Temporal Relevance Profiles for Tweet Search

Jimmy Lin[1] and Miles Efron[2]

[1] The iSchool, University of Maryland, College Park
[2] Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign

jimmylin@umd.edu, mefron@illinois.edu

ABSTRACT

When searching tweets, users may know something about the temporal characteristics of the information they're after. For example, based on external knowledge, a searcher might prefer more recent results or results within a particular time interval. However, most search applications do not allow the user to explicitly supply this information, and neither do most retrieval models have a mechanism to incorporate this additional evidence. In this paper, we introduce the notion of a temporal relevance profile, which a user explicitly includes alongside a keyword search query. We propose alternative representations of temporal relevance profiles and how existing retrieval models might take advantage of this data. Oracle experiments on microblog track data from TREC 2011 and 2012 empirically demonstrate that this approach has the potential to significantly increase the quality of retrieved results.

1. INTRODUCTION

Twitter has become an indispensable communications platform through which hundreds of millions of users around the world witness breaking news events and participate in the global conversation in real time, 140 characters at a time. To access relevant content, users often turn to search. Almost by definition, searching for tweets has an important temporal component.

The temporal distribution of relevant tweets for an information need is frequently non-uniform, and thus it is important for retrieval systems to model the temporal characteristics of the query, retrieved documents, and the collection as a whole. This is an insight shared by many researchers [6, 3, 5, 4, 2, 9], which provides a starting for our study.

In this paper, we explore the value of temporal signals for tweet search in the following progression:

- We begin with an empirical characterization of the temporal distribution of relevant documents for various information needs.
- We experimentally quantify the value of this temporal signal with an oracle that is given access to the (*post hoc*) distribution of relevant documents.
- We show that an obvious approach to exploiting this temporal signal—using the empirical distribution of retrieved documents—does not appear to be effective.
- We conclude with an alternative approach that reframes the challenge of exploiting temporal signals as an interactive retrieval problem.

In this paper we focus on microblog search because of the important role of temporal signals. However, our proposed methods would likely generalize to other domains and retrieval tasks with a strong temporal component (e.g., news search), although we leave such explorations for future work.

2. TEMPORAL RELEVANCE

The context for our study is the recent microblog tracks at TREC [8, 12]. The 2011 and 2012 evaluations used the Tweets2011 corpus,[1] which consists of an approximately 1% sample (after some spam removal) of tweets from January 23, 2011 to February 7, 2011 (inclusive), totaling approximately 16 million tweets. 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. There are 49 topics for TREC 2011 and 60 topics for TREC 2012. Each topic consists of a query and an associated timestamp, which indicates when the query was issued. Using a standard pooling strategy, NIST assessors evaluated a total 114K tweets and assigned one of four judgments to each: spam, not relevant, relevant, and highly-relevant. For the purpose of our experiments, we considered both relevant and highly-relevant tweets "relevant".

We begin with simple visualizations that characterize the distribution of relevant documents for various TREC microblog topics in Figure 1. Each topic is associated with a query time, represented by the right edge of each graphic. The x-axis shows time *prior* to the query time, in days. Dots show tweets that were retrieved by participating teams and evaluated by assessors (i.e., the pools): green dots are relevant, red dots are highly relevant. The vertical position of the dots has no meaning; jitter is added only to prevent overlap. The underlying blue bars shows the distribution of relevant and highly-relevant tweets as a histogram. We see that for topic 29 "global warming and weather", relevant tweets are distributed relatively evenly from a temporal

[1] `http://twittertools.cc/`

TAIA'13 August 1, 2013, Dublin, Ireland.

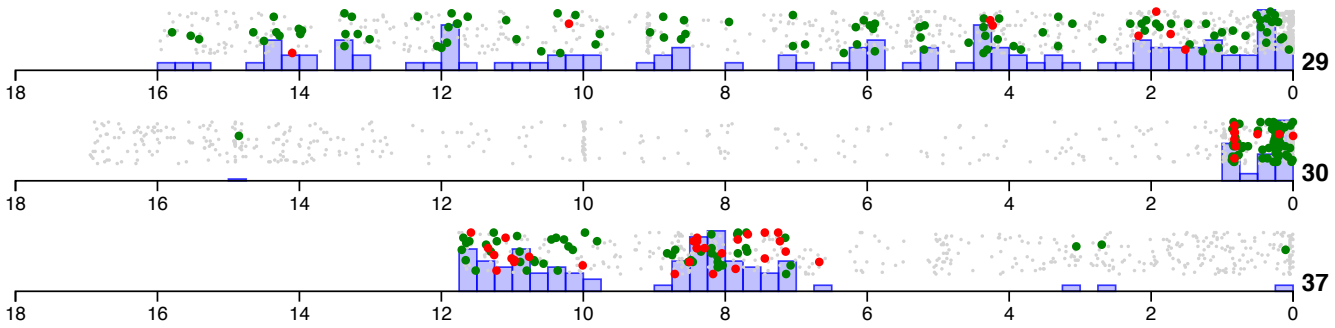

Figure 1: Visualizations illustrating the temporal distribution of retrieved documents and relevant documents for three topics from the TREC 2011 microblog track: topic 29, "global warming and weather", topic 30 "Keith Olbermann new job", and topic 37 "Giffords recovery". The timeline is measured in days, anchored by the query time on the right edge. Green dots represent relevant documents, red dots represent highly-relevant documents, and gray dots represent irrelevant documents. The bar graphs show bucketed distributions of the relevant and highly-relevant documents.

perspective; for topic 30 "Keith Olbermann new job", with one exception all relevant tweets are very close to the query time; for topic 37 "Giffords recovery", most relevant tweets are clustered in two temporal intervals that occur several days prior to the query time. Due to space limitations, only three topics are shown here, but visualizations for all topics from TREC 2011 and TREC 2012 are available online.[2] Nevertheless, these three timelines are fairly representative of the shapes of the distribution we see across all topics.

How might we take advantage of the temporal signal illustrated by these visualizations? Let us extend the basic language modeling approach [10, 1] (specifically, query-likelihood) to incorporate temporal signals using the framework of Dakka et al. [2] and building on the temporal ranking extension of Efron and Golovchinsky [4]. For ranking, let us consider two types of features for a given document (tweet): first, w_d, the lexical terms in the document and second, the timestamp t_d. We can then decompose the likelihood function as follows:

$$P(D|Q) = P(w_d, t_d|Q) \quad (1)$$
$$= P(t_d|w_d, Q)P(w_d|Q) \quad (2)$$

Making the simplifying assumption that the temporal relevance of D does not depend on the document's content, we can drop w from the joint probability in the above equation, giving us:

$$P(D|Q) \propto P(w_d|Q)P(t_d|Q) \quad (3)$$

which is identical to the standard query-likelihood model, but with the addition of the probability of observing a time t_d given the query Q. Intuitively, this can be understood as, for a given query Q, "where would I expect the relevant tweets to show up in time?" Formally, we denote this as an arbitrary function f_t to indicate that it is not necessarily a probability distribution:

$$P(D|Q) \propto P(w_d|Q)f_t(t_d) \quad (4)$$

To establish an effectiveness upper bound and to quantify the value of this temporal signal, we define an oracle condition. Suppose an oracle told us the timestamps of all the relevant tweets. If this information were available, we would simply use the empirical distribution to estimate f_t. Formally, we accomplish this using kernel density estimation, a non-parametric way to estimate the probability density function of a random variable.

Let $\{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample drawn from some distribution with an unknown density f. We are interested in estimating the shape of this function f. Its kernel density estimator is:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=0}^{n} K\left(\frac{x - x_i}{h}\right) \quad (5)$$

where $K(\cdot)$ is the kernel—a symmetric but not necessarily positive function that integrates to one—and $h > 0$ is a smoothing parameter called the bandwidth. If we choose a Gaussian kernel, as we do here, then as Silverman [11] has shown, the optimal bandwidth is:

$$h^* = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{-\frac{1}{5}} \quad (6)$$

where $\hat{\sigma}$ is the sample standard deviation. It is important to note that the choice of a kernel function is mainly a matter of convenience, carrying with it no implications of underlying parametric forms of the data. We select the Gaussian due to its wide use and its ready definition of an optimal bandwidth.

All experiments reported below were performed with Indri. The effectiveness of the oracle experiment is shown in Table 1. The baseline condition (QL) is standard query likelihood with Dirichlet prior ($\mu = 2500$), retrieving 1000 results per topic. For the oracle condition, we retrieved the top 5000 results using query-likelihood and then reranked the results using the kernel density estimation method described above. During collection preparation, we eliminated all retweets since they are by definition not relevant according to the assessment guidelines.[3]

As an additional reference condition, we report the effectiveness of a technique commonly known as recency priors.

[2] `https://github.com/lintool/trec-mb-vis/`

[3] Note that removal of retweets has a substantive impact on effectiveness—since retweets usually appear temporally close to the original tweet, unless we discarded retweets the oracle condition would also be promoting non-relevant documents.

Table 1: Effectiveness results on TREC microblog data. Statistical significance (i.e. $p < 0.05$ on a paired t-test) is shown with the following symbols: †(improve on QL baseline), ‡(improve on QL and recency prior baselines), × (decline with respect to recency prior and QL).

(a) TREC 2011

	MAP	Rprec	P30
QL	0.2980	0.3264	0.3673
Recency prior	0.3082†	0.3362	0.3796
Oracle	0.3612‡	0.4004‡	0.4041‡
Empirical density	0.2694×	0.2969×	0.3660×

(b) TREC 2012

	MAP	Rprec	P30
QL	0.1930	0.2525	0.3446
Recency prior	0.1969	0.2532	0.3429
Oracle	0.2260‡	0.2969‡	0.3751‡
Empirical density	0.1826×	0.2459×	0.3402×

In early work on integrating time into language modeling, Li and Croft [7] defined a prior distribution over documents that favors recent documents in the following form:

$$P(D) = \lambda e^{-\lambda t_D} \tag{7}$$

where λ is the rate parameter of an exponential distribution and t_D is the age of document D. Following Efron and Golovchinsky [4] we use $\lambda = 0.01$ and document ages measured as fractions of days before query time (cf. [9]).

Results show that while the recency prior approach helps for TREC 2011 topics, the improvements are not statistically significant for Rprec and P30. For topics from TREC 2012, the recency prior doesn't have any significant impact on effectiveness. Given the visualizations in Figure 1, this result is not surprising: although for some topics the relevant documents are indeed clustered right before the query time, this is not universally true. The effectiveness of the recency prior is dependent simply on the prevalence of topics that display a relevant tweet distribution similar to topic 30 in Figure 1.

The oracle condition is significantly better than both the query-likelihood baseline and the recency prior condition in all metrics. This is expected, and gives a picture of the value of the temporal signal contained in the distribution of relevant documents: +21% gain in MAP for TREC 2011 and +17% for TREC 2012.

3. EMPIRICAL DENSITY

We've shown that there is substantial value in knowing the distribution of relevant documents for a topic. The obvious next question is: can we somehow approximate the distribution without access to relevance judgments, just as we may conduct pseudo-relevance feedback when we lack relevance judgments? Instead of using oracle data, let us estimate the density f based on the top k results obtained from an initial run. This is similar to other approaches based on analysis of documents' empirical distribution (e.g., [2, 6]).

We attempted this empirical density approach with $k = 5000$, the results of which are shown in Table 1 (last row). Experiments show that this approach is ineffective, and significantly worse than the query-likelihood baseline. The results are relatively insensitive to the setting of k. This finding is somewhat surprising, as the density of retrieved documents has been found by other researchers to provide valuable relevance signals (see cites above).

Of course, this is only an initial attempt at computing correlates of the ground truth distribution of relevant documents—no doubt there are other statistical measures that are worth examining. However, we propose an alternative framing of the problem based on interactive retrieval, which we turn to next.

4. REFRAMING THE PROBLEM

Though fully-automated approaches to modeling time in search are obviously valuable, we argue that especially in tweet search, it is reasonable to obtain temporal information via user interactions instead of statistical estimation.

Let us consider a journalist who is searching an archive of tweets as part of a retrospective piece on the impact of social media on the course of the Egyptian revolution. Let's say she is particularly interested in activists using Twitter for "on the ground" reporting purposes—informing the world of live developments. In this case, the journalist has a concrete idea when the relevant tweets should occur: they are more likely to be posted when protesters were gathered in Tahrir Square or during some other organized event. Furthermore, she is likely to know when exactly these mass protests were—based on world knowledge, domain expertise, etc. It would be very desirable if the journalist could somehow impart this knowledge to the search system. For convenience, let us call this the "archive search" scenario.

Consider another scenario in which the journalist is investigating a sports scandal that has been brewing for the last several weeks. She has just now gotten news of a breaking development, and turns to searching tweets to find out more details: what exactly happened, reactions from fellow athletes, commentary from analysts, etc. Since this particular news story has been developing for several weeks, any keyword search involving the athlete's name might bring up results from many different points in time. It would be desirable if the journalist could specify that she is only interested in the most recent tweets. For convenience, let us call this the "recency search" scenario.

The commonality between these two scenarios is that the user begins the search knowing quite a bit about the temporal characteristics of the expected relevant results and is able to articulate them (since they are sophisticated searchers). Yet, in most search applications, there is no way for the user to provide this knowledge to the system. To fix this, we propose obtaining and using a temporal relevance profile $\mathcal{U}$, an explicit representation of the temporal characteristics of the results expected by the user. This specification is part of the input the searcher provides to the system.

Acquiring the profile $\mathcal{U}$ is non-trivial. It must be expressive enough to communicate the information need's temporal dynamics, but it must be relatively simple to obtain, so

as to avoid imposing too much burden on the user. For example, asking the user to directly specify the distribution, as in Figure 1, is likely too burdensome.

Balancing these issues, we propose four different types of temporal relevance profiles that can be fairly easy to solicit:

- *Soft interval.* Tweets occurring within a specified interval t_{min} and t_{max} are preferred, but this is not an absolute requirement. That is, the user specifies start and end times between which the relevant results are expected. This interval specification is particularly useful for the archive search scenario described above.
- *Hard interval.* Tweets *must* occur within a time interval (in contrast to the *preference* above). As with the soft interval, the user specifies a start t_{min} and end time t_{max}.
- *Recency bias.* Recent tweets are preferred, but this is not an absolute constraint. Note that although this condition could technically be encompassed by a soft interval, we view it as being conceptually distinct. This temporal relevance profile is typically anchored by t_q, the time when the query was issued, and is appropriate for the recency search scenario described above.
- *No bias.* Retrieved tweets should not be temporally biased one way or another.

We believe that expressing $\mathcal{U}$ along the lines of the options specified above would require little effort on the user's part. For example, specification of either hard or soft intervals could be accomplished by dragging sliders—such interface widgets are familiar to users in the context of browsing time series such stock prices, and would require little learning. These user-supplied temporal relevance profiles can then be converted into estimates of $f_t(t_d)$ and incorporated into the retrieval model per Equation (4). Of course, specifying a temporal profile would be optional; in the absence of specific temporal constraints, a system would fall back to some default behavior.

In conclusion, we believe that reframing the challenge of exploiting temporal signals for microblog search as an interactive retrieval problem is a promising avenue for future work. Why spend substantial effort trying to automatically infer the temporal characteristics of the information need if a system can simply elicit this information from the user in a lightweight manner?

5. ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grant Nos. 1144034, 1217279, and 1218043. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional support was provided by the Dutch National Institute for Mathematics and Computer Science (CWI), where the first author was a visiting researcher during Summer 2013. We'd also like to thank Arjen de Vries and Jiyin He for helpful discussions.

6. REFERENCES

[1] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pages 222–229, Berkeley, California, 1999.

[2] W. Dakka, L. Gravano, and P. G. Ipeirotis. Answering general time-sensitive queries. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):220–235, 2012.

[3] M. Efron. Linear time series models for term weighting in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(7):1299–1312, 2010.

[4] M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. In *Proceedings of the 34rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, pages 495–504, Beijing, China, 2011.

[5] J. L. Elsas and S. T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010)*, pages 1–10, New York, New York, 2010.

[6] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems*, 25(3):Article 14, 2007.

[7] X. Li and W. B. Croft. Time-based language models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM 2003)*, pages 469–475, New Orleans, Louisiana, 2003.

[8] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the TREC-2011 Microblog Track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*, Gaithersburg, Maryland, 2011.

[9] M.-H. Peetz and M. de Rijke. Cognitive temporal document priors. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR 2013)*, pages 318–330, Moscow, Russia, 2013.

[10] J. M. Ponte and W. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pages 275–281, Melbourne, Australia, 1998.

[11] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, Boca Raton, 1996.

[12] I. Soboroff, D. McCullough, J. Lin, C. Macdonald, I. Ounis, and R. McCreadie. Evaluating real-time search over tweets. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM 2012)*, pages 579–582, Dublin, Ireland, 2012.