Cognos: Crowdsourcing Search for Topic Experts in Microblogs

ABSTRACT

Finding topic experts on microblogging sites with millions of users, such as Twitter, is a hard and challenging problem. In this paper, we propose and investigate a new methodology for inferring topical experts in the popular Twitter social network. Our methodology relies on the wisdom of the Twitter crowds – it leverages Twitter `Lists`, which are often carefully created by individual users to include experts on topics that interest them and whose meta-data (List names and descriptions) provide valuable semantic cues to experts' domain of expertise. We mined List information to build `Cognos`, an expert search system for Twitter. Detailed experimental evaluation based on a real-world deployment shows that: (a) Cognos infers a user's expertise more accurately and comprehensively than state-of-the-art systems that rely on the user's bio or tweet content, (b) Cognos scales well due to built-in mechanisms to efficiently update its experts' database with new users, and (c) Despite relying only on a single feature, namely crowdsourced Lists, Cognos yields comparable, if not better, results in user tests, as compared to the official Twitter experts search engine for a wide range of queries. Our study highlights Lists as a potentially valuable source of information for future content or expert search systems in Twitter.

1. INTRODUCTION

Microblogging sites, out of which Twitter is the most popular, have emerged as an important platform for exchanging real-time information on the Web. Recent estimates suggest that 200 million active Twitter users post 150 million tweets (messages) daily [1, 17]. These messages contain a wide variety of information, varying from conversational tweets to highly relevant information on niche topics. The users posting these messages range from globally popular news organizations and celebrities to locally popular community organizers or activists and from domain experts in fields like computer science and astrophysics to spammers that fake the identities of well-known users.

As a result, the quality of information posted in Twitter is highly variable and finding the users that are recognized sources of relevant and trustworthy information on specific topics (i.e. topical experts) is a key challenge. Identifying topic experts is also the first step towards finding authoritative information on the topic. Recognizing this, Twitter itself has created a topical expert search system (known as the Twitter Who To Follow (WTF) service [15]). However, as we show later in this paper, the results from this service leave a lot of scope for improvement.

In this paper, we present **Cognos**, a system for finding topic experts in Twitter. Cognos is based on a new methodology for inferring users' expertise. Traditional approaches to identify topical experts in Twitter rely either on the information provided by the user herself (e.g., user bio) [16] or on analyzing the network characteristics and tweeting activity of users [10, 19]. Cognos takes a different approach to identify topical experts in Twitter utilizing *crowdsourced* topical annotation of experts. Specifcally, Cognos exploits the *Lists* feature in Twitter, using which any user can group Twitter accounts that tweet on a topic that is of interest to her, and follow their collective tweets. We observe that many users carefully create Lists to include other Twitter users who they consider as experts on a given topic. Furthermore, they generate meta-data, such as List names and descriptions, that provide valuable semantic cues to the topical expertise of the users included in the List. Our key idea is to analyze the meta-data of the Lists containing a user to infer the user's topics of expertise, which in turn enabled us to identify topical experts.

To build Cognos, we address three key challenges: (1) How to accurately and comprehensively infer individual user's topics of expertise from Lists? (2) How to rank the relative expertise of different users identified as experts on a given topic? and (3) How to crawl the Lists mata-data for hundreds of millions of Twitter users efficiently and scalably? The main contributions of this paper lie in the methodologies we propose to tackle the above challenges.

We present an extensive evaluation of Cognos based on user feedback obtained using a real-world deployment, which can be accessed at `http://139.19.103.35/who-to-follow/`. [1] To summarize a few highlights from our evaluation: We find that Cognos performs as good as or better than the official Twitter WTF service in more than 52% of the queries. Cognos yields particularly better search results in cases in which experts do not have an account bio, or whose bio does not contain information about the user's topic of expertise. Moreover, Cognos rarely produces entirely irrelevant results, unlike the Twitter WTF service whose top results at times include a few users who are not related to the given query, but whose name or bio contains the terms in the query. Furthermore, as Cognos is based on the use of a single and simple feature (Twitter Lists) it is far more scalable as compared to prior approaches, which use computationally intensive machine learning algorithms over graph and content-based metrics [10, 19].

[1] We anonymized the URL to IP address to preserve the anonymity of author's home institutions.

2. RELATED WORK

As the number of users and information shared in Twitter has increased exponentially, different information retrieval tools, such as search [13] and recommender systems [15], are becoming very popular ways to find trend topics, users, and valuable content. A critical component of such mechanisms consists of identifying users who are important sources of information on specific topics (topical experts).

There have been several attempts to measure the influence of Twitter users and hence to identify influential users or experts [3, 4, 8, 12]. However, none of the above mentioned efforts attempts to identify experts in any *specific topic*. To the best of our knowledge, there has been only two efforts that have approached the problem of identifying experts in *specific topics* [10,19]. Weng *et. al.* [19] proposed a Page-Rank like algorithm TwitterRank, that uses both the Twitter graph and processed information from tweets to identify experts in particular topics. On the other hand, Pal *et. al.* [10] used clustering and ranking on more than 15 features extracted from the Twitter graph and the tweets posted by users.

Apart from the above research studies, there also exist some *services* for identifying topical experts in Twitter. Recognizing the importance of searching for experts on specific topics, Twitter itself provides an official "who to follow" (WTF) service [15] where one can search for experts on a given topic (query). Though the exact details of implementation of the service are not publicly known, it is reported that Twitter WTF uses several factors such as the profile information (e.g. name and bio) of users, their social links, their level of engagement in Twitter, and so on [16] to identify topical experts.

It can be noted that all the above approaches primarily rely on the information provided by a user herself (e.g. her account name and bio, the tweets posted by her) and her social graph, to infer the topics in which she is an expert. In contrast, the present work uses an entirely different methodology to infer the topics of expertise of an individual Twitter user, which relies on the 'wisdom of the Twitter crowd' (i.e. how others describe this user), collected through crowdsourced Lists. Further, all of the above mentioned research studies use fixed Twitter datasets collected at a certain point in time. To the best of our knowledge, this study is the first to address the challenge of keeping an OSN-based search / recommender system up-to-date, a challenge that has become essential given the phenomenal rate of increase of population in today's OSNs [2].

Finally, it is important to mention that a few prior studies have used Twitter Lists for different purposes, such as identifying seed nodes for sampling algorithms or topic-sensitive Pagerank-like algorithms [18, 20] or for contextualizing a user [11]. The present study provides an in-depth analysis about Lists and uses Lists for a fundamentally different purpose, as stated above.

3. METHODOLOGY AND CHALLENGES

In this section, we first propose our methodology for finding topic experts using a recently introduced Twitter feature called `Lists`. Later we identify the key design challenges in designing a search system based on the methodology. We address these challenges in the subsequent sections.

3.1 Methodology: Leverage Twitter Lists

Our methodology is based on the Twitter Lists feature. In late 2009, Twitter introduced Lists to help users organize their followings (i.e. the people whom a user follows) and the information they post [7]. By creating a List, a user can group other Twitter users, and view the aggregated tweets posted by all the listed users in the List timeline. When creating a List, a user typically provides a List name (free text, limited to 25 characters) and optionally add a List description. For instance, a user can create a List namely "celebrities" and add celebrities to this List. Then, the user can view tweets posted by these celebrities in the List timeline.

Table 1 presents illustrative examples of Lists, extracted from Twitter users. The key observation here is that the List names and descriptions provide valuable semantic cues to the topic of expertise of the members of the Lists. For example, using List meta-data, we can associate Barack-Obama with Politics and Politicans, Eminem with music and musicians, and Daniel Tunkelang with SIGIR. Thus, Lists provide a way to annotate Twitter users with their topics of expertise. Interestingly, these annotations are generated by arbitrary Twitter users and so they reflect the collective wisdom of the crowds.

Our methodology relies on *extracting the information contained in the crowdsourced Lists to build an expert search system*. Specifically, it has three parts: (i) gather crowd-created Lists for all Twitter users, (ii) mine List meta-data to infer the topical expertise of individual Twitter users, and (iii) for a given query topic, rank the relative expertise of the users, whose topical expertise matches the query.

3.2 Key open questions and design challenges

Our proposed methodology for building a search system for experts in Twitter raises a number of important questions and key design challenges, which we enumerate below:

1. How to infer users' topics of expertise from Lists? Do Lists contain sufficient information to infer the various topics of expertise of individual Twitter users both accurately and comprehensively?

2. How to rank the relative expertise of different users identified as experts on a given topic?

3. How to crawl the Lists meta-data for tens of millions of Twitter users (experts) created by hundreds of millions of other users? How to keep the Lists data up-to-date as several tens to hundreds of thousands of new users join and new Lists are created every single day [2]?

We address the above research challenges in each of the subsequent sections. In Section 4, we describe how we use crowdsourced Lists to infer the topics of expertise of individual Twitter users. In Section 5, we present Cognos, a topical expert search system for Twitter that leverages the topical expertise inferred using Lists to identify experts on a given topic and rank them. In Section 6, we propose efficient strategies that minimize the number of Lists that we need to crawl to keep Cognos system up-to-date. We conduct an extensive evaluation of our proposals by comparing their performance with two systems: (a) the state-of-the-art research system for identifying topical experts in Twitter [10] and (b) the official Twitter Who-To-Follow service [15].

List Name	Description	Members
News	News media accounts	nytimes, BBCNews, WSJ, cnnbrk, CBSNews
Music	Musicians	Eminem, britneyspears, ladygaga, rihanna, BonJovi
Tennis	Tennis players and Tennis news	andyroddick, usopen, Bryanbros, ATPWorldTour
Politics	Politicians and people who talk about them	BarackObama, nprpolitics, whitehouse, billmaher
SIGIR2010	People tweeting from SIGIR 2010	Daniel Tunkelang, Maria Grineva, Ian Soboroff, James Caverlee

Table 1: Examples of Lists, their description, and some members

4. USING LISTS TO INFER EXPERTISE

In this section, we first describe our methodology of inferring the expertise of individual Twitter users and then evaluate the accuracy and expressiveness of the inferred expertise.

4.1 Mining List meta-data to infer expertise

Our strategy consists of extracting frequently occuring topics (words) from the List meta-data (names and description) and associating these topics with the listed users. The intuition behind our strategy is that a user listed by many other users under a certain topic is very likely to be an expert on that topic. Previous efforts that analyzed Twitter Lists showed that *nouns* and *adjectives* in list names and descriptions are particularly useful for this purpose [11]. So our strategy to extract topics from List meta-data consists of the following steps:

1. We first apply common language processing techniques, such as case-folding, stemming, and removal of stop words. In addition to the common stop words, a set of domain-specific words are also filtered out, such as Twitter, list, and formulist (a tool frequently used to automatically create Lists).
2. Since list names cannot exceed 25 characters, users often combine multiple words using *CamelCase* (e.g. TennisPlayers). Thus, we separate these words into individual words.
3. We identify nouns and adjectives using a part-of-speech tagger.
4. As a number of list names and descriptions are in languages other than English, we group together words that are very similar to each other (based on edit-distance among words), e.g. politics and *politica*, journalist and *jornalistas*, etc.
5. As list names and descriptions are typically short, we consider only unigrams and bigrams as topics.

The above strategy produces a set of topics for each user, as well as the frequency with which a topic appeared in the names and descriptions of the Lists containing the user.

4.2 Evaluating quality of expertise inference

When evaluating the quality of inferred expertise, we check for two metrics: (i) accuracy: is the user really an expert in the inferred topics of expertise? (ii) expressiveness: do Lists comprehensively capture all the different topics in which a user has expertise?

For our evaluation, we need to gather ground truth information about Twitter users' expertise. Since such ground truth is difficult to obtain for a random set of Twitter users, we consider the following strategies: First, we evaluate for a select set of *popular* users whose true topics of expertise are generally well-known or easily verifiable. Second, for a given set of topics, we collect the top experts identified by the state-of-the-art research system for identifying topical authorities [10], and by the official Twitter WTF service [15]. We then check if our methodology identifies these users as experts in the given topics. The results not only demonstrate the high quality of our expertise inference, but they also uncover drawbacks of competing state-of-the-art methods.

4.2.1 Inferred expertise for selected popular users

Table 2 shows the top 10 topics (obtained using our List-based method) for Twitter users whose expertise is well-known. It is evident that the main topics accurately describe the topics of expertise of the users. The inference is accurate and comprehensive not only for users with millions of followers, but also for users with hundreds or thousands of followers. For instance, for Mark Sanderson (Program Committee Chair at SIGIR 2012), even though his Twitter account is included in only *12 Lists*, the inferred topics identify that he is a researcher in computer science ("cs"), specializing in information retrieval, machine learning ("ml"), search and so on. Again, for US senators (two examples shown in Table 2 – Chuck Grassley and Claire McCaskill), this methodology could accurately identify a variety of topics, for instance, their political party (Republicans / Democrats), their state, their gender ('women' in case of Claire McCaskill), their political ideology (conservative / progressive) and even a number of the senate committees of which each senator is a member (e.g. 'health' in case of Chuck Grassley). We verified the accuracy of our inference using the Wikipedia pages for these people, and found them to be almost always accurate. Thus, List meta-data is often sufficiently rich to yield very high quality expertise inference for users over a large range of popularity (number of followers).

4.2.2 Comparing with the state-of-the-art research

Next we compare the extent to which the experts identified by a state-of-the-art research system built by Pal *et. al.* [10] can be recalled by our methodology. Pal *et. al.* use more than 15 features extracted from the Twitter social graph and the content of the tweets posted by users to identify topical experts. Though an implementation of this system is not publicly available, their paper lists the top 10 experts identified for three specific topics – *iphone*, *oil spill* and *world cup*. We test whether the topics inferred by our methodology for these experts match with the topic reported by Pal *et. al.*

We find that for a majority of the top 10 experts in each of the three topics, the set of topics inferred by us includes the topic for which they are reported by Pal *et. al.* – for 8 out of 10 for "iphone", for 7 out of 10 for "world cup", and for 6 out of 10 for "oil spill". Table 3 shows some of these experts, along with their bio.

However, for the rest of the cases, the topics inferred using Lists do *not* contain the topic reported by Pal *et. al.*. Table 4 lists these users along with their bios. Examining their bio, it is evident that these users are, in fact, *not* specifically related to the topic of the corresponding query. For

User	# followers	Most frequent topics
Barack Obama	12,481,245	politics, celebs, government, famous, president, news, leaders, noticias, current events
Ashton Kutcher	9,479,352	celebs, actors, famous, movies, stars, comedy, funny, music, hollywood, pop culture
Mark Sanderson	320	information retrieval, ir, cs, ml, semantic, analysis, search, research, nlproc, tech
Chuck Grassley	34,710	politics, senator, congress, government, republicans, iowa, gop, officials, conservative, health
Claire McCaskill	63,687	politics, senator, government, congress, democrats, missouri, dems, officials, progressive, women
BBC News	574,035	media, news, noticias, journalists, politics, english, newspapers, current, periodicos, london
Linux Foundation	46,718	linux, tech, open, software, libre, gnu, computer, developer, ubuntu, unix
Yoga Journal	71,689	yoga, health, fitness, wellness, magazines, media, mind, meditation, body, inspiration

Table 2: The most common topics of expertise of some well-known Twitter users, as identified from Lists

User	Extracts from Bio
Query: iphone	
macworld	Mac, iPod, iPhone experts
TUAW	Unofficial Apple Weblog
Query: oil spill	
kate_sheppard	Reporter covering energy, environment
LATenvironment	Environmental news from California
Query: world cup	
FIFAWorldCupTM	FIFA soccer world cup tweets
itvfootball	News from ITV football

Table 3: Some of the top 10 results reported by Pal *et. al.* [10], for whom the topics inferred using Lists include the query-topic (iphone / oil spill / world cup)

User	Extracts from Bio
Query: iphone	
teedubya	Social Strategy Shaman, SEO
macTweeter	*Account no longer exists in Twitter*
Query: oil spill	
Reuters	latest news from around the world
CBSNews	official Twitter feed of CBS News
TIME	Breaking news and current events
huffingtonpost	The Internet Newspaper
Query: world cup	
nikegoal	marketing, music, education, sport
Flipbooks	News, Random Information
channel4news	exclusive stories & breaking news

Table 4: The top 10 results reported by Pal *et. al.* [10], for whom the topics inferred using Lists does *not* include the query-topic (iphone / oil spill / world cup)

example, a social media entrepreneur and technology blogger *teedubya* was identified as an expert on "iPhone", even though he is not a specialist on Apple products. Similarly, *Reuters*, *CBSNews* and *channel4news* are general news media and authoritative sources of information on a variety of topics, but they are not related specifically to the topics 'oil spill' or 'world cup'. It is likely that the algorithm used by Pal *et. al.* identified these users as experts because a number of their tweets were related to the topic in question during the period when the evaluation was done.

It is worth noting that Pal *et. al.* explicitly set out to discover experts that are not just overtly general and highly followed authorities like popular news media accounts. They highlight the discovery of dedicated specialists that mostly post tweets related to their specialization. Interestingly, our methodology has successfully recalled all such experts (i.e., 100% recall), even though it is based on a single feature (Lists). In comparion, Pal *et. al.* rely on 15 features, which indicates the relative advantages of using crowdsourced Lists to identify users' expertise.

4.2.3 *Comparing with Twitter's official WTF service*

The official Twitter Who-To-Follow (WTF) service helps to search for topical experts for a given topic (query), and is reported to use several factors such as the profile information (e.g. name and bio) of users, their social links, their level of engagement in Twitter, and so on [16] to identify experts. As part of a user survey to evaluate our system (detailed in Section 5.3.2), we obtained the top 20 experts returned by the Twitter WTF service for a few hundred queries generated by users. We investigated the extent to which our methodology would recall these experts.

We find that out of the 3495 users returned by Twitter (top 20 results for some given query), the topics inferred using Lists include the corresponding topic (word in the given query) for 83.4% (2916) of the users. However, the topics inferred by the List-based methodology for the other 16.6% (579) users did *not* contain the topic (word) in the query. To understand these missing experts better, we manually verified 50 randomly selected users out of the 579 users.

We found 9 out of these 50 users (i.e. 18%) to be relevant experts on the query topics. Our methodology infers topics very similar to the query, but none matching the exact query-word. Table 5 shows two such examples. For the official Twitter account of the 'dineLA' restaurant, the inferred topics include 'food' and 'restaurant' but not the query-word 'dining' (for which it was returned by Twitter WTF). Similarly, for the Twitter user 'HubbleHugger77' who is a space explorer and directed the film 'Saving Hubble', we identify 'space', 'cosmology' and 'nasa' but not the query-word 'hubble'. This would appear to suggest that a user's name and bio occasionally contain clues to the user's expertise.

However, in 29 out of the 50 cases (i.e. 58%), we found that the official Twitter WTF service returns *wrong* results, i.e., the returned user is not at all related to the topic of the query for which he is returned. Interestingly, this is most possibly because the query-word appears in the name or bio of the user. For instance, the well-known comedian Jimmy Fallon has (mockingly) described himself as an astrophysicist in his bio, as a result of which he shows up in the top 20 Twitter WTF results for the query 'astrophysicist'. Table 5 shows other examples of users who are wrongly included within the top 20 results returned by Twitter WTF. We were not able to infer the relevance of the expert to the query in the remaining 12 out of the 50 (24%) manually verified user accounts, as we found the query to be ambiguous.

Thus, not only does our methodology recall a vast majority (83.4%) of the experts identified by the official Twitter WTF, but also a majority of the missing experts were incorrectly identified by Twitter. Our List-based methodology fails to recall only a small fraction of experts who are actually related to the given query, and even in those cases, we

Query	User	Extracts from Bio	Major topics obtained from Lists
Users for whom topics inferred from Lists contain very similar words but not the exact query-word			
dining	dineLA	official Twitter account of dineLA	restaurant, food, los angeles, chefs, recipes
hubble	HubbleHugger77	Space Explorer, Director of Film Saving Hubble	science, tech, space, universe, cosmology, nasa
Wrong results in Twitter WTF top 20 results			
astrophysicist	jimmyfallon	astrophysicist	celebs, comedy, funny, actors, famous, humor
cooking	danecook	When I tweet, I tweet to kill	celebs, comedy, funny, famous, actors
origami	ScreenOrigami	Web developer from Germany	webdesign, webkrauts, html, designers

Table 5: Examples of (i) users for whom topics inferred from Lists contain very similar words but not the exact query-word (ii) wrong results within Twitter WTF top 20 results.

identify topics that are quite similar to the query word.

4.2.4 Summary

Our evaluation demonstrates that our proposed methodology of utilizing crowdsourced List meta-data provides an accurate and comprehensive inference of topics of expertise of individual Twitter users. We also show that in many cases, the List-based methodology is more accurate, as compared to the existing techniques of inferring topics of a user from his profile data or his tweets. In the next section, we describe how we utilize the topics inferred using Lists, to build a search system for topical experts in Twitter.

5. COGNOS EXPERT SEARCH SYSTEM

In this section, we leverage our previously discussed methodology to infer users' expertise to build *Cognos* [2], a search system for topical experts in Twitter. Cognos using crowdsourced Lists as the *only* source of information and so its performance illustrates the potential uses of Lists in finding experts. We first describe how we rank experts in Cognos and then present an extensive evaluation of the Cognos system.

5.1 Ranking experts

Ranking of users related to a given topic is a well-studied problem, and over the years, several ranking algorithms have been proposed for the Web [6], online topical communities [21], and even for topical experts in Twitter [10,19]. The expert ranking schemes in Twitter take into account several metrics extracted from the social graph and the content of the tweets posted by users. In contrast, we decided to evaluate a ranking scheme that is based solely on the Lists feature, since one of our objectives is to evaluate crowdsourced Lists as the *only* source of information for topical experts – we have already shown that Lists can be used to accurately infer topics of expertise, now we investigate whether Lists are also an effective metric to rank topical experts.

Using the method described in the previous section, we obtain for each individual user, a set of topics as well as the frequency of occurrence of each topic in the names and descriptions of the Lists containing the user. Thus, for each user we obtain a vector of topics and we store this in a database. Given a query, we compute a topical similarity score between the topic vector for a user and the given query vector, using the algorithm in [5] which computes the *cover density ranking* between the vectors. We chose this similarity score (which is suited to queries containing one to three terms) since queries to expert search systems are almost always short, hence using cosine similarity on tf-idf based representations may not be very effective [9, 10]. Finally, we multiply the topical similarity score for a user with the logarithm of the number of Lists containing the user – the intuition behind this is that a user who is included in more number of Lists (by other users) is likely to be more popular in Twitter.

Thus, given a query (topic), Cognos identifies the set of experts related to the topic using the List-based methodology discussed in Section 4, and then ranks them using the algorithm described above. In the remainder of this section, we extensively evaluate this List-based methodology of identifying and ranking topical experts in Twitter.

5.2 Building the Cognos experts database

To populate the Cognos expertise database, we started to crawl all the Lists containing all Twitter users. We quickly realized that a brute-force crawl of all Lists for all users would be prohibitively expensive and would not scale. So we only crawled the Lists containing all the 54 million Twitter users in a complete snapshopt of the Twitter social network taken in August 2009 [4]. This is only a small fraction of the estimated 465 million Twitter users as of January 2012 [2]. We address the challenge of crawling Lists efficiently and scalably to include experts that joined after 2009, in Section 6.

Of the 54 million Twitter users, we found that **6,843,466** users have been listed at least once. In order to reliably infer topical expertise of a user from Lists, it is important that a user has been listed at least a few times. So we considered only the **1,333,126** users who were listed at least 10 times. Due to rate-limitations in accessing the Twitter API, we collected the information of at most 2000 Lists for a given user. Overall for the 1.3 million users, we gathered a total of 88,471,234 Lists. Out of these, 30,660,140 (34.6 %) Lists had a description, while the others had only the List name.

5.3 Evaluating Cognos expert search system

Judgements on the quality of the results returned by a search system are to an extent subjective. So we chose to evaluate Cognos through an extensive user study where a set of human evaluators judged the relevance of the results returned by Cognos, using a web-based feedback service (available at `http://139.19.103.35/who-to-follow/`) [3]. We also gathered another set of human evaluations where the results returned by Cognos were directly compared with those returned by the official Twitter WTF service [15]. We also compared the top experts returned by Cognos with those

[2]The name is derived from the word *cognoscenti*, i.e. people who are considered to be especially well informed about a particular topic.

[3]The URL has been anonymized due to double blind process.

Category	Sample queries
News	politics, sports, entertainment, science, technology, business
Journalists	politics, sports, entertainment, science, technology, business
Politics	conservative news, liberal politicians, USA / German / Brasilain / Indian politicians
Sports	F1, baseball, soccer, poker, tennis, NFL, NBA, Bundesliga, LA Lakers
Entertainment	celebrities, movie reviews, theater, music
Hobbies	hiking, cooking, chefs, traveling, photography
Lifestyle	wine, dining, book club, health, fashion
Science	biology, astronomy, computer science, complex networks
Technology	iPhone, mac, linux, cloud computing
Business	markets, finance, energy

Table 6: The 55 sample queries used for evaluation of Cognos.

returned by the state-of-the-art research system [10].

The above URL was publicly advertised to all people in three academic institutes located across three different continents, inviting a few hundred people at each of the institutes to evaluate the system. It is to be noted that we preferred such an *in-the-wild* evaluation (instead of a controlled evaluation, e.g. with a fixed set of evaluators and few selected queries, as used by [10]) since this actually resembles a real-world deployment of the search system.

5.3.1 Evaluating quality of Cognos results

In this evaluation, an evaluator issues a query, for which she is shown the top 10 results returned by Cognos. Then the evaluator gives a binary judgement on each of the top 10 results as to whether it is relevant to the given query. The queries used for the evaluations could be selected from a given set of 55 sample queries spread over the 10 categories shown in Table 6. Fig. 1 shows the distribution of the number of times each query was asked, the 5 most frequently asked queries being "computer science", "cloud computing", "movie reviews", "technology news", and "travelling". In the rest of this section, we use the term 'evaluation' to indicate a relevant / non-relevant judgement for an individual result given by Cognos for a particular query.

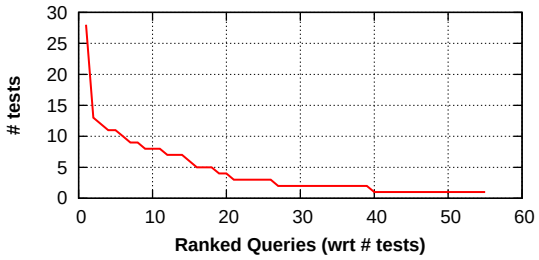

Figure 1: Distribution of the number of times a query was asked (out of 55 sample queries); queries ranked w.r.t. this number

Overall, we obtained 2136 relevance judgements[4] over the top 10 results for the 55 sample queries, out of which 1680 (78.7%) judged the result (topical expert shown by Cognos) to be relevant to the query. We found that the fraction of evaluations that judged a result as relevant, for each individual rank out of the top 10 (i.e. considering the results shown at a certain rank for any of the 55 queries) to be largely invariant – the top 4 results were judged to be relevant in more than 80% of the evaluations, while the results ranked 5–10 were judged relevant in more than 75% of the evaluations.

[4]Despite our request, some of the evaluators did not evaluate all 10 results for a particular query.

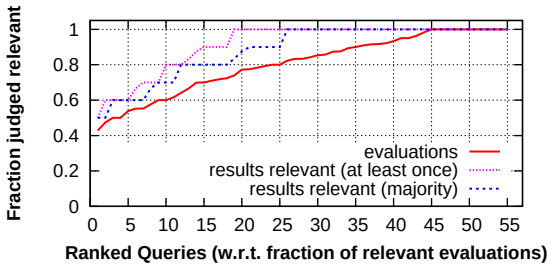

Figure 2: Fraction of evaluations / individual results that were judged relevant (queries ranked w.r.t. fraction of relevant judgements)

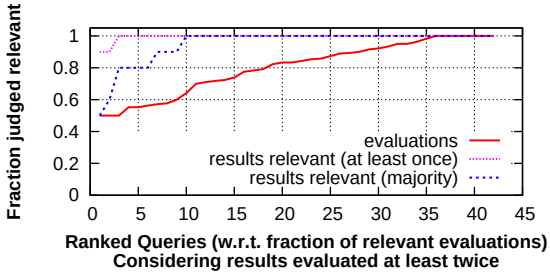

Figure 3: Fraction of evaluations / individual results that were judged relevant (queries ranked w.r.t. the fraction of relevant judgements) – considering only those results for a query, which were evaluated at least twice

Next we examined the Cognos results that received the 456 (21.3%) 'non-relevant' judgements. We found that a large amount of subjectivity in these judgements driven by whether a particular user recognizes another user as a top expert on a given topic. We found a number of cases where the same result for the same query was judged relevant by some evaluator and non-relevant by others. For example, for the query 'cloud computing', Werner Vogels, who is one of the principal architects of Amazon's approach to cloud computing, was rated as relevant in 4 evaluations, and as non-relevant in 6 evaluations, possibly because the name was unknown to these evaluators.

To understand the subjectivity in our judgements, we consider for each particular query, (i) what fraction of evaluations judged a result for this query as relevant, (ii) what fraction of the top 10 results were judged relevant at least once, and (iii) what fraction of the top 10 results were judged

Cognos results			Results by Pal *et. al.*		
User	**Extracts from bio**	**followers**	**User**	**Extracts from bio**	**followers**
Query: oil spill					
BP_America	BP America	35,505	NWF	National Wildlife Federation	76,796
TheOilDrum	energy, peak oil, sustainability	26,257	TIME	Breaking news, current events	3,231,359
GOHSEP	Emergency Preparedness	5,295	huffingtonpost	The Internet Newspaper	1,574,848
usoceangov	National Ocean Service	37,866	NOLAnews	Latest news and updates	29,433
USCG	US Coast Guard	20,513	Reuters	Latest news	1,491,852
Query: iphone					
p0sixninja	iPhone Hacker	127,631	macworld	Mac, iPod, iPhone experts	182,248
iH8sn0w	made f0recast, iREB, iFaith	105,015	Gizmodo	Technologies that change	347,667
chronicdevteam	Hax	107,541	macrumorslive	Updates from Apple events.	170,813
MuscleNerd	iPhone hacker	330,625	macTweeter	*Account not found in Twitter*	
iPhone_News	iPhone news and notes	153,024	engadget	Twitter account of Engadget	419,583
Query: world cup					
worldcupscores	Live 2010 World Cup Scores	10,866	TheWorldGame	Australia's football website	11,541
EdsonBuddle	Soccer playerFC Ingolstadt	30,808	GrantWahl	Sports Illustrated writer	180,290
thefadotcom	Website for England Football	102,536	owen_g	Guardian's Olympics editor	14,930
nytimesgoal	New York Times Soccer Blog	11,699	guardian_sport	Sport news from Guardian	121,095
herculezg	US National Team Forward	31,454	itvfootball	News from ITV football	54,395

Table 7: Top 5 results by Cognos and by Pal *et. al.* [10] for the three queries evaluated by Pal *et. al.*, along with their bio and number of followers

relevant in the *majority* of evaluations. Fig. 2 shows the distribution of these fractions for all queries (where queries are ranked by the fraction of evaluations that judged a result as relevant). It can be seen that for 37 out of the 55 queries, every result was judged relevant by at least one evaluation, and for 30 out of the 55 queries, every result was judged relevant by the majority of the evaluations for that particular result.

The effects of subjectivity is seen even more clearly in Fig. 3 where we plot the above three fractions for each query, considering only those results that were evaluated at least twice. Note that there are 13 queries (out of the 55) for which no individual result was evaluated twice, hence Fig. 3 shows the other 42 queries. For as many as 40 out of these 42 queries, every result (that was evaluated at least twice) was judged relevant by at least one evaluation, and for 33 out of these 42 queries, every result (that was evaluated at least twice) was judged relevant by the majority of the evaluations for that result.

The above statistics show that a vast majority of the results returned by Cognos were judged topically relevant to the given query (topic) by at least some evaluators. Thus, Cognos can successfully identify relevant experts over a wide variety of topics.

5.3.2 Comparing Cognos with state-of-the-art research system

As discussed in Section 4, Pal *et. al.* [10] list the top 10 experts identified by their algorithm for three specific queries: *oil spill*, *iphone*, and *world cup*. For these queries, Table 7 compares the top 5 results from Cognos and the top 5 results reported by Pal *et. al.*, along with the bio and number of followers of each user. Note that while the top results reported by Pal *et. al.* contain some general news media sites (as also discussed in Section 4.2.2), the top Cognos results are much more topic-specific, even if they are not as popularly followed as the news media sities. Interestingly, in their paper, Pal *et.al.* explicitly set out to discover such specialized topic-specific experts, even if they are highly visible. Cognos achieves this goal better than the state-of-the-art system.

Given that Cognos uses only a single feature as compared to more than 15 network and content-based features used by Pal *et. al.* [10], these results further demonstrate the potential of crowdsourced Lists in identifying topical experts in Twitter.

5.3.3 Comparing Cognos with Twitter WTF

In this evaluation, when an evaluator issues a query, she is simultaneously shown the top 10 results returned by Cognos as well as the top 10 results returned by the official Twitter WTF service for the same query. results are anonymized, i.e. the evaluator is not told which result-set is from which service, in order to prevent bias in judgement. Then the evaluator indicates which set of results is better for the given query, or whether both result-sets are equally good or equally bad [5]. It is to be noted that since Cognos uses a Twitter dataset crawled in 2009 (see Section 5.2), for this comparison to be fair, we filtered out from the Twitter WTF results those user-accounts which were created *after* 2009 [6]. In order to test the performance of Cognos 'in-the-wild', we allowed the evaluators to issue any query of their choice.

We obtained relevance judgements for 325 total queries of which 259 are distinct. These queries are evaluator-chosen and they cover a wide variety of topics. Given the high subjectivity observed in user relevance judgements in the previous section, we choose to focus our evaluation on the 27 distinct queries that were asked at least two times. In total, these 27 queries were asked 93 times.

Table 8 shows the 27 queries that were asked at least twice. For each query, we consider the verdict – Cognos better / Twitter WTF better / tie – based on majority voting. The queries for which there was a unanimous verdict (i.e. all evaluations for this query agreed that one was better) are italicized in Table 8. Cognos was judged to be better for 12 out of the 27 queries, while Twitter WTF was judged better for 11, and there was a tie for 4 queries. The fact that Cognos was judged to be better than the official Twitter WTF service for 44% of the queries, clearly indicates the potential

[5]The search engines corresponding to the result-sets are revealed to the evaluators *after* the evaluation is done.

[6]The date on which an account was created is available from the profile information.

	Cognos better	Twitter WTF better	Tie
Queries	Linux, computer science, mac, India, Apple, Facebook, *internet*, ipad, markets, *windows phone*, photography, politic journalist	politic news, music, *Sachin Tendulkar*, Twitter, *Alka Yagnik*, *Anjelina Jolie*, cloud computing, *Delhi*, *Harry Potter*, metallica, ***	Microsoft – Cognos better: 1, Twitter better: 1, both good: 1, both bad: 1 Dell, Kolkata – Cognos better: 1, Twitter better: 1 Sanskrit as an official language – both bad: 2
Average overlap in top 10 results	**1.83**	**2.1**	**3.0**

Table 8: Evaluator-chosen queries (which were asked at least two times) for comparison of Cognos and Twitter WTF, where the verdict (Cognos better / Twitter better / tie) is given by majority voting. Queries in italics are the ones for which there is a unanimous verdict. One query is not shown as it is the name of one of the institutions of the authors of this paper (due to the double blind process)

Cognos results			Twitter WTF results		
User	**Extracts from bio**	**followers**	**User**	**Extracts from bio**	**followers**
Query: music					
Katy Perry	i kissed a girl ...	15,016,823	iTunes Music	Music updates for U.S.	1,903,343
Lady Gaga	mother monster	19,203,867	YouTube	YouTube news, trends, videos	9,220,791
taylorswift13	*Bio not written*	10,994,066	SonyMusicGlobal	home of Sony Music	102,753
jtimberlake	Official Justin Timberlake	8,451,967	50cent	It's the kid 50 Cent	5,861,243
Pink	it's all happening	7,128,708	guardianmusic	Squashing music	107,167
Query: windows phone					
BrandonWatson	developers on Windows Phone	12,462	Windows Phone	Official Windows Phone	130,925
wmpoweruser	Windows Phone Power Users	10,402	pocketnow.com	Windows Phone news	42,134
Charlie Kindel	Founder, CTO, Mentor	8,026	WP Dev Team	Windows Phone Dev Team	37,344
joebelfiore	Runs team doing W. Phone 7	15,542	WindowsPhoneNL	Windows Phone in Nederland	2,785
pocketnow.com	Windows Phone news	42,134	WPCentral	All thing Windows Phone 7	18,266

Table 9: Top 5 results by Cognos and by Twitter WTF for the queries "music" and "windows phone". While top Cognos results mostly contain personal accounts, top Twitter WTF results mostly contain organizations / business accounts.

of crowdsourced Lists (the only feature used in Cognos) in identifying topical experts in Twitter. It can be noted that a significant fraction of the cases where Twitter was unanimously judged better are names of individuals (celebrities) or organizations. Since such names appear very rarely in the List names / descriptions, Cognos does not handle these queries well.

It can also be noted from Table 8 that the top 10 Cognos results show very low overlap with top 10 Twitter WTF results across all queries. This is in spite of the fact that 83.4% out of the Twitter WTF top 20 results for some query (topic), were inferred by out List-based methodology to be related to the same topic (as reported in Section 4). This implies that the low overlap between the top Cognos results and Twitter WTF results is primarily due to the List-based *ranking* used in Cognos. We observe that in general, the top Twitter WTF results mostly include organizations / business accounts while the Cognos top results mostly include personal accounts. We present some examples in Table 9 for the queries "music" (for which the majority voted Twitter WTF better), and "windows phone" (for which the majority voted Cognos better). This is possibly because the Twitter WTF considers the name and bio of users [16], and organizational / business accounts are more likely (compared to personal accounts) to have names or bios which contain terms related to their topics of expertise. As such, these examples again bring out the subjective nature of human judgement, where some evaluators preferred the personal accounts while others preferred the organizational accounts.

5.3.4 Summary

Our evaluation of the Cognos search system shows that a vast majority of its results are relevant for a wide variety of topics. In fact, Cognos rarely produces irrelevant results for user queries. Comparing Cognos with state-of-the-art research system by Pal *et. al.* and official Twitter WTF service highlights the advantages of relying on crowdsourced Lists to identify experts. Cognos yields particularly better search results in the cases when the bio or tweets posted by a user does not correspond to or contain informaiton about the user's topic of expertise. In fact, Cognos performs as good as or better than the official Twitter WTF service for more than 52% of the queries, even though it is based on a single and simple feature (Lists).

6. FINDING EXPERTS EFFICIENTLY

In the section, we address the practical challenge of keeping our Cognos system up-to-date, even as hundreds of thousands new Twitter accounts and new Lists are created every day.

6.1 Scalability problem with crawling Lists

We begin by analyzing the scalability of a simple updation strategy that relies on periodically crawling all the Twitter users and the Lists that contain them. Recent reports indicate that 200 million new users joined Twitter in the last 9 months [2], which roughly amounts to 740,000 new users joining per day. Twitter rate-limits the number of profile crawls from a single machine (IP address) to 150 API requests per hour [14], i.e., to 3600 user profile crawls per day. For each user, we would need to make at least one extra request to crawl her Lists. In fact, Twitter returns only 20 Lists per request. For instance, for a user with more than 2000 lists, it would be necessary to make 100 requests to

Twitter API. Thus, just to keep the system up-to-date, a lower-bound rate limit would be of at least 1,480,000 requests per day. Fortunately, three of our machines were *white-listed* by Twitter, which allows each of them to crawl at a significantly higher rate of 20,000 user profiles per hour. Thus, we can fetch at most 1,440,000 (20,000 × 3 × 24) user profiles per day from all three of our white-listed machines. Note that our maximum crawl rate is still lower than the lower-bound rate we would need to gather the Lists of all new users joining Twitter. Given that we would need to periodically crawl the new Lists for the already existing 465 million Twitter users [2], it becomes quite evident that our simple strategy of crawling all users' Lists would not scale.

Next, we estimated the number of highly listed users amongst the 465 million Twitter accounts as of January 2012. Since Twitter assigns userids in an integer sequence starting from 1, we took a random sample of 300,000 integers in the range 1 to 465 million, and attempted to crawl the profiles of Twitter userids in the sample. The distribution of experts within this large random sample can be expected to be similar to the distribution of experts among all Twitter users. For instance, only 363 out of the 300,000 sampled users (i.e. 0.12%) were Listed 100 or more times; hence we expect the total number of Twitter users who are Listed 100 or more times to be 0.12% of the entire Twitter population. Thus, only a small fraction of all Twitter users are highly listed experts and once they are identified, it would be possible to crawl the Lists containing these experts periodically. The key challenge, however, lies in efficiently identifying these experts from the large Twitter user population.

6.2 Crawling experts efficiently

Our discussion above showed that we cannot exhaustively crawl Lists for all Twitter users. However, we can crawl Lists for the small fraction of *expert* users, if we somehow identified them from the Twitter user population. We now propose and evaluate a strategy to efficiently identify expert users.

We begin by observing that the Twitter social network consists of a number of *hubs*, users who follow a large number of popular experts and include them in Lists. Our strategy is to first identify popular hubs in an older snapshot of the network (when the network was considerably smaller) and then leverage the Lists created by the top hubs in order to find new authorities. It can be noted that this strategy also relies on crowdsourcing – we expect the Twitter crowd (in particular, the top hubs) to discover experts who newly join Twitter, and we can utilize their discovery by periodically crawling the Lists created by the top hubs.

We used the well-known HITS algorithm to identify the top hubs in the snapshot of the Twitter network gathered in 2009 [4] (introduced in Section 5.2), when the network had only 54 million users. We then crawled the Lists created by the top *1 million* hubs in the network to efficiently discover experts. In all, the top 1 million hubs had created 479,129 Lists, which taken together contained 4,100,367 unique users. Out of these, 2,064,373 (i.e. 50.3%) have been included in 10 or more Lists. In comparison, only 1.13% of all the users in our large random sample of Twitter users are listed 10 or more times. The difference clearly indicates that our strategy is effective in focusing our crawls on experts in Twitter. Also, the crawl for the top 1 million hubs took about 3 weeks (January 20 – February 8, 2012) using the machines whitelisted by Twitter, and hence can be repeated every month to discover new experts.

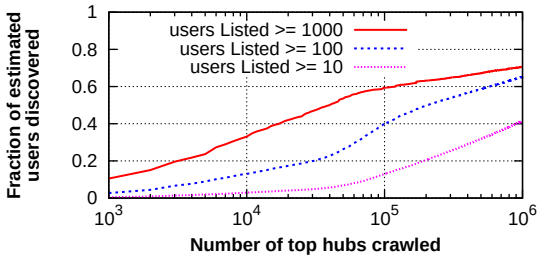

Figure 4: Fraction of estimated number of experts who are included in at least K Lists, that is discovered in the hub-based crawl, for K = 10, 100, 1000.

6.3 Evaluating coverage of our crawls

In this section, we estimate the fraction of experts covered by our strategy to crawl Lists created by top hubs.

6.3.1 Coverage of most listed users

We measure the fraction of most Listed users in Twitter, that is covered by our methodology as follows. First, we estimate the number of Twitter users listed at least K times by computing the number of such users in our 300,000 random sample of users, and then scaling it to the total Twitter user population of 465 million users. Next we calculate the fraction of the estimated number of users Listed at least K times, that are actually discovered by crawling the Lists created by the top hubs.

Figure 4 plots the fraction of experts discovered, against the number of top hubs crawled. We find that by crawling the Lists created by the top 1 million hubs, we discovered 25,887 experts who are Listed 1000 or more times, which is 70.6% of our estimated total number of experts Listed at least 1000 times in Twitter. Further, we find that crawling the Lists created by only the top 100,000 hubs is sufficient to discover 53.3% of the estimated number of experts Listed 1000 or more times in Twitter. Thus, the hub-based updation methodology can be used to efficiently discover a large fraction of new experts in Twitter.

6.3.2 Coverage of newly joined experts

Account	Bio / Description	Listed	Created
MartiRiverola	F.C.Barcelona	67	Feb 6
annekirkbride	English Actress	23	Feb 4
AaronAStanford	Canadian Actor	32	Feb 1
Shay Given	Ireland goalkeeper	107	Jan 27
CourteneyCox	American actress	294	Jan 24
PMOIndia	Prime Minister India	309	Jan 23

Table 10: Examples of very recently created expert accounts discovered by our Hub-based crawl (which ended on Feb 8, 2012)

Our expert discovery strategy is effective in discovering newly joined experts. For example, even though our top hubs were selected using a 2009 snapshot of the Twitter network, more than 42.3% of the 4,100,367 users in the Lists created by these hubs have joined Twitter after 2009. Further, we show some examples of very recently created Twitter accounts that our hub-based crawl could discover, in Ta-

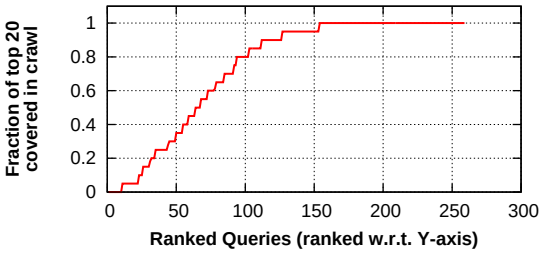

Figure 5: Distribution of the fraction of the Twitter WTF top 20 results that is covered in our hub-based crawl (for the queries discussed in Section 5)

ble 10. Our crawl of Lists created by the top 1 million hubs, which ended on February 8, 2012, discovered some experts who joined Twitter as recently as Feb 6 or Feb 4 (i.e. while the crawl was going on). This validates our hypothesis that the top hubs quickly discover newly joined experts and add them to Lists, and hence shows the effectiveness of the hub-based updation strategy.

6.3.3 *Coverage of experts identified by other systems*

We evaluate whether our updation methodology can discover topical experts returned by the Pal *et. al.* research system and Twitter WTF service. Out of the 30 topical experts stated by Pal *et. al.* (for the three topics "oil spill", "world cup" and "iPhone"), 29 are included in the crawls of Lists created by the top 1 million hubs (the remaining account no longer exists in Twitter). Next, we consider the top 20 experts returned by Twitter WTF service for all the 259 queries obtained by our user-survey (discussed in Section 5) and calculate what fraction of these experts are covered by our hub-based crawls. Figure 5 plots the distribution of the fraction of Twitter WTF top 20 results included in our hub-based crawls, across all queries. It is seen that our crawls include all Twitter WTF top 20 results for more than 50% of the queries and at least 15 out of the Twitter WTF top 20 results for close to 80% of the queries.

The above results indicate that the hub-based strategy – periodically discovering experts through the Lists created by top hubs – can be used to efficiently discover newly joined experts (even very recently joined ones), and thus keep an expert search system up-to-date in the face of rapid increase in the Twitter population.

7. CONCLUSION

As Twitter emerges as a popular platform for users to search for interesting topical content, an important research challenge lies in identifying experts in specific topics. In this paper, we show that an effective solution to this hard problem lies in exploiting wisdom of the Twitter crowds. We observe that individual Twitter users, for their own convenience, annotate and classify experts in various topics using the Lists feature. We show that by aggregating the *List* information for a Twitter user, we can discover an extremely rich and varied characterization of the topical expertise of the user as perceived by the Twitter crowds. Based upon this methodology, we build and deploy Cognos, a topical expert search system. Through extensive evaluation, we demonstrate that even though Cognos is built utilizing *only* the Lists feature, it can compete with the commercial who-to-follow system deployed by Twitter itself. We believe that crowdsourced Lists provide a valuable foundation for building future content search / recommendation / discovery services in Twitter.

8. REFERENCES

[1] There Are Now 155m Tweets Posted Per Day, Triple the Number a Year Ago. `http://rww.to/gv4VqA`, April 2011.

[2] Twitter to hit 500 million accounts by February. `http://bit.ly/twitpopulation`, Jan 2012.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on Twitter. In *ACM WSDM*, pages 65–74, 2011.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *AAAI ICWSM*, May 2010.

[5] C. Clarke, G. Cormack, and E. Tudhope. Relevance ranking for one to three term queries. *Information Processing and Management*, 36:291–311, 2000.

[6] T. H. Haveliwala. Topic-sensitive pagerank. In *ACM WWW*, pages 517–526, 2002.

[7] N. Kallen. Twitter blog: Soon to Launch: Lists. `http://blog.twitter.com/2009/09/soon-to-launch-lists.html`, Sep 2009.

[8] C. Lee, H. Kwak, H. Park, and S. Moon. Finding influentials based on the temporal order of information adoption in Twitter. In *ACM WWW*, 2010.

[9] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. In *ECIR*, 2007.

[10] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *ACM WSDM*, pages 45–54, 2011.

[11] R. Pochampally and V. Varma. User context as a source of topic retrieval in Twitter. In *Workshop on Enriching Information Retrieval (with ACM SIGIR)*, Jul 2011.

[12] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *ACM WWW*, pages 113–114, 2011.

[13] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: a comparison of microblog search and web search. In *ACM WSDM*, pages 35–44, 2011.

[14] Rate Limiting | Twitter Developers. `https://dev.twitter.com/docs/rate-limiting`.

[15] Twitter: Who to Follow. `http://twitter.com/#!/who_to_follow`.

[16] Twitter Improves "Who To Follow" Results & Gains Advanced Search Page. `http://selnd.com/wtfdesc`.

[17] L. Rao, Twitter Seeing 90 Million Tweets Per Day, 25 Percent Contain Links, *TechCrunch*, 2010. `http://tinyurl.com/27x5cay`.

[18] M. J. Welch, U. Schonfeld, D. He, and J. Cho. Topical semantics of Twitter links. In *ACM WSDM*, 2011.

[19] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *ACM WSDM*, 2010.

[20] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on Twitter. In *ACM WWW*, pages 705–714, 2011.

[21] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: Structure and algorithms. In *ACM WWW*, 2007.