Evaluating Real-Time Search Over Tweets

Ian Soboroff[1], **Dean McCullough**[1], **Jimmy Lin**[2], **Craig Macdonald**[3], **Iadh Ounis**[3], **Richard McCreadie**[3]
[1]NIST, {ian.soboroff,dean.mccullough}@nist.gov
[2]University of Maryland, jimmylin@umd.edu
[3]University of Glasgow, {craigm,ounis,richardm}@dcs.gla.ac.uk

Abstract

Twitter[1] offers a phenomenal platform for the social sharing of information. We describe new resources that have been created in the context of the Text Retrieval Conference (TREC) to support the academic study of Twitter as a real-time information source. We formalize an information seeking task—real-time search—and offer a methodology for measuring system effectiveness. At the TREC 2011 Microblog Track, 58 research groups participated in the first ever evaluation of this task. We present data from the effort to illustrate and support our methodology.

Introduction

Twitter is a communications platform on which users can send short, 140-character messages, called "tweets", to their "followers". Conversely, users can receive tweets from people they follow via a number of mechanisms, including web clients, mobile clients, and SMS. As of Fall 2011, Twitter has over 100 million active users worldwide, who collectively post over 250 million tweets per day.

Increasingly, Twitter is being used to share critical information with substantive, meaningful impact on society: for example, as a tool for grassroots organization and mobilization in the "Arab Spring" democracy movement since December 2010; or as a post-disaster coordination mechanism following the Japanese earthquake and tsunami in March 2011. We need support for searching, filtering, distilling, and summarizing these large volumes of messages.

Twitter itself supports real-time search (Busch et al. 2012), which allows users to see what others are tweeting about *right now*. The service has just begun to address the issue of tweet relevance,[2] but there is much room for improvement. This paper is concerned with a small but critical variation in task: find the *essential tweets* relating to the search terms. Such a search should reveal not what people are saying right now, but what the searcher needs to know about that topic to get up to speed on the conversation.

[1]This article discusses data from the company Twitter solely in a scientific context and does not recommend, endorse or promote any commercial product or service.

[2]engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html

Consider for example the death of Apple co-founder and CEO Steve Jobs on the afternoon of October 5th, 2011. Searching "Steve Jobs" or clicking on a related trend would confirm his death, with scores of new tweets appearing every second that repeat this fact. Twitter's current relevance algorithms do not appear adequate to "cut through the noise" if the reader wanted to know precisely what had happened or further details. Imagine a search for "Steve Jobs" that would retrieve the most relevant tweets about the topic, which might be top retweeted URLs containing details about his last days or reactions from colleagues and family of Jobs. This is the kind of search that this paper explores.

The Text REtrieval Conference (TREC)[3] is a workshop series that aims to improve the state-of-the-art in information access effectiveness through building sharable *test collections* and collaboratively developing measurement practices. A test collection consists of a set of things to be searched (usually called "documents"), a set of defined information needs, which might include actual user queries, and an indication of which documents should be retrieved in response to each information need. In information retrieval experiments, test collections are used to measure the effectiveness of a search system in isolation from a production environment (Voorhees and Harman 2005). This laboratory setting allows different experimental conditions to be reasonably compared on shared data. Results that are determined in a test collection experiment can then be tested in a user study setting (Hersh et al. 2000).

The Microblog Track is a focus area within TREC to examine search issues in Twitter that builds upon previous experiences from the TREC Blog Track 2006–2010 (Macdonald et al. 2010). For the first year of the Microblog Track in 2011, we gathered a sample of tweets and determined a way to share those tweets among an arbitrary community of researchers within the bounds of the Twitter terms of service. We designed a real-time search task, which was operationalized in 49 topics, each created by a real user to represent an actual information need. At TREC 2011, 58 research teams from around the world attempted the task, by submitting to TREC what they considered to be the top tweets that should be returned for each topic. This paper describes our efforts, which complements the analysis in (Ounis et al. 2012).

[3]trec.nist.gov

The Corpus

Tweets are available from one of two sources, a REST-based API and a streaming API. The streaming API—colloquially known as "the firehose"—offers near real-time access to user tweets as they are created. Unfortunately, this service is not generally available to the research community, and furthermore, data from the firehose is ephemeral, in that the client accessing the API must manage, organize, and store the received data itself (it is, for example, impossible to "rewind" the firehose to access tweets created from last week). The REST API, in contrast, provides access to all available tweets as well as common Twitter functionality including posting new tweets, retweeting, following a user, and searching. Twitter's REST API is generally available to the public, although by default it is rate limited.[4] This restriction makes it impractical to gather large numbers of tweets for offline processing. Historically, Twitter has lifted the API request limit for some clients, but this capability is no longer offered.

Beyond the technical restrictions, Twitter's terms of service forbids third parties from data redistribution, which means that a researcher who has gathered a tweet dataset cannot legally share it. As a result, previous studies on Twitter data have used *ad hoc*, one-off collections from individual research groups. Each study typically adopts a custom methodology for gathering tweets, and the collections examined are usually mutually exclusive, making it difficult to compare findings across different studies and generalize results. This, naturally, is problematic for academic research, which is built on reproducibility of research findings.

An explicit goal of the TREC Microblog Track is to develop *reusable* collections of tweets, so that multiple groups can work on the *same* data, leading to comparability of results as well as replicability of findings. The solution to the constraints above is to distribute not the tweets themselves but rather pairs of (username, tweet id) tuples and associated software for *reconstructing* the tweets. The software is an asynchronous HTTP fetcher that downloads each tweet individually from Twitter. Researchers with access to the REST API without rate limit restriction can fetch the tweets in JavaScript Object Notation (JSON) format. For researchers without this access, the fetcher crawls raw HTML pages from the twitter.com site and reconstructs the tweets in JSON format via scraping. Some fields present in the JSON retrieved from the REST API are not available from the HTML, but we consider this limitation an acceptable trade-off for wider accessibility.

The collection consists of an approximately 1% sample (after *some* spam removal) of tweets from January 23, 2011 to February 7, 2011 (inclusive). Major events that took place within this time frame include the massive democracy demonstrations in Egypt as well as the Super Bowl in the United States. To ensure a representative multi-lingual tweet retrieval environment, no language filtering was performed. The ids for the resulting 16 million tweets are available under a usage agreement from NIST.[5]

This Tweets2011 corpus is one of the major contributions of our work, beyond the real-time search task that we explored in the context of TREC—for the first time, it is now possible for researchers to work on common data, thus bringing all the benefits associated therewith.

[4] dev.twitter.com/docs/rate-limiting

[5] trec.nist.gov/data/tweets/

```
<top>
<num> Number: MB01 </num>
<title> Wael Ghonim </title>
<querytime> 25th February 2011
        04:00:00 +0000 </querytime>
<querytweettime> 3857291841983981
       </querytweettime>
</top>
```

Figure 1: Example topic for the TREC Microblog Track.

The Real-Time Task

Recency or freshness is key in various search tasks. Mishne and de Rijke (2006) found that blog search queries for named entities were tied to real-world events; these *context queries* aimed to identify the news context around the occurrences of the target of the query. Often, social media search engines rank posts in reverse chronological ordering instead of in order of predicted relevance (Mishne 2006; Thelwall and Hasler 2007).

Following the current organization of Twitter search, we believe that results should be ordered reverse chronologically. That is, the system should answer a query by providing a list of relevant tweets ordered from newest to oldest, starting from the time the query was issued.

While returning all recent matching tweets is one way to address the user's search need, it risks drowning the user in tweets that are only marginally relevant. For example, issuing a search on a trending topic tends to retrieve tweets mentioning that the topic is trending, which is probably not what the user wants to know. We propose that the user wants to get up to speed on the topic, and thus wishes to retrieve highly relevant tweets that will provide context, and perhaps link to important outside resources.

Topic Creation

In TREC, a *topic* is an articulation of a user's search need, including sufficient information for the topic's creator to later judge whether the retrieved tweets are relevant to the topic or not. In the experiment, systems are presented with some subset of the topic and create a query based on that, possibly with human assistance.

Forty-nine topics were generated by a group of analysts at NIST. For each topic, the analyst created a title, a description defining the information that was being requested, a "trigger" tweet that indicates the time of the search and that might cause someone to be interested in this topic, and an estimate of the number of relevant tweets in the collection preceding the trigger tweet in time. Analysts tried to identify topics with a small number of relevant tweets, as topics with hundreds or thousands of relevant tweets can unbalance the test collection and be a source of assessment error.

TREC participants were only provided the topic title (as a query) and trigger tweet. The remaining information was used to guide the selection of topics to use as part of the evaluation set, and to remind the analysts when evaluating the

results what the question was. An example topic is shown in Figure 1. Note that the time that the search is issued is expressed as the trigger tweet itself.

During topic development, the analysts searched the Tweets2011 collection using a tool based on Apache Lucene. The tool accepted queries in Lucene's query format and would return both matching tweets as well as an indication of the density of matching tweets on each day of the collection's two-week epoch. As an example, for topic 3, "Haiti Aristide Return", the topic question was "What is the news on Aristide possibly returning to Haiti, and what is the reaction to this?" The trigger tweet was "Haiti concede passaporte a Aristide." Note that in this case the trigger tweet contained relevant information, but this was not a requirement. The analyst performed the following searches:

1. Haiti (finding 1136 matching tweets)
2. +haiti +Aristi* (=> 55)
3. +haiti +return (=> 29)
4. +haiti +return -arist* (=> 6, all of these discussing Duvalier's return, not Aristide's)

In the second query's results, 15 of the first 25 were found to be relevant, and thus it was estimated that about 50 tweets were relevant; 38 were found by participants' systems.

Topic Evaluation

At TREC 2011, 58 groups participated in the task, submitting 184 "runs" or rankings of the top tweets (max 1000) for each topic. For each topic, the top 30 tweets (by score, rank, and identifier) from each run were gathered and the resulting list was de-duped and retweets were removed. In TREC, this list is called the "pool" and is presented to the analyst who decides which tweets are relevant, highly relevant (adding interesting information), or not relevant. Novelty was not considered. By pooling results from multiple diverse retrieval systems, a high quality test collection that identifies many relevant items can be created (Voorhees and Harman 2005).

To improve judgment consistency, the resulting pool for each topic was ordered to have similar tweets near each other. This reordering was done by first generating a similarity score between each pair of tweets. The score used was the number of matching character 6-grams between them, after removing non-letters and converting to lower case. An arbitrary tweet was selected as the starting point, and a Hamilton path was generated by taking as the next tweet the tweet that has not yet been selected and that was closest (scored highest) to the previous two tweets. A Hamilton path is a path in a graph that visits each node exactly once, but does not return to the original tweet. In our case, the graph is completely connected, and the approach is a greedy nearest neighbor procedure. Thus, this path results in a list of all the tweets, with adjacent tweets being as similar as possible.

Each tweet was judged with respect to the topic title and description formulated earlier, by the analyst who developed the original topic (to the extent possible). Analysts could choose from four labels for each tweet in the pool for a given query (with number of judgments of each type): Highly Relevant (561), Relevant (2404), Not Relevant (47243), Spam (116). The distinction between "highly relevant" and "relevant" was left to the analyst to decide, on a topic-by-topic basis. As a result, the distinction between these two judgment levels is not consistent across topics. Labeling of spam tweets was guided based on a small number of examples, primarily tweets containing many hashtags; these judgments should not be taken as definitive or complete. Non-English language tweets were judged as "not relevant".

In judging the relevance of a tweet that contains a link to other web content, that content is considered part of the tweet. However, subsequent links from the landing page were not considered or examined. This results in some degradation of the corpus, as web links have become invalid in the intervening time. The evaluation was made on what was available; if the link was no longer valid, then the web link was not considered in the evaluation.

Retweets were not considered to be relevant. In our construction of the real-time search task, where the searcher wishes to get up-to-date on a subject being discussed, retweets might signal relevance but are not considered to be relevant themselves. Partial retweets, that is, tweets which contain a comment followed by a retweet, were evaluated only on the portion preceding the retweet (and thus were almost always considered not relevant).

On a closer inspection of the pooled tweets from each run, after de-duping, there were a total of 50324 tweets across all 49 topics, of which 2965 were marked as relevant or highly relevant (5.9%). The number of relevant or highly relevant tweets for each topic found ranged from 1 to 177, with a median of 42. Of the retrieved tweets, 9940 were retweets, and another 1858 were partial retweets (indicated by the string "RT @" not at the beginning of the tweet). The number of relevant partial retweets found was 47 or 2.5%. Overall, this suggests that by reducing the number of retweets retrieved, participanting groups could have markedly enhanced their retrieval effectiveness.

Measuring Effectiveness

For the first iteration of the Microblog Track, our goal was to provide a simple effectiveness metric as the starting point for discussion and refinement in subsequent evaluations. Thus, we ordered participants' rankings in descending order of tweet identifier (as a close proxy to reverse chronological ordering) and computed precision at 30 tweets. Precision is defined as the fraction of a set which is relevant; for this task, we compared the systems' precision scores considering alternatively tweets that were minimally topical ("relevant") and tweets that met the "interestingness" threshold ("highly relevant"). A complete analysis of participant scores appears in (Ounis et al. 2012).

Topic Categories

To analyze the 49 topics in the collection, we manually categorized each topic in three ways:

1. **News Categories:** Each topic was classified into 11 standard categories used by news providers to distinguish news article types.

News Category	# Topics	All-Rel		High-Rel	
		Mean MAP	Mean P@30	Mean MAP	Mean P@30
Arts	3	0.1770	0.1991	0.2520	0.0622
Business	3	0.1521	**0.4948**	0.0715	**0.1643**
Entertainment	8	**0.2116**	0.2477	**0.2647**	0.0984
Food	1	0.0082	0.0313	N/A	N/A
Health	1	0.0832	0.0248	0.0832	0.0248
Politics	6	0.0937	0.3054	0.0609	0.1083
Shopping	1	0.1423	0.1706	0.1432	0.1189
Sport	5	0.1869	0.1646	0.2063	0.0376
Travel	2	0.0709	0.2859	0.0295	0.0443
U.S.	8	0.1394	0.2000	0.0316	0.0224
World	11	0.1476	0.2667	0.1222	0.1005

Table 1: News categories in the Microblog 2011 task topics and the mean system effectiveness for each category.

Topic Geography	# Topics	All-Rel		High-Rel	
		Mean MAP	Mean P@30	Mean MAP	Mean P@30
International	21	**0.1874**	0.2529	**0.2143**	**0.0896**
National	22	0.1569	**0.2669**	0.1271	0.0844
Neutral	6	0.0905	0.0960	0.1815	0.0594

Table 2: Topic Geographical Interest categories in the Microblog 2011 task topics and the mean system effectiveness for each category.

2. **Geographical Interest:** Topics were classified as either being of interest to a national (U.S.) audience, international audience or as geographically neutral.
3. **Topic Target:** Each topic was classified with regard to the target of the information need, either an named entity (single/multi-term), an ambiguous entity (e.g., "Kate and William"), an acronym (e.g., "NIST"), a location or no obvious target (None).

Tables 1, 2 and 3 report the mean effectiveness across all submitted runs for topics belonging to the different News, Geographical Interest and Topic Target categories, respectively. Effectiveness is reported in terms of Mean Average Precision (MAP), which measures a system's ability to effectively rank all relevant tweets for a topic, and Precision at rank 30. Scores are shown both for the condition where any relevant tweet counts towards the measure (denoted All-Rel), and where only highly-relevant tweets are counted (High-Rel).

From Table 1 (News Categories), we observe that participating systems were most effective on topics pertaining to the Business and Entertainment news categories. This is intuitive, as these categories hold high interest stories (and hence are high impact) to Twitter users, resulting in significant coverage.

We observe from Table 2 that participating systems did not favor national or international stories under P@30. However, effectiveness is distinguishable under the more recall-oriented MAP measure, especially when only highly relevant tweets are counted as correct. Moreover, it appears that topics with no clear geography of interest, e.g., topic 47 "organic farming requirements", are more difficult.

Table 3 reports the mean effectiveness for topics containing specific types of target terms. We observe the following: First, topics with no target (None) appear to be difficult, with low effectiveness achieved by runs. Second, systems performed better on topics containing named entities (all types), especially under the high-rel assessments. Third, locations within topics lead to higher precision but not higher overall MAP. Fourth, participating systems performed better on

Topic Target	# Topics	All-Rel		High-Rel	
		Mean MAP	Mean P@30	Mean MAP	Mean P@30
None (Ambiguous)	7	0.0547	0.1977	0.0387	0.0665
Acronym	4	0.1680	0.1897	0.0294	0.0708
Location	10	0.1037	**0.2660**	0.0723	**0.1116**
Single Term Named Entity	17	0.1857	0.2589	0.1805	0.0790
Multi-Term Named Entity	9	0.1352	0.2586	0.1589	0.0591
Ambiguous Named Entity	2	**0.4052**	0.1034	**0.4642**	0.0825

Table 3: Topic Target categories in the Microblog 2011 task topics and the mean system effectiveness for each category.

topics with single term entities than on multi-term entities, indicating that some systems were not leveraging n-grams effectively. Last, ambiguous entities in topics resulted in particularly high MAP but not overall precision.

Conclusions

This paper describes the methodology and provides a brief analysis of the results from the first reusable test collection for search on Twitter, the real-time search task of the inaugural TREC Microblog Track in 2011. This task was investigated by 58 research groups from many countries, indicating the wide research interest in reusable Twitter evaluation data—this number represents a track record in the 20 years of TREC. Our methodology enables the evaluation of real-time search systems where rankings must be most recent items first. In particular, we detail how topics that are appropriate for the timeframe of the developed Tweets11 corpus can be created, and how near-duplicate detection can be used to reduce the assessor error. Our analysis illustrates the difficulties faced by participants in building effective approaches for a real-time search task. Indeed, we show that topics that do not mention name entities are among the most difficult. The Microblog Track will run again in TREC 2012, such that advances in both approaches for real-time search and the corresponding evaluation methodology can be achieved.

Reusable datasets remain a challenge for the social media research community, and we hope that the Tweets2011 collection might mark a shift towards collaboratively-designed tasks and measurements on open data. This in turn would promote corroboration of experimental results, a shared vision of future work directions, and more significant results for the research community as a whole.

References

Busch, M.; Gade, K.; Larson, B.; Lok, P.; Luckenbill, S.; and Lin, J. 2012. Earlybird: Real-time search at Twitter. In *ICDE*.

Hersh, W.; Turpin, A.; Price, S.; Chan, B.; Kramer, D.; Sacherek, L.; and Olson, D. 2000. Do batch and user evaluations give the same results? In *SIGIR*.

Macdonald, C.; Santos, R.; Ounis, I.; and Soboroff, I. 2010. Blog track research at TREC. *SIGIR Forum* 44(1):57–74.

Mishne, G., and de Rijke, M. 2006. A study of blog search. In *ECIR*.

Mishne, G. 2006. Information access challenges in the blogspace. In *Intl. Workshop on Intelligent Information Access*.

Ounis, I.; Macdonald, C.; Lin, J.; and Soboroff, I. 2012. Overview of the TREC 2011 Microblog Track. In *TREC*.

Thelwall, M., and Hasler, L. 2007. Blog search engines. *Online Information Review* 31(4):467–479.

Voorhees, E. M., and Harman, D. K. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.