluated and presented in Section 4.5.

4.1 Definition and Rubrics

Given a tweet with question marks, our task is to determine whether this tweet conveys a real information need or not (i.e., real questions). A formal definition is needed to describe what we mean by "a real information need." Inspired by the literature of how people ask questions on Twitter and Facebook [29, 27], we provide the following definition and rubrics of "real questions:"

A tweet conveys an information need, or is a real question, if it expects an informational answer from either the general audience or particular recipients.

Therefore, a tweet conveys an information need if

- **it requests for a piece of factual knowledge, or a confirmation of a piece of factual knowledge**. A piece of factual knowledge can be phrased as a claim that is objective and fact-checkable (e.g., "Barack Obama is the 44th president of the United States").
- **it requests for an opinion, idea, preference, recommendation, or personal plan of the recipient(s), as well as a confirmation of such information.** Here the information been requested is subjective, which is not fact checkable at the present.

A tweet does not convey an information need if it doesn't expect an informational answer. This includes rhetorical

questions, expressions of greeting, summary of the content (eye attractors), imperial requests (to be distinguished from invitations), sarcasm, humor, expressions of emotion (complaints, regrets, anger, etc), or conversation starters.

Figure 1 shows some examples of tweets conveying information need and tweets which don't. Using the description we proposed above, a human annotator can easily classify a tweet. In the following subsections, we introduce how we extract features from the tweets, how we select features using the state-of-the-art feature selection techniques, and how we train classifiers using a single type of feature and then combine them using boosting.

4.2 Human Annotation

Based on the rubrics, we developed a codebook[5] and recruited two human annotators to label a random sample of tweets. We sampled 5,000 tweets randomly from our collection, each of which contains at least one question mark and self-reported as English. Finally, 3,119 tweets are labeled as real tweets in English and have same labels by the two coders. Among the 3,119 tweets, 1,595 are labeled as conveying an information need and 1,524 are labeled not conveying an information need. The inter-rater reliability measured by Cohen's kappa score is 0.8350, the proportion of the agreements in all the results is 91.5%. The 3,119 labeled tweets will be used to train and evaluate the classifier of information needs.

4.3 Text Classification

4.3.1 Feature Extraction

The classification of tweets is a particularly challenging because of the extremely short length of content (i.e., a tweet has a limited length of 140 characters). This makes the textual features in an individual tweet extremely sparse. To overcome this challenge, we not only utilize lexical features from the content of the tweets, but also generalize them using the semantic knowledge base WordNet [24, 10]. It is also our intent to include syntactical features as well as metadata features. We extracted four different types of feature from each tweet, i.e., lexical ngrams, synonyms and hypernyms of words (obtained from the WordNet), ngrams of the part-of-speech (POS) tags, and light metadata and statistical features such as the length of the tweet and coverage of vocabulary(i.e., number of different words used in a tweet divided by the number of different words in the whole dataset), etc..

Lexical Features

We included unigrams, bigrams, as well as trigrams. The start and end of a tweet are also considered in the ngrams. This gives us great flexibility to capture features that reflects the intuitions from qualitative analysis. For example tweets beginning with the 5Ws (who, when, what, where, and why) are more likely to be real questions. All lexical features are lowercased and stemmed using the Krovetz Stemmer [18]. Hashtags are treated as unique keywords. To eliminate the noise of low frequent words, a feature is dropped if it appears less than 5 times. This resulted in 44,121 lexical features.

WordNet Features

To deal with the problem of data sparsity, we attempt to generalize the lexical features using the synonyms and the hypernyms of the words in tweets. We hope this approach would connect different features sharing relevant semantics in different tweets. By doing this, our algorithm can also handle words that haven't been seen in the training data, thus is anticipated to achieve a higher performance with limited training data.

In [22], the authors studied how different types of relevant words from WordNet influence the results of text classification. In most cases, using only synonyms and hypernyms can improve classifiers such as Support Vector Machine (SVM) the most. We explored different WordNet features in our task and drew the same conclusion. We therefore adopt only synonyms and hypernyms of words in a tweet as additional features. Note here we actually excluded this semantic generalization for nouns in a tweet. This is because our task is to discover patterns of how people ask questions, instead of what they ask. 23,277 WordNet features are extracted.

Part-of-Speech Features

Compared to a statement, questions present special patterns of syntactic structure. Therefore we attempt to include syntactic features into consideration. Syntactic parsing of billions of tweets appears to be costly and probably unnecessary, since the quality of parsing is compromised given the inaccurate use of language in social media. We thus seek for features that capture light syntactic information. We first obtain part-of-speech of the words in a tweet, and then extract ngrams of these part-of-speech tags. That is, given a tweet with n words, $w_1, w_2, \ldots, w_n$, we extract grams from the part-of-speech sequence of the tweet, is $t_1, t_2, \ldots, t_n$, and then extract unigrams, bigrams and trigrams from this part-of-speech sequence as additional features of the tweet. 3,902 POS features are extracted in total.

Meta Features

We also include 6 metadata features and simple statistical features of the tweet such as the length of the tweets, the number of words, the coverage of vocabulary, the number of capitalized words, whether or not the tweet contains a URL, and whether or not it mentions other users. We believe these features are possibly indicative of questions.

4.3.2 Feature Selection

The four types of extracted features represent each tweet as a vector with a very large number of dimensions. This is not surprising given the huge and open vocabulary in Twitter. Even though we can reduce the number of features by various heuristics of post-processing, the number of features remaining is still far larger than the number of training examples. Therefore, it is essential to conduct feature selection and further reduce the dimensionality of the data.

In this paper, we adopt the state-of-the-art feature selection method named Bi-Normal Separation (BNS) proposed in [11]. In this work, the author proved that the proposed metric for feature selection outperformed other well-known metric such as Information Gain and Chi-distance. Specifically, let tp and tn be the number of positive cases with and without a given feature, fp and fn be the number of negative cases with and without the feature. Let tpr be the sample true positive ratio (i.e., $tpr = tp/(tp + fn)$) and fpr be the sample false positive ratio (i.e., $fpr = fp/(fp + tn)$).

The BNS metric of a given feature can be calculated by

$$\|F^{-1}(tpr) - F^{-1}(fpr)\|, \tag{1}$$

where F is the Normal cumulative distribution function.

[5]The codebook is made available at `http://www-personal.umich.edu/~zhezhao/projects/IN/codebook.html`

4.4 Training Classifier

After feature selection, we move forward and train four independent classifiers using the Support Vector Machine (SVM) [7], based on each of the four types of features. We then combine the four classifiers that represent four types of features into one stronger classifier using boosting. This is done through the Adaptive Boosting method called Adaboost [12].

Adaboost is an effective algorithm that trains a strong classifier based on several groups of weak classifiers. Usually Adaboost can obtain one classifier better than any of the weak classifiers. However, when the performances of the weak classifiers are higher than a certain level, it is hard to use this algorithm to generate a better classifier. This situation seems to apply to our scenario, since the SVM classifiers are sufficiently strong. In [20], the authors indicated that the reason why this problem occurs is that after several iterations, when the combination of weak classifiers starts to achieve a higher performance, the diversity inside the combination is getting lower. That says, new weak classifiers are likely to make same predictions as the old ones. To solve this problem, they add a parameter to control for the diversity of the weak learners in each iteration. We aslso adopt this technique to combine the four SVM classifiers.

We define parameter div as the threshold of a minimum diversity of a new weak classifier to be added in each iteration in the Adaboost. The diversity that a new classifier could add in iteration t is defined as follows:

$$div_t = \frac{1}{N}\sum_{i=1}^{N} d_t(x_i) \tag{2}$$

$$d_t(x_i) = \begin{cases} 0 & \exists k, f_k(x_i) = f_t(x_i) \\ 1 & \forall k, f_k(x_i) \neq f_t(x_i) \end{cases} \tag{3}$$

Here $d_t(x_i)$ is the diversity of classifier to be added in iteration t to data point x_i. N is the size of the training set. $f_k(x_i)$ is the predicted result of the classifier in iteration k for data point x_i. Our information need detection algorithm uses this modified Adaboost named AdaboostDIV. The diversity of a classifier represents how much new information it could provide to a group of classifiers that have already been trained in Adaboost. This value will be smaller and smaller when there are more classifiers adopted. In each iteration of AdaboostDIV, we examine the diversity of a new classifier. If the diversity of this classifier is higher than minimal threshold div, we accept this classifier into the group of classifiers. Otherwise we terminate the algorithm.

4.5 Evaluation of the Classifier

We train and evaluate our algorithm using the manually labeled set of 3,119 tweets. 10-fold cross validation and the metric of classification accuracy are adopted to evaluate each candidate classifier.

Before feature selection, there are 44,121 ngram lexical features, 23,277 WordNet features, 3,902 Part-of-Speech features, and 6 meta features. In Table 1, we compare the performance of the four SVM classifiers using each of the four types of features and various feature selection algorithms. The findings are consistent with the conclusions in [11]. Feature selection using the BNS metric outperformed two other metrics, namely accuracy (ACCU) and Information Gain, both of which improved over the classifiers without feature selection. Among the four types of features alone, ngram lexical features appear to provide the best performance, while the six meta features provide the weakest result which is also far better than random.

Feature Type	Lexical	WordNet	POS	Meta
Raw	0.745	0.610	0.668	**0.634**
ACCU	0.790	0.673	0.718	/
Information Gain	0.804	0.676	0.723	/
BNS	**0.856**	**0.702**	**0.745**	/

Table 1: Results of SVM classifiers. Lexical features performed the best. Feature selection improved classification accuracy.

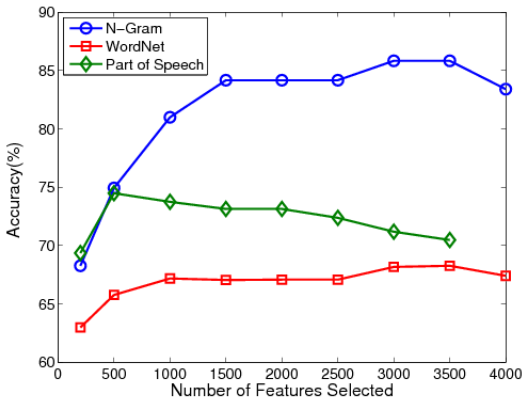


Figure 2: Feature selection using BNS

Figure 2 shows a fine tuning of the number of features selected using BNS. Clearly, when too few or too many features are selected, the classification performance drops because of insufficient discriminative power and overfitting, respectively. Based on our experiment results, we select 3,795 top ranked lexical features, 3,119 top WordNet features, as well as 505 top Part-of-Speech features.

At last, we combined the four SVM classifiers, representing four types of features, using AdaboostDIV. The accuracy of the classifier (with 10-fold cross validation) improved from 85.6% to 86.6%. The small margin suggests that the lexical features are strong enough in detecting information needs, while other types of features add little to the success. Using Adaboost instead of AdaboostDIV compromised the performance, which is consistent to the findings in [20].

Finally, the best performing classifier (four SVM classifiers combined with AdaboostDIV, with feature selection with BNS) is adopted to classify all the tweets in our collection. In our evaluation, the improvements made by feature selection and AdaboostDIV passed the paired-sample t-test at the 5% significance level.

5. ANALYZING INFORMATION NEEDS

After applying the text classifier above to the entire collection of tweets, we detected 136,841,672 tweets conveying information need between July 10th 2011 to June 31st 2012. This is roughly a proportion of 3% of all tweets, and 28.6% of tweets with question marks. With this large scale collection of real questions on Twitter, we are able to conduct a comprehensive descriptive analysis of user's information

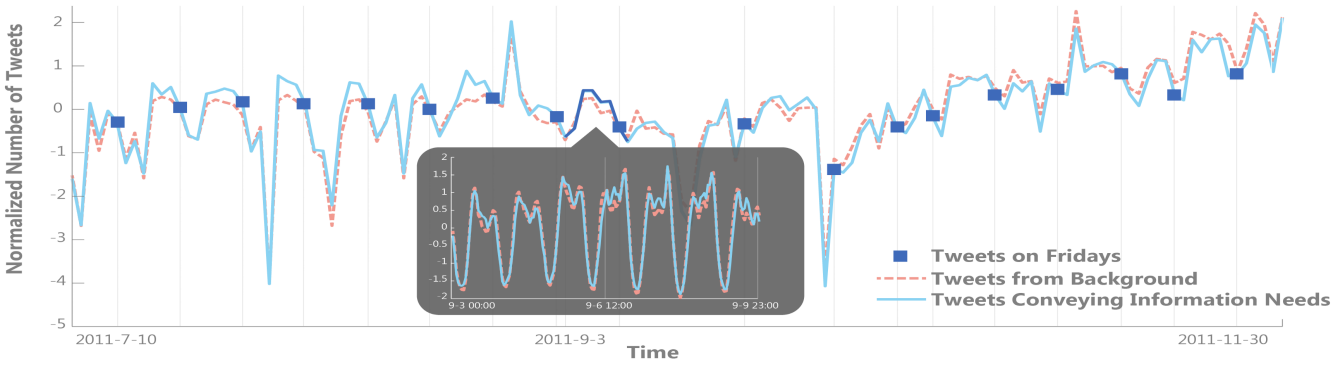


Figure 3: Questions and background tweets over time.

needs. Without ambiguity, we call all the tweets collected as the BACKGROUND tweets, whether they are questions or not. We call tweets that convey information needs as INFORMATION NEEDS (or short as IN), or simply QUESTIONS.

5.1 General Trend

Once we are able to accurately identify real questions (or information needs), the first thing to look at is how many questions are being asked and how they are distributed. Below we present the general trend of the volumes of questions being asked comparing to the total number of tweets in the background. For plotting purposes, we choose to show the trend of the first 5 months from this entire time scope, from July 10th 2011 to November 30th, 2011. Most of the events occurred during this period of time, so plotting the whole year's time series would take more space and cannot be shown distinctly. These 5 months contain a collection of 1,640,850,528 tweets, in which 51,263,378 conveyed an information need. We use this time period for all visualization and time-series analysis below.

Since there is a huge difference between the raw numbers of information needs and the background tweets, we normalize the time series so that the two curves are easier to be aligned on the plot. Specifically, we normalized all the time series using the Z-normalization. That is, for the i^{th} data point valued x_i in the time series, we transform the value by following equation:

$$x_i' = \frac{x_i - \mu}{\sigma} \tag{4}$$

Where μ and σ are the mean and standard deviation of all data points in this time series. This simple normalization doesn't change the trend of the time-series, but allows two series of arbitrary values being aligned to the same range. In the plot, a positive value means the daily count of IN/background tweets is above the average count over time, and a negative value means the count is below the mean. An actual value x on one day indicates that the count of that day is x standard deviations away from the average.

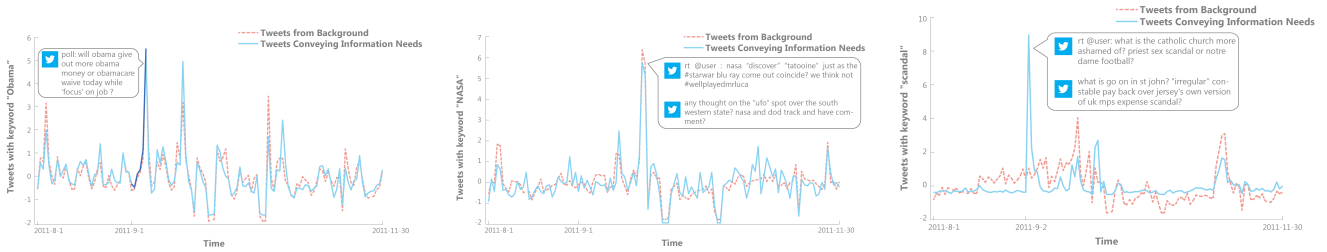
From Figure 3, we observe that both the number of tweets and the number of questions are increasing over time. There are observable but weak days-of-week patterns, which differ search engine logs which present significant weekly patterns (more queries on weekdays than weekends) [23]. The trend is much more sensitive than that of query logs [23], with obvious and irregular spikes and valleys scattered along the time line. This implies that user's information seeking behaviors on Twitter are more sensitive to particular events than the behaviors on search engines. The subfigure presents a strong daily pattern, where both the total number of tweets and information needs peak in late morning and early evening, leaves a valley after noon, and sinks soon after midnight.

In general, the trend of information needs correlates with the trend of the background, which means the information needs on Twitter are likely to be social driven but not information driven. This is not surprising since real world events are likely to stimulate both the demand and supply of information. The more interesting signals in the plot are the noticeable differences between the two curves. On some days there is a significantly overrepresented "demand" of information (i.e., questions) than the "supply" (i.e., background), where there appears a noticeable gap between the two curves. This offers opportunities to analyze what people want, provide better recommendations, and develop propaganda. It is also an interesting observation from the hours-of-day trend that information needs are always overrepresented between the two peaks, before and after noon.

5.2 Keywords

With a sense of the general trend of how people ask, the next question is what people ask. Previous literature has provided insights on the categorization and distribution of topics of questions [26, 29]. Here we are not repeating their efforts, but to provide a finer granularity analysis of the keywords. After removing stopwords from the tweets, we extracted all unigrams and bigrams from tweets classified as conveying information needs. We trim this list of keywords by keeping those appeared every day of our time frame. We believe these keywords that the most representative of the *everyday* information needs of users in Twitter instead of information needs only triggered by particular events. For each of these keywords, we keep the daily count of the number of background tweets containing the keyword and the number of information needs containing the keyword.

With these counts, we can distinguish keywords that appeared frequently in information needs and those appeared frequently in the background. Table 2 lists a subset of keywords that are significantly overrepresented in information needs (i.e., have a much larger frequency in IN than in Background tweets, normalized by the maximum of the two frequencies), compared to the keywords significantly overrep-



(a) Trend of tweets conveying information need with keyword "obama"

(b) Trend of tweets conveying information need with keyword "nasa"

(c) Trend of tweets conveying information need with keyword "scandal"

Figure 4: Trend of tweets conveying information need with different keywords

Frequent in IN	Frequent in BACKGROUND
noyoutube	http
butterfly fall	user video
pocket camera	follow back
Monday	retweet
skype	beautiful
any suggestion	photo
waterproof phone	good night
any recommend	god bless

Table 2: Overrepresented keywords in information needs and background

resented in the background. One can observe from the table that keywords about technology (e.g., "noyoutube," "pocket camera," "skype," "waterproof phone") and recommendation seeking (e.g., "any suggestion," "any recommend") have a high presence in questions while URLs (e.g., "http"), greetings (e.g., "good night," "god bless") and requests (e.g., "follow back") are more frequent in the background. This finding is consistent with the quantitative analysis in literature [26, 29].

We further dropped the keywords that appeared less than 10 times a day in average, from which we obtained 11,813 keywords. For these keywords, we generated time series that represent the demand of information about these keywords, by counting the number of questions and general tweets containing particular keywords everyday.

Figure 4 presents the trends of information needs and background tweets containing three particular keywords, namely "Obama," "NASA," and "scandal." In Fig. 4(a), we can see that the trend of information needs closely correlates with the background, with several noticeable bursting patterns. These patterns generally correspond to real world events. For example, the largest spike around September 8th was correlated with President Obama's speech about the $450 billion plan to boost jobs. Such types of major events are likely to trigger both questions and discussions in online communities, thus have caused a correlated spike of both information needs and the background.

The trends of the keyword "NASA" present a different pattern. The questions and the background align well around the big spike, but disjoin in other time periods. In general, the trend of information needs is more sensitive than the background discussions, presenting more fluctuation. These smallish spikes are not triggered by major events, but rather reflecting the regular demands of information. The trends of the keyword "scandal" is even more interesting. Even the major spikes don't correlate with questions and with

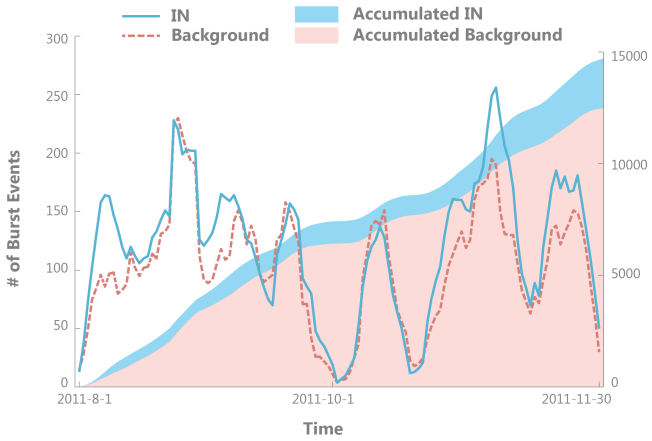


Figure 5: Bursts detected from IN and background

the background. For example, the big spike in information needs was triggered by a widespread cascade of tweets that connects "Priest sex scandal" with "Notre Dame football," which is more like a cascade of persuasion, rumor, or propaganda instead of an real event.

5.3 Burstiness

The anecdotal examples above presented interesting insights in understanding the different roles of bursting patterns in the time series. This is done by comparing individual spikes in information needs with the pattern in the background in the same time period. A different perspective of investigating such bursting patterns is to compare them longitudinally. How many spikes are like the spike caused by Obama's job speech? If a similar bursting pattern can be found among the information needs of a different keyword, that means there is an event that have made a similar impact with the president's speech in terms of triggering the users' behaviors of information seeking.

Literature has thrown light on how to detect real events based on burst detection in social media [38, 37]. In our analysis, we adopt a straightforward solution to detect **similar** burst events in the time series of information needs and the background. Specifically, we select a signature bursting pattern of a real event as a query (e.g., the spike corresponding to Obama's job speech in Figure 4(a)) and retrieve all similar spikes in the time series of other keywords. The

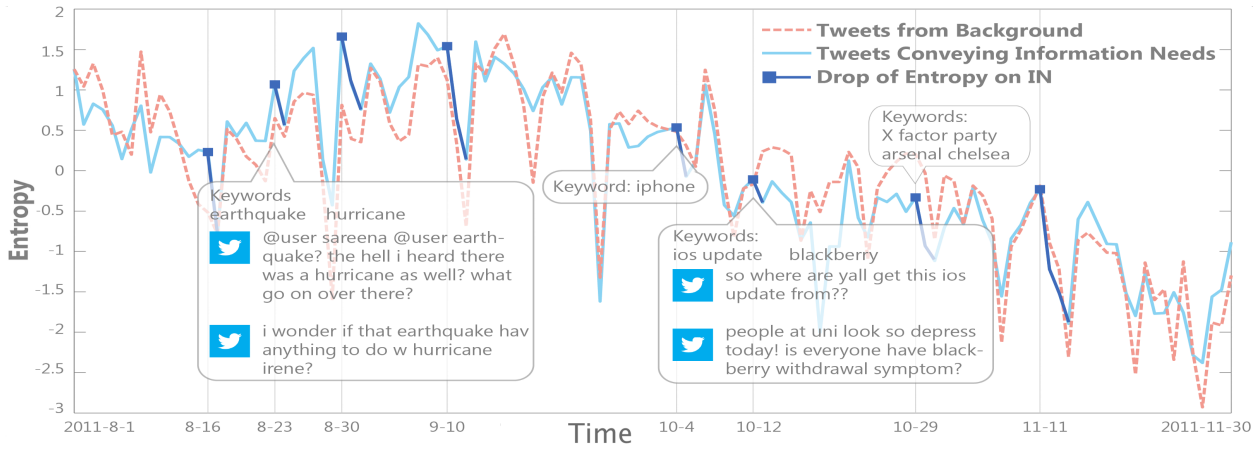


Figure 6: Entropy of word distributions in questions and background

similarity measurement is the Euclidean distance between Z-normalized time series. By doing this, we found 14,640 burst patterns in the time series of information needs and 12,456 burst patterns in the background of all keywords. Figure 5 plots the number of burst events that have a similar impact as the Obama speech, aggregated from the time series of all different keywords. Apparently, there are more such spikes in the time series of information needs rather than in the background, which reassures our finding that the behavior of question asking is more sensitive than the narrative discussions of events. The number of bursting patterns tops in late August and the month of October, which coincides with the two series of events related to "Hurricane Irene" and "Occupy D.C."

5.4 Entropy

The investigation of bursting patterns provides insights about understanding the impact of real events on Twitter users' information seeking behaviors. The impact is featured by the sudden increase of information needs (or background tweets, or both) containing certain keyword. Another way to measure the impact of an event is to look at how it influences the content of information people are tweeting about and asking for. Shannon's Entropy [13] is a powerful tool to measure the level of uncertainty, or unpredictability of a distribution. It is well suited for sizing challenges, compression tasks, as well as the measure of diversity of information. We apply Shannon's entropy to the information needs detected, by measuring the entropy of the word distribution (a.k.a., the language model) in all background tweets and in all questions every day. Clearly, a lower entropy indicates a concentration of discussions on certain topics/keywords, and a higher entropy indicates a spread of discussions on different topics, or a diversified conversation.

Our intuition is that if a major event influences the discussion and information seeking behaviors, the topics in the background or in the questions on that day will concentrate on the topics about that event. Thus we are likely to observe a decreased entropy. Figure 6 plots the entropy of the language models of all information needs, and of all tweets in the background over time. We mark several points in the time series where we observe a sudden drop of entropy on the next day, which indicates a concentration of topics being discussed/asked. We selected these points by the significance of the entropy drop and the differences between the entropy of IN and the entropy of background. We then extract the keywords that are significantly overrepresented in the day after each marked point, which give us a basic idea about the topics that have triggered this concentration. These keywords are good indicators of the actual events that have triggered the concentration (e.g., "the hurricane Irene," "arsenal chelsea" and "the rumor about the release date of iphone 5").

It is especially interesting to notice that on some particular days, entropy drops in information needs but increases in the background. We believe these are very indicative signals for monitoring what the public needs. For example, on October 12th, 2011, there was a sudden drop of entropy in information needs which didn't occur in the background tweets. The discussions concentrated on keywords like "ios," "update," and "blackberry." Indeed, on that day Apple released the new operation system iOS 5, which triggered massive questions about how to get the updates. During the same time, there was a series of outages which caused a shutdown of the Blackberry Internet Service. Such an event has contributed in the concentrations of questions about Blackberry. It is interesting to see that these events about technology indeed had a larger impact in questions instead of the background tweets, which is again consistent with the statistics in literature [26, 29]. Clearly, analyzing the entropy of information needs provides insights on detecting events that have triggered the concentration of information needs. Such discoveries indicate compelling opportunities for search and recommender services, advertising, and rumor detection.

Interestingly, we found entropy analysis not only a powerful tool for macro-level analysis of the impact of events, but also effective in micro-level analysis of the information seeking behaviors of individual users. Indeed, we can also compute the entropy of the distribution of the number of questions that a user asks among different hours of a day. Behaviors of users with a low entropy are more predictable than behaviors of users with a high entropy. Below we show the two behavior patterns from two specific users. One is with high entropy and the other is with low entropy in Figure 7 and 8 respectively. In these two figures, the x-axes represent the 30 days in September, 2011, and the y-axes

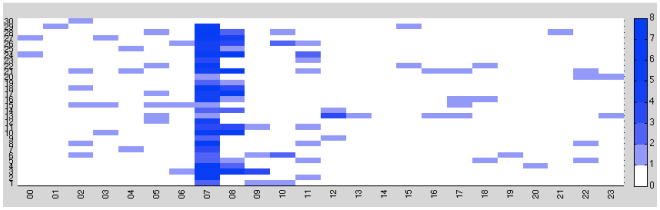


Figure 7: Questions of a user of low entropy.

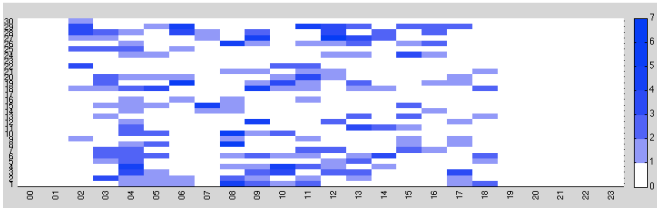


Figure 8: Questions from a user of high entropy.

represent the 24 hours in each day. The different colors in these two figures represent different numbers of posts (the legends are shown on the right side of the figures).

Clearly, the user with low entropy is fairly predictable: he always asks questions at 7am. By looking into his tweets, we found that this is an automatic account that retweets open questions from Yahoo! Answers. The second user is much less predictable, who seemed to be asking questions all over the hours of a day except for the bed time. By looking into his tweets, we found that this is a user who uses Twitter as an instant message platform, who chats with friends whenever he is awake. This user-level analysis on entropy of information needs presents insights on characterizing different individual behaviors.

5.5 Predictive Power

Up to now, we have presented many interesting types of analysis, mostly on the longitudinal patterns of information needs. We see various insights about how to make use of the analysis of information needs in Twitter. Previous literature has visioned the different and complementary roles of social networks and search engines in information seeking. What if we compare the information needs (questions) posted on Twitter and the information needs (queries) submitted to search engines? Is one different from the other? Can one predict the other? If interesting conclusions can be drawn, it will provide insight to the search engine business.

To do this, we compare the trends of information need in Twitter with the trends of Google search queries. Figure 9(a) shows the time series of the Twitter questions containing the keyword "Justin Bieber" and the Google trend of the query "Justin Bieber". We use this query as an example because it is one of the most frequent search queries in Google 2011 and is also contained in a large number of Twitter questions. We can see that information needs in Twitter is more sensitive to bursting events, while the same queries in Google presents a more periodic pattern (e.g., days-of-week pattern).

We then move forward to test whether the information needs from one platform can predict those in the other, using the Granger causality test. The Granger causality test is a statistical hypothesis test for determining whether a time series is useful in forecasting another [15]. In [3], it is used to test whether the sentiment of Twitter users can predict stock market.

Specifically, we selected a subset of keywords and manually downloaded the trends of these keywords as queries submitted to Google [6]. The subset of keywords contains twenty keywords that have a high frequency in background tweets, twenty keywords that have a high frequency in the questions, and twenty keywords from the most popular search queries in Google. To select this subset, we sorted all the keywords by frequency from the three different sources and select the top 20 named entities and nouns. If there is an overlapping keyword from multiple sources, we simply add a new keyword from the source with lower frequency of the overlapping keyword[7]. We then use the Granger causality test to test whether the three trends (Twitter background, Twitter information needs, and Google queries) of each keyword can predict each other. By changing the parameters in Granger causality test, we can test the prediction power of one time series to the other for different lags of time. In this paper, we only show the results with lag of 5 days due to the limitation of space. We obtained similar results with other different lags.

Results show that the trends of information needs in Twitter have a good predictive power in predicting trends of Google queries and are less likely to be predicted by the Google trends. This is measured by "*of how many keywords, one type of time series can predict another type of time series, given certain significance level.*" From Figure 9(b), we see that the information needs in Twitter have a better predictive power than the background in predicting Google trends. From Figure 9(c), we see that the information needs in Twitter have a better predictive power in predicting Google trends rather than the other way around. Between information needs in Twitter and Google trends, the questions in Twitter have a stronger predictive power of Google queries, which successfully predicts the Google trends of more than 60% of the keywords with a significance level of 0.05. Among these keywords, 9 of them are from the popular Google queries. This is a promising insight for search engine practitioners to closely watch the questions in Twitter and improve the search results of targeted queries whenever a bursting pattern is observed.

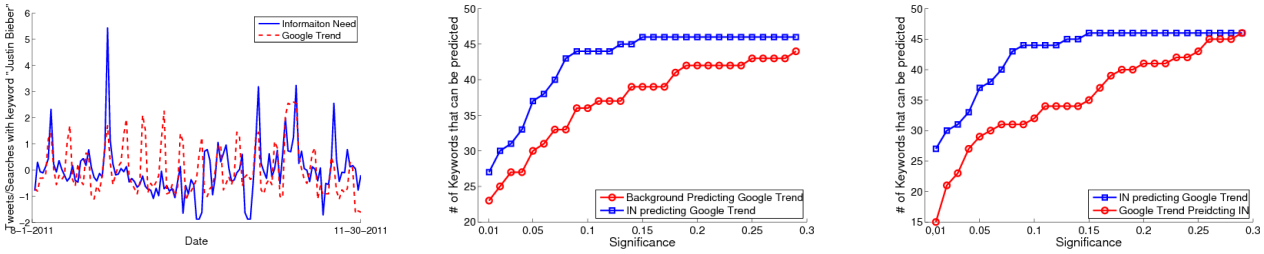
5.6 Implications and Discussion

In this section, we presented various investigations of the questions, or information needs, in Twitter. Most analyses presented are longitudinal, based on the time series of particular statistics and comparisons between the questions and background tweets. The analysis provided interesting implications to social behavior observers, search engine practitioners, and researchers of social search.

To summarize, we confirmed that the behaviors of information seeking (questions) are substantially different from the behaviors of narrative conversations (background) in Twitter. We also find that it differs from behaviors in Web search. Some of the findings reconfirmed the conclusions in literature, such as the overrepresented topics in Twitter

[6]Since we don't have access to real search logs, we used the Google trend: `http://www.google.com/trends/`

[7]The List of the keywords can be found at `http://www-personal.umich.edu/~zhezhao/projects/IN/wlist`



(a) Keyword: Justin Bieber

(b) Background v.s. information needs in predicting Google trends. The higher the better

(c) Information need v.s. Google trends in predicting each other. The higher the better

Figure 9: Twitter information needs can predict search queries.

questions. Interesting patterns emerge when comparing the questions with the background, which implies opportunities for providers of information service. This includes the patterns when the demands (e.g., questions) are significantly higher than the supply (e.g., background), when the information needs concentrate on particular topics, and when the spikes in information needs do not agree with those in the background.

We found that information needs in Twitter are sensitive to real world events (more sensitive than search queries). By comparing the patterns of individual keywords in questions and in the background, we foresee a new and meaningful taxonomy to discriminate the types of information needs. The comparative analysis also provides new ways to differentiate real world events and cascades of persuasion. This implies useful tools to detect propaganda and rumors from social media.

Entropy analysis provided a good way to detect real events and their impact in the topics being asked about in Twitter. It also provided a unique perspective to understand and discriminate the behaviors of individual users. Such analysis implies new tools for business providers and advertisers, which can help them to come up with a better social monetization strategy such as targeted advertising or content recommendation.

With a limited but representative set of keywords and trend information extracted from the Google trend, we found that the information needs in Twitter has a predictive power of search queries in Google. Although this conclusion has to be reevaluated when large-scale query log data is available (which we don't have access to), it implies interesting action moments for search engines. When spikes of information needs are observed in Twitter, the search engine practitioner has time to strategically optimize the search results for the corresponding topics.

Despite the interesting implications, we do see potential limitations of this analysis. For example, all our analysis is done on a **random** sample of tweets. This makes it difficult to answer questions like "*how many questions are answered,*" "*how many questions are distributed (i.e., retweeted),*" or "*consequential user behaviors after information seeking.*" These questions can only be answered with the availability of the complete set of tweets, or subset of tweets sampled in a different way (e.g., all tweets of a subnetwork of users). We leave these questions about questions for future work.

6. CONCLUSION

Information needs and information seeking behaviors through social platforms attracted much interest because of its unique properties and complementary role to Web search. In this paper, we present the first large-scale analysis of information needs, or questions, in Twitter. We proposed an automatic classification algorithm that distinguishes real questions from tweets with question marks with an accuracy as high as 86.6%. Our classifier makes use of different types of features with the state-of-the-art feature selection and boosting methods.

We then present a comprehensive analysis of the large-scale collection of information needs we extracted. We found that questions being asked on Twitter are substantially different from the topics being tweeted in general. Information needs detected on Twitter have a considerable power of predicting the trends of Google queries. Many interesting signals emerge through longitudinal analysis of the volume, spikes, and entropy of questions on Twitter, which provide valuable insights to the understanding of the impact of real world events in user's information seeking behaviors, as well as the understanding of individual behavioral patterns in social platforms.

Based on the insights from this analysis, we foresee many potential applications that utilizes the better understanding of what people want to know on Twitter. One possible future work is to develop an effective algorithm to detect and predict what individual users want to know in the future. By doing this one may be able to develop better recommender systems on social network platforms. With the presumption of accessing large scale search query logs, a promising opportunity lies in a large-scale comparison of social and search behaviors in information seeking. On the other hand, improving the classifier to detect tweets with implicit information need such as tweets that is not an explicit question or without a question mark is also a potential future work. Furthermore, it is interesting to do some user-level analysis, such as studying the predictive power of different groups of users to see whether there exists a specific group of users that contributes to predicting the trend most.

Acknowledgement We thank Cliff Lampe, Paul Resnick and Rebecca Gray for the useful discussions. This work is partially supported by the National Science Foundation under grant numbers IIS-0968489, IIS-1054199, and CCF-1048168, and partially supported by the DARPA under award number W911NF-12-1-0037.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM, 2006.

[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology-EDBT 2004 Workshops*, pages 395–397. Springer, 2005.

[3] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

[4] A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, D. Metzler, L. Riedel, and J. Yuan. Online expansion of rare queries for sponsored search. In *Proceedings of the 18th international conference on World wide web*, pages 511–520. ACM, 2009.

[5] E. H. Chi. Information seeking can be social. *Computer*, 42(3):42–46, 2009.

[6] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun. Finding question-answer pairs from online forums. pages 467–474, 2008.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[8] M. Efron and M. Winget. 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, 2010.

[9] B. Evans and E. Chi. Towards a model of understanding social search. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 485–494. ACM, 2008.

[10] C. Fellbaum. Wordnet: An electronic lexical database. *Cambridge, MA: MIT Press*, 38(11):39–41, 1998.

[11] G. Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, 2003.

[12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting, 1995.

[13] R. Gallager. Claude e. shannon: A retrospective on his life, work, and impact. *Information Theory, IEEE Transactions on*, 47(7):2681–2695, 2001.

[14] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, February 2009.

[15] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, July 1969.

[16] B. Hecht, J. Teevan, M. R. Morris, and D. Liebling. Searchbuddies: Bringing search engines into the conversation. *ICWSM*, pages 138–145, 2012.

[17] R. Jones, R. Kumar, B. Pang, and A. Tomkins. I know what you did last summer: query logs and user privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 909–914. ACM, 2007.

[18] R. Krovetz. Viewing morphology as an inference process. *16th ACM SIGIR Conference*, pages 191–202, 1993.

[19] B. Li, X. Si, M. R. Lyu, I. King, and E. Y. Chang. Question identification on twitter. pages 2477–2480, 2011.

[20] X. Li, L. Wang, and E. Sung. Adaboost with svm-based component classifiers. *Eng. Appl. Artif. Intell.*, 21(5):785–795, 2008.

[21] Z. Liu and B. Jansen. Almighty twitter, what are people asking for? *ASIST*, 2012.

[22] T. N. Mansuy and R. J. Hilderman. Evaluating wordnet features in text classification models. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference*, pages 568–573. AAAI Press, 2006.

[23] Q. Mei and K. Church. Entropy of search logs: how hard is search? with personalization? with backoff? In *Proceedings of the international conference on Web search and web data mining*, pages 45–54. ACM, 2008.

[24] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[25] M. Morris and J. Teevan. Exploring the complementary roles of social networks and search engines. *Human-Computer Interaction Consortium Workshop(HCIC)*, 2012.

[26] M. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1739–1748. ACM, 2010.

[27] M. R. Morris, J. Teevan, and K. Panovich. A comparison of information seeking using search engines and social networks. *Proceedings of 4th International AAAI Conference on Weblogs and Social Media*, 42(3):291–294, 2010.

[28] J. Nichols and J. Kang. Asking questions of targeted strangers on social networks. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 999–1002. ACM, 2012.

[29] S. A. Paul, L. Hong, and E. H. Chi. Is twitter a good place for asking questions? a characterization study. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 18(11):578–581, 2011.

[30] C. Shah. Measuring effectiveness and user satisfaction in Yahoo! answers. *First Monday*, 16(2-7), 2011.

[31] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. In *ACm SIGIR Forum*, volume 33, pages 6–12. ACM, 1999.

[32] J. Teevan, E. Adar, R. Jones, and M. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158. ACM, 2007.

[33] J. Teevan, S. Dumais, and D. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 163–170. ACM, 2008.

[34] J. Teevan, D. Ramage, and M. Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international Conference on Web search and Data Mining*, pages 35–44. ACM, 2011.

[35] K. Wang and T.-S. Chua. Exploiting salient patterns for question detection and question retrieval in community-based question answering. pages 1155–1163, 2010.

[36] J. Yang, M. R. Morris, J. Teevan, L. A. Adamic, and M. S. Ackerman. Culture matters: A survey study of social q&a behavior. 2011.

[37] J. Yao, B. Cui, Y. Huang, and X. Jin. Temporal and social context based burst detection from folksonomies. In *AAAI*. AAAI Press, 2010.

[38] J. Yao, B. Cui, Y. Huang, and Y. Zhou. Bursty event detection from collaborative tags. *World Wide Web*, 15(2):171–195, 2012.