Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media

A Study of "Churn" in Tweets and Real-Time Search Queries

Jimmy Lin and **Gilad Mishne**
Twitter, Inc.
@lintool @gilad

Abstract

The real-time nature of Twitter means that term distributions in tweets and in search queries change rapidly: the most frequent terms in one hour may look very different from those in the next. Informally, we call this phenomenon "churn". Our interest in analyzing churn stems from the perspective of real-time search. How do we "correctly" compute term statistics, considering that the underlying distributions change rapidly? In this paper, we present an analysis of tweet and query churn on Twitter, as a first step to answering this question. Analyses reveal interesting insights on the temporal dynamics of term distributions on Twitter and hold implications for the design of search systems.

Introduction

Twitter is a communications platform through which millions of users around the world send short, 140-character tweets to their followers. Particularly salient is the real-time nature of these global conversations, which rapidly evolve to reflect breaking events such as major earthquakes (e.g., Japan, March 2011) and deaths of prominent figures (e.g., Steve Jobs, October 2011). This paper analyzes the temporal dynamics of tweets and real-time search queries. We focus specifically on the notion of "churn", informally characterized as the process by which terms and queries become prevalent and then "drop out of the limelight". To our knowledge, this is the first large-scale study of this phenomenon.

We are interested in churn primarily from the perspective of real-time search. A defining property of search in this context is the speed at which relevance signals change, and thus a proper understanding and treatment of this phenomenon is instrumental to search and derived tasks (event tracking, query spelling correction, and so on). This paper falls short of delivering solutions, but presents a characterization of the phenomenon. We hope that these observations will be useful to the community.

Methods

Let us begin with definitions of metrics. We define a very simple measure of churn at rank r as the fraction of terms in the top r terms (as ordered by frequency) at time interval t_i that are no longer in the top r at time t_j. From this definition, churn at rank r of 0 indicates the top r terms are exactly the same (but may be in different relative order), whereas at the opposite end of the spectrum churn at rank r of 1 indicates that none of the top r terms are in common between the two time periods. In our analyses, we consider different intervals (daily and hourly) and examine interval-over-interval changes (t_1 vs. t_2, t_2 vs. t_3, t_3 vs. t_4, etc.).

This definition of churn does not capture changes in rank within r, or the actual term frequencies. To better characterize differences between term distributions represented by successive time intervals, we use Kullback-Leibler (KL) divergence. Since KL divergence is not a symmetric measure, to be precise, we compute $D_{KL}(S_{t+1}||S_t)$, where S_t and S_{t+1} are the term distributions at time interval t and $t+1$, respectively. Following most information retrieval applications, we use base 2 for the log. To prevent the problem of zero probabilities (e.g., from out-of-vocabulary terms), we smooth the maximum likelihood estimate term probabilities using Bayesian smoothing with Dirichlet priors (using the average of the two distributions as the background model). The smoothing hyperparameter μ is arbitrarily set to 10,000.

Finally, to quantify the impact of previously unseen terms, we compute an out-of-vocabulary (OOV) rate, also defined in terms of a rank r. An OOV rate at r is the fraction of terms in the top r (sorted by frequency) from time interval t_i that is not observed in the top r during time interval t_j. An OOV rate of 0 means that all top r terms have been previously observed (hence, collection statistics exist for them). A non-zero OOV rate means that a retrieval engine must explicitly handle query terms that may not exist in the collection (smoothing, backoff, defaults, etc.); otherwise, the retrieval model may produce non-sensical results.

Our analyses span the entire month of October 2011, both at the daily and hourly level (all times are provided in UTC). We consider all tweets created during that time, as well as all search queries submitted to the twitter.com site. Since October has 31 days, this corresponds to 30 different data points for the day-over-day analysis and 743 data points for the hour-over-hour analysis. Unfortunately, we are unable to provide exact statistics on the size of our complete dataset, except to note that according to publicly available figures (as of Fall 2011), Twitter users create approximately a quarter

of a billion tweets a day, and Twitter search serves over two billion queries a day (although this includes API requests).

To highlight topics gathering attention in tweets, Twitter introduced Trending Topics: algorithmically-identified emerging trends and topics of discussion (both worldwide and geographically-focused). The algorithm is based on term statistics: evidence for a term's recent prominence is continuously collected and compared with longer-term statistics about the term. Terms (including hashtags and phrases) that are significantly more prevalent in recent data, compared with past data, are marked as trending topics and surfaced. Note that, critically, phrases or hashtags become trending topics primarily as a result of velocity (i.e., the rate of change in prevalence), *not* sheer volume. Trending topics are featured prominently both on the twitter.com site and on Twitter clients, and a click on a trend leads to a Twitter search for tweets containing the trend. This results in significantly elevated query volumes, and these terms and phrases often become top-searched queries. Since this affects our query churn observations, we repeat our analyses twice: once using all queries, and once using all queries except those issued by clicking on a trend.

To provide some context and to highlight differences in churn between a real-time search engine and a general-purpose one, our results should be compared to similar numbers obtained from web search logs (from the same time period). Unfortunately, large-scale collections of web search queries are generally not available for research purposes, with the exception of the AOL query set (Pass, Chowdhury, and Torgeson 2006). As a point of comparison, this corpus is far from ideal: it is relatively old (2006); a sample of the full search stream; and drawn from a search engine that no longer had dominant market position at the time the data was collected. For completeness, we report churn and OOV figures, but alert the reader to the caveats above. We omit the hourly analyses from this source as the query frequencies are too low for meaningful analysis (the top hourly terms are observed only a handful of times).

Results

The day-over-day analysis in terms of KL divergence is shown in the top graph of Figure 1. There are three subgraphs: results over tweets (top), queries (middle), and query unigrams (bottom). The difference between the last two is worth explaining: for analysis in terms of queries, we consider the entire query string (which might consist of multiple terms) as a distinct event. This would, for example, consider "steve jobs" and "jobs" distinct events. For analysis in terms of query unigrams, all queries are tokenized into individual terms, and we consider the multinomial distribution over the term space. Both analyses are useful: when building a query model or extracting query-level features for learning to rank, estimates over the event space of queries would yield more signal, although due to sparsity, one would typically need to back off to unigram statistics. For both the middle and bottom subgraphs, we further break down analysis in terms of all queries (thin red line) and with trends discarded (thick blue line). We make a few observations: First, there does not appear to be cyclic patterns in day-over-day churn (e.g., day of week effects). Second, eliminating trend queries reduces KL divergence, i.e., successive distributions appear more "similar"—this makes complete sense given the nature of trending topics.

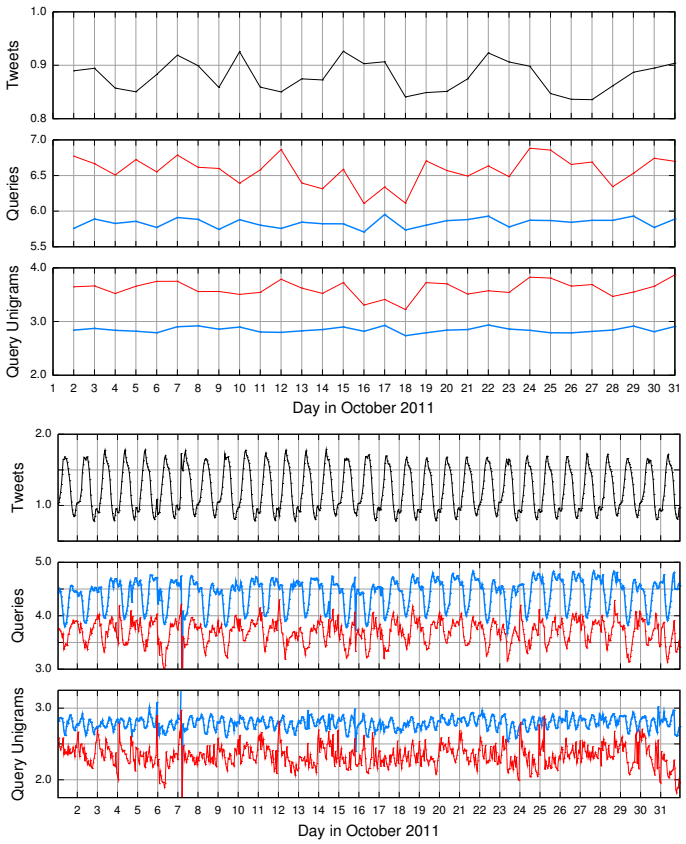

Figure 1: KL divergence: day-over-day (top graph) and hour-over-hour (bottom graph). Both graphs are similarly organized: tweets (1st subgraph), queries (2nd subgraph), and query unigrams (3rd subgraph); analysis without trends (thick blue line) and including trends (thin red line).

The hour-over-hour analysis in term of KL divergence is shown in the bottom graph of Figure 1. For tweets and queries, but not query unigrams, we observe strong daily cyclic affects. This is driven by a combination of the rhythm of users' daily activities and the international nature of Twitter: for example, as users in the United States go to bed and users in Japan wake up, the composition of tweets and queries naturally changes, resulting in churn. Interestingly, we observe that removing trends actually *increases* churn, i.e., we observe higher KL divergence values. This suggests that the typical lifespan of a trending topic is longer than an hour, i.e., trending topics churn "naturally" at a rate that is slower than hourly, such that removing those events from the distribution increases overall churn. The time that a particular topic trends is a function of many factors: for news, factors include significance of the news event and interactions with competing stories; for internet memes, the lifespan of a hashtag is often idiosyncratic. Although not in the Twitter context, see (Leskovec, Backstrom, and Kleinberg 2009) for a quantitative analysis of this "news cycle".

	Churn Rate				OOV Rate			
	10	100	1000	10000	10	100	1000	10000
Tweets	0.0167	0.0233	0.0407	0.0682	0.0000	0.0000	0.0008	0.0012
Queries	0.8067	0.8180	0.6413	0.3199	0.4500	0.4073	0.2678	0.0702
Queries (−T)	0.4433	0.3807	0.3166	0.2722	0.0400	0.0450	0.0317	0.0313
Q. Unigrams	0.8067	0.6937	0.4105	0.1754	0.1633	0.0960	0.0732	0.0297
Q. Unigrams (−T)	0.2500	0.1360	0.1254	0.1319	0.0100	0.0043	0.0051	0.0070
Web Queries	0.1107	0.2139	0.3670	0.7584	0.0000	0.0000	0.0290	0.5402
Web Q. Unigrams	0.0410	0.1608	0.1740	0.3401	0.0000	0.0000	0.0570	0.1641

Table 1: Day-over-day analysis, showing query churn (left half of the table) and OOV rate (right half of the table) at various ranks r for: tweets, all queries (± trends), all query unigrams (± trends), and web queries from the AOL dataset.

	Churn Rate				OOV Rate			
	10	100	1000	10000	10	100	1000	10000
Tweets	0.0349	0.0349	0.0653	0.1004	0.0000	0.0000	0.0000	0.0015
Queries	0.3055	0.3094	0.2880	0.2880	0.0424	0.0967	0.0452	0.2479
Queries (−T)	0.2899	0.2768	0.3336	0.5281	0.0096	0.0158	0.0329	0.2599
Q. Unigrams	0.3257	0.2930	0.1687	0.3059	0.0073	0.0302	0.0137	0.0598
Q. Unigrams (−T)	0.1783	0.1639	0.1657	0.3144	0.0027	0.0019	0.0032	0.0622

Table 2: Hour-over-hour analysis, showing query churn (left half of the table) and OOV rate (right half of the table) at various ranks r for: tweets, all queries (± trends), all query unigrams (± trends).

Table 1 presents churn rates and OOV rates at rank $r = \{10, 100, 1000, 10000\}$ for the day-over-day analysis, averaged across the entire month, for all five experimental conditions: tweets, queries, and query unigrams (± trends for the last two). Table 2 shows similar results for the hour-over-hour analysis.

From the search perspective, the OOV rates are interesting, in that they highlight a challenge that real-time search engines must contend with. Take, for example, the day-over-day query unigram OOV rate at rank 1000: results tell us that 7.32% of query terms were not observed in the previous day. This means that for a non-trivial fraction of query unigrams, we have no query-level features: query frequency, clickthrough data to learn from, etc. This is the result at rank 1000, which represents queries pretty close to the head of the distribution. Of course, this particular analysis includes trends (and removing trends reduces the OOV rate substantially), but trend queries remain an important class of queries for which we would like to return high quality results.

Zooming In

In the context of term churn, rapidly-unfolding events such as natural disasters or political unrest are of particular interest. In such scenarios term frequencies may change significantly over short periods of time as the discussion evolves. Our next analysis examines one such event, the death of Steve Jobs, the co-founder and CEO of Apple, in the afternoon hours of October 5th (around midnight UTC).

Figure 2 shows the KL divergence at 5-minute intervals over a period of 12 hours surrounding the event (thick blue line), contrasting it with the 5-minute KL divergence values over the same hours in the previous day (thin red line). Note the sharp drop in divergence as the real-time query stream focuses on the event. A few hours later, divergence converges to a pattern close to that observed in the previous day, although actual values are around 10% less, as significant portions of the query stream continue to discuss the event.

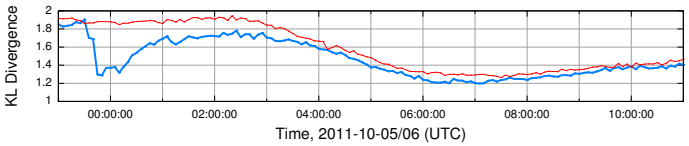

Figure 2: KL divergence of the query stream, in intervals of 5 minutes, over a 12 hour period during a major event (in blue, thick line); the overlay (red, thin) shows the 5-minute KL divergence during the same hours in the preceding day, for reference.

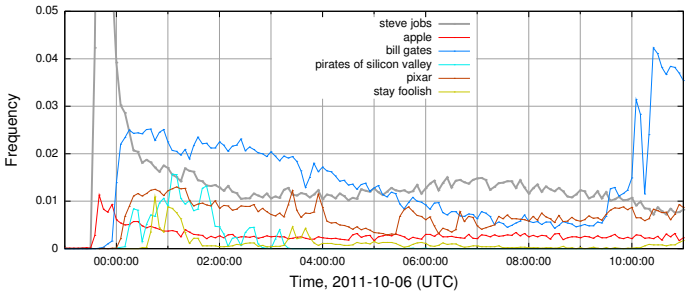

Figure 3: Frequencies of queries related to Steve Jobs' death over a 12 hour period in 5-minute intervals, normalized to the total number of queries in the interval. At its peak, the query "steve jobs" reaches 0.15 (15% of the query stream); for readability of the other query frequencies, the scale is not stretched to include this point.

However, even within a particular event, churn of individual queries and terms vary. Figure 3 shows the frequency of several queries related to the event over the same time period, in 5-minute intervals. Note that the patterns do not necessarily correlate in timespan or shape, displaying a range of exponential, linear, and irregular decays. It appears that a simple approach to account for changes in term frequency over time, such as applying a decay function over the frequency, may not be sufficient. Additionally, it is clear that, at least during major events, sub-hour updates to various col-

lection and query term statistics are essential.

Finally, we manually examined the terms responsible for high churn rates in tweets and in the query stream. One interesting observation is that term churn in tweets appears more cyclic and predictable than that in the query stream. While churn levels are high in both, in the case of the query stream this is largely driven by news events, whereas in tweets this is driven by a combination of shifting cyclical interests (e.g., weekends) and news. This suggests that while both tweets and queries experience large amounts of churn, they are *qualitatively* different. For example, terms like "weekend" and "party" in tweets have frequencies that rise and fall predictably, whereas variations in query frequencies appear to be far less predictable.

Related Work

In the domain of temporal information retrieval, a large body of work was driven by TDT—the Topic Detection and Tracking initiative (Allan 2002). Studies demonstrate that document retrieval benefits from incorporating temporal aspects of the collection into the ranking, e.g., via a language model (Li and Croft 2003), by utilizing additional statistics about term frequencies over time (Efron 2010), or by taking into account the temporal distribution of results (Jones and Diaz 2007). A body of recent work focuses on the temporal dynamics of Twitter *content* (rather than search queries). For example, Petrović et al. (2010) apply a TDT task—first story detection—to Twitter data, while Wu et al. (2011) develop predictors for term churn in tweets.

In the context of web search, Beitzel et al. (2004) track the relative popularity of queries throughout the day, as well as the hourly overlap between queries. While exact figures vary by query category, they observe relatively high correlations between the queries of a given hour and the hour following it, with few exceptions related to breaking news. A later study by Kulkarni et al. (2011) groups queries by the shape of their popularity over time, showing significant differences over time. However, they focus on a small set of manually-selected queries rather than examining properties of the full query stream.

Of particular interest to us is the work of Teevan et al. (2011), who analyze several aspects of Twitter queries and compare them to the web query stream. Interestingly, they observe a *lower* churn rate on Twitter than on the web. This is counter-intuitive, and may be attributed to the relatively limited amount of data analyzed (our collection is several orders of magnitude larger, as well as annotated for the presence of trends, cleaned of spam, and so on).

Also related to our study is the extensive analysis of modifications to web documents over time, presented in (Adar et al. 2009). The types of change patterns we observe appear very different from the patterns the authors identify on web pages (e.g., "hockey stick" curves), which naturally makes sense given the context.

Conclusions

This paper examines changes to term distributions on Twitter over time, focusing on the query stream and the implications for ranking in a real-time search system. Our observations can be summarized as follows:

- **Churn.** Term distributions change rapidly—significantly faster than in web search for the head, even after discounting trending terms. For the tail, churn drops quickly, and appears to be lower than that observed in web queries.
- **Unobserved terms.** Similarly, OOV rates are higher for top Twitter queries, but lower at the tail of the distribution. This translates to rapid changes in the top user interests, but relative stability in the topics for which users seek real-time results.
- **Update frequency.** Although query churn is consistently high, during major events it can further increase dramatically, as queries change minute by minute. In fact, to maintain accurate collection statistics requires frequent term count updates—in intervals of 5 minutes or less, according to our data.
- **Churn patterns.** The time period in which a query remains a top one varies, as does its decay pattern; naïve approaches such as fixed term frequency decays may not be able to correctly model frequency changes over time.
- **Predictability.** Anecdotal evidence suggests that some query churn may be predicted from past observations, providing a potential source for addressing this issue.

The growing importance of real-time search brings several challenges; this paper frames one such challenge, that of rapid changes to term distributions, particularly for queries. In follow-up work we plan to evaluate techniques for handling the volatility of the real-time search stream and the limited collection statistics that exist for new queries.

References

Adar, E.; Teevan, J.; Dumais, S.; and Elsas, J. 2009. The web changes everything: understanding the dynamics of web content. In *WSDM*.

Allan, J., ed. 2002. *Topic detection and tracking: event-based information organization*. Kluwer.

Beitzel, S.; Jensen, E.; Chowdhury, A.; Grossman, D.; and Frieder, O. 2004. Hourly analysis of a very large topically categorized web query log. In *SIGIR*.

Efron, M. 2010. Linear time series models for term weighting in information retrieval. *JASIST* 61:1299–1312.

Jones, R., and Diaz, F. 2007. Temporal profiles of queries. *TOIS*.

Kulkarni, A.; Teevan, J.; Svore, K.; and Dumais, S. 2011. Understanding temporal query dynamics. In *WSDM*.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD*.

Li, X., and Croft, W. 2003. Time-based language models. In *CIKM*.

Pass, G.; Chowdhury, A.; and Torgeson, C. 2006. A picture of search. In *InfoScale*.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story detection with application to Twitter. In *HLT*.

Teevan, J.; Ramage, D.; and Morris, M. R. 2011. #TwitterSearch: a comparison of microblog search and web search. In *WSDM*.

Wu, S.; Tan, C.; Kleinberg, J.; and Macy, M. 2011. Does bad news go away faster? In *ICWSM*.