Popularity-Based Temporal Relevance Estimation for Micro-Blogging Retrieval

Manar F. Alohaly
The College of Saint Rose
432 Western Ave.
Albany, NY 12203 USA
alohalym473@strose.edu

James D. Teresco
The College of Saint Rose
432 Western Ave.
Albany, NY 12203 USA
+1 (518) 485-3755
terescoj@strose.edu

ABSTRACT

Finding relevant information among the vast amounts of data generated continuously by modern micro-blogging platforms has opened new challenges in information retrieval. Recent studies on time-based retrieval have shown that identifying the relevant time periods to be incorporated into the retrieval process is promising; by relevant time period we mean the peak time of a query that satisfies the temporal needs of the user's query. Or in other words, a time period at which the potential to find accurate matches for the query in a set of retrieved documents is relatively high when compared with other time periods. We refer to this as temporal relevance. In this paper, using data collected from Twitter, we propose a new temporal relevance estimation technique based on tracking documents published by the popular users, who have high indegree (i.e., number of followers). In this study we concentrate on queries that are short (one word) and popular, i.e., constantly consumed by micro-bloggers. We choose the simple frequency-based technique to estimate the relevant time period as a baseline against which we evaluate our technique. The results of our technique either match or suggest a better time period as the most relevant one, when compared with the baseline. In fact, for the type of queries in our study, narrowing our focus to the documents published by popular users produces a query-to-documents matching pattern that uncovers some information about temporal relevance that might otherwise be hidden. Also, our matching pattern reflects the nature of the real-world news events that are related to the user's query more so than the baseline, thus revealing the important time frames more clearly.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Relevance feedback.*

General Terms

Algorithms, Experimentation.

Keywords

Temporal relevance, Micro-blogs, Micro-blogging Retrieval

1. INTRODUCTION

Micro-blogging is an emerging form of interpersonal communication that has become very popular in recent years. Micro-blogging services allow users to publish short text messages. These messages are broadcasted to the user's "followers" in real-time, and are generally accessible to all users of the platform (if the micro-blogger's account is public).

In this work, we are focusing on Twitter, which is the most popular micro-blogging service, estimated to have more than 554 million users as of December 2013[1]. Its users post limited-length messages, restricted to be no longer than 140 characters per posting, referred to as "tweets", in order to instantly share their current status or personal opinions. Beyond its micro-blogging features, Twitter is also a social networking service. Each micro-blogger has lists of other users who are his/her "followers" and whom he/she is "following". Being a follower means the user will receive the tweets that are either published or retweeted by those whom the user follows. Among the most frequently used terms in the Twittersphere are "hashtag" and "mention". A hashtag is one word or multiple words with no spaces in between, preceded by hash symbol #, and it is used to identify a topic of interest and facilitate a search for it. A mention is a way to reference a Twitter user within the tweet context by including his/her name preceded by the @ symbol.

Studies, e.g. [6], have shown that people searching for temporally relevant information often target Twitter with short queries related to popular topics. The distinctive features of the query and the real time nature of Twitter require the temporal relevance estimation to be quick and accurate in order to enhance the retrieval process.

We build our work based on the fact that a micro-blogging service contains a large amount of real-time information that makes it ideal to serve as an information-spreading media. The information is predominantly published by a group of users referred to as the *information providers*, among which is a subset of *popular users* who have a tremendous audience. Indeed, one study [4] conducted using Twitter as a micro-blogging service, has shown that the popular users are most often either celebrities or mass media. Thus, it is possible to make an analogy between the action of publishing breaking news in traditional news media and the action of publishing tweets in Twitter by popular micro-bloggers since they act as mass media. In fact, when a breaking event occurs, all traditional news media, e.g., television, radio and newspapers, are likely to cover the event at the same time. Thus, this time would be the one we would want to consider as the most relevant for this event. In this work, we hypothesize that we can estimate the relevant time period of a topic (query) of interest by identifying time intervals with the highest concentration of popular users who act as information-spreading media and publish documents that discuss the same topic at the same time.

Our method addresses the temporal relevance estimation for queries that are relatively short (often one word, e.g., "apple" or "twitter"), popular, ambiguous, and unevenly distributed over time. That is, there is always some ongoing discussion of a topic at a more or less uniform level, but with spikes in discussion activity that coincide with the occurrence of real-world events relevant to the topic. The queries of our interest lack any clue from which one could reasonably infer temporal information. In fact, documents containing the word(s) of such a query are very numerous, but they may or may not be semantically relevant to it. For example, this tweet from our dataset:

```
"Microsoft Word - Ch 3 Solutions IE done pdf
ebook:  http://t.co/nZTGU900 #google"
```

includes the hashtag "#google", even though it has no real relevance to Google. We refer to tweets of this kind as *irrelevant*. Such a tweet will not be of interest to a user who is searching for information about Google. Limiting our focus to documents that are published by popular users reduces the noise resulting from inappropriate use of hashtags and mentions, e.g., in an attempt to attract attention to a spam tweet. The results discussed below are promising, as the incorporation of the popularity of users makes our approach less likely to consider such irrelevant tweets as compared to the baseline mechanism. That is, a popular user is less likely to publish irrelevant tweets.

Before we detail our mechanism, we discuss some related work in Section 2 and describe the dataset in Section 3. In Section 4, we describe our approach, and present our experimental results in Section 5. Section 6 describes a generalization of our idea, and Section 7 suggests some conclusions and ideas for future work.

2. RELATED WORK

Works that have been conducted to detect temporal relevance revolve around three criteria. These are document and query text analysis, recency, and frequency distribution. Our work is a variation of the frequency distribution approach.

2.1 Text Analysis

Analyzing the query and document content to infer the temporal information is useful when the document creation time stamp is missing or unavailable. This is not the case with tweets. Also, the short text length of each tweet and the informal writing styles often used to compose tweets present significant challenges when applying this technique to Twitter data.

2.2 Recency

The recency criterion favors the most recently published documents. Early work by Xiaoyan and Bruce [5] has shown that incorporating document recency into ranking retrieval systems satisfies a group of queries targeting documents that are basically relevant to recent information or new events. Recency suggests that if a document is outdated then it is not relevant for search. However, in Twitter, the recent moment might not be rich enough for the type of queries used in micro-blogging. It is appropriate in this context to consider documents published further in the past, as those documents may be much more relevant to the query than recent ones.

2.3 Frequency Distribution

The third criterion to estimate the relevant time periods is frequency distribution based on the document publication time stamp. The basic form of this approach is the simple frequency-based technique which groups documents containing the query words based on the publication time, regardless how semantically relevant those documents are, e.g., a document might contain the word "#google" even if that hashtag was included out of context. Then it estimates temporal relevance of a time frame based on the count of matching documents in the time frame. The higher the count, the more relevant the time frame is assumed to be. Previous work conducted by Jones and Diaz [2] incorporated the document relevance into this technique, aiming to improve its accuracy. They defined the temporal relevance as the normalized sum of the relevance scores of documents that are published at time t for a query q. In fact, Jones and Diaz showed that this approach works very well in the context of retrieval of relevant news stories. However, in our context of micro-blogging, one should also consider that the average length of a query targeted to Twitter has been measured by one study [6] at 1.64 words, which limits the efficiency of a ranking based solely on the query. For this reason, we believe that using the temporal relevance could be a mechanism to improve the rankings of matching documents compared to using ranking scores to estimate the temporal relevance as in Jones and Diaz. Another study, by Dakka et al [1], was done using news article datasets to estimate the temporal relevance in order to address what they call the time-sensitive query. Their approach was built upon the frequency distribution mechanism. More specifically, they computed the temporal relevance of documents matching a query based on a sorting of the time periods (they use days) during which the documents were published, into bins of different priority levels. The highest priority bin contained the time periods with the highest document frequency. Then, they estimated the value of the temporal relevance based on the assignment of time periods to bins.

Exclusively relying on the document creation time stamp to plot the frequency distribution of the documents, and then using its outcomes to estimate the relevant time periods is a reasonable approach when the documents are news stories, since we know in advance that those articles are published by journalists, which gives confidence in the validity of the content. However, in Twitter anyone can publish documents so we have no confidence in the validity of documents if we consider those published by any user. Inspired by the environmental differences between the previous work and the micro-blogging environment (i.e., that we are unsure of the validity of many of the documents), we estimate the relevant time periods based on frequency distribution of those documents written only by popular users. By this, we aim to diminish the possibility of considering irrelevant documents, e.g., documents that include query words merely to gain advantage from such words' popularity. A popular user is more likely to publish a relevant document, and unlikely to include an intentionally irrelevant hashtag or mention (such as "@apple", which is one of the queries in our study) solely to try to attract attention to an otherwise irrelevant document. In fact, in the dataset described below, we find that 12% of tweets are marked as *irrelevant*, but only 3% of tweets by popular users are marked as *irrelevant*.

3. DATASET

We use the Sanders-Twitter Sentiment Corpus[2], which consists of 5513 Tweet ids from 15-20 of October 2011. The actual tweets were previously hand-classified as positive, neutral, negative, or irrelevant with respect to one of four different topics: Apple, Microsoft, Twitter and Google. Search terms were @apple, #microsoft, #twitter and #google, respectively.

Table 1. Classification categories of tweets in the Sanders-Twitter Sentiment Corpus.

Positive	The tweet is on-topic and has a positive indicator toward the query
Neutral	The tweet is on-topic, and its content is neutral toward the query
Negative	The tweet is on-topic and has a negative indicator toward the query
Irrelevant	The tweet contains the query word, but it is not on-topic or is out of context, e.g. spam

We use the Twitter API to fetch the actual tweets, publication time, the publishers, and the count of each publisher's followers. As some tweets had been deleted since this corpus was constructed, we were left with 4832 tweets.

4. EXPERIMENT

We start with filtering, to eliminate non-English tweets. For filtering, we use a language detection library[3] that is based on a Naive Bayesian filter. Its precision is over 99% for 53 languages. This left 3520 tweets in our dataset.

Since the tweets were already hand-classified to positive, neutral, negative, or irrelevant, we consider a tweet that is marked as one of the three former classes as *relevant* while those in the latter class are semantically *irrelevant*. We group the tweets by the hour during which they were published. For each single hour, we find the total number of tweets that correspond to each query of our four search terms considering (*i*) all tweets, which is the baseline against which we compare, (*ii*) only semantically relevant tweets and (*iii*) only those published by popular users, which is our primary mechanism. We point out that the second category, the semantically relevant tweets (those marked as positive, negative or neutral) uses information that, in general, is not known from a simple retrieval process. This makes it unreasonable to use as a general mechanism in the absence of such classification. Our conjecture is that the query-to-documents matching frequencies that result from our mechanism might help in two areas: identifying the relevant time periods for the query without being faked by considering irrelevant documents (in other words, the spam tweets), and better reflecting news events occurring in the real world above and beyond the usual ongoing discussion among users. We then give credits to each slot (hour) based on these counts. The count is proportional to the temporal relevance of the corresponding time interval, i.e., time frames with higher numbers of these tweets have the potential to be the more relevant time periods.

To measure a publisher's popularity we use indegree (the publisher's number of followers) as a popularity indicator. Despite the fact that this approach has its vulnerabilities since the indegree could be faked, Kwak et al. [4] found this popularity measurement to be similar to PageRank. Unlike PageRank, using indegree measurement does not require an iterative procedure to converge to the final result. This makes it faster and more appropriate for the real time nature of Twitter. Based on preliminary experiments, we use an indegree of 1000 as a threshold that indicates user popularity. The *Twitter Counter tool* suggests that an indegree of about 3000 is a better indication of popularity. However, our experimental results when using 3000 or higher indegree as the threshold gives a matching pattern with only subtle changes over the time. We find that using 1000 indegree as the threshold for popularity increases the chance to observe differences in matching frequency over the time. As this may depend on the specifics of our data set, it may be necessary when using larger datasets, to adjust the threshold to a higher indegree so that the popular users, who play the role here of mass media journalists in traditional settings, are identified appropriately.

For most of our experiments we use a one-hour slot for the frequency distribution over the time. Since the documents corresponding to the query "apple" are spread along a larger time period, we apply the same experiment using a one-day slot. The results in both cases, the one-hour and one-day time slots, are promising. In contrast, in both queries #twitter and #google, the retrieved documents were concentrated in only few hours during the day. Since identifying the peak time from such a short time frame is not likely to be reasonable thing to do, we show their results but do not use them to draw any meaningful conclusion.

5. EVALUATION

We plot the following histograms using the retrieved documents. Each histogram shows the simple frequency distribution for all retrieved documents (red line), the relevant documents (blue line), and those that are published by popular users (green line) during a day. The red line corresponds to the baseline while the green line corresponds to our mechanism. The purpose of excluding the irrelevant tweets and showing the distribution of only relevant ones (blue line) is to examine whether the peak time that is suggested by either of these mechanisms is faked due to the inclusion of irrelevant tweets. Also, this emphasizes our mechanism's ability to precisely reflect the real-world events and disregard the noise that results from the ongoing discussions. Therefore, it generates a matching pattern that better identifies the important time periods.

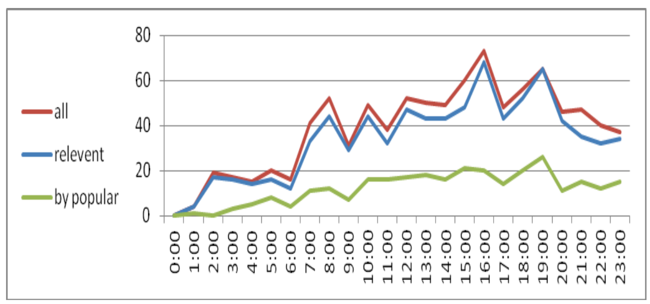

Figure 1. The first (red) line shows the frequency distribution of all English tweets that are retrieved using #microsoft as a query term, which includes both semantically relevant and irrelevant tweets. The second (blue) line shows the distribution of only relevant tweets. Finally, the third (green) line shows the distribution of tweets that are written by popular users. These tweets were published on October 19, 2011.

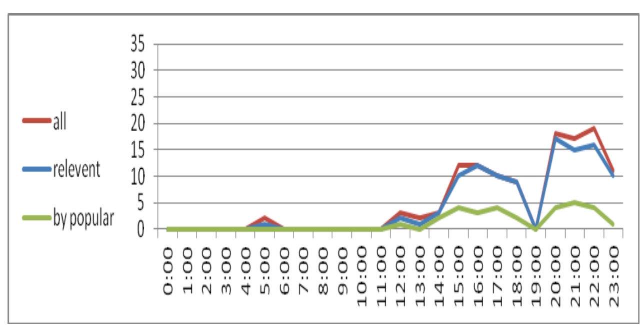

Figure 2. Here, we consider the tweets retrieved using @apple as a query term, in the same color scheme as used in Figure 1. These tweets were published on October 15, 2011.

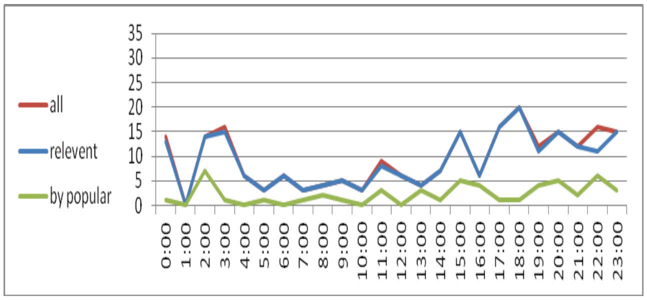

Figure 3. Here, we consider the tweets retrieved using @apple as a query term, in the same color scheme as used in Figure 1. These tweets were published on October 16, 2011.

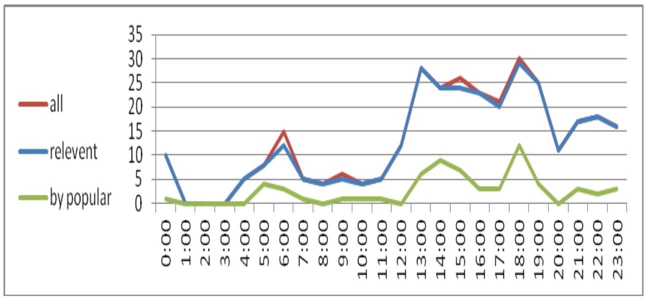

Figure 4. Here, we consider the tweets retrieved using @apple as a query term, in the same color scheme as used in Figure 1. These tweets were published on October 17, 2011.

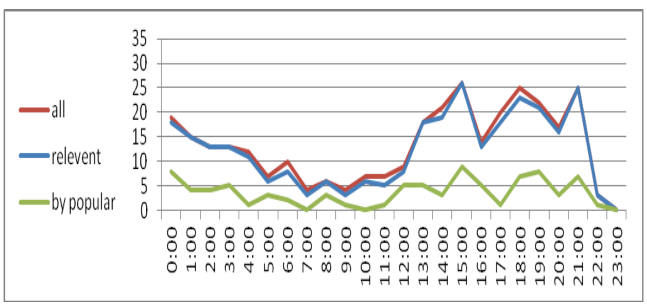

Figure 5. Here, we consider the tweets retrieved using @apple as a query term, in the same color scheme as used in Figure 1. These tweets were published on October 18, 2011.

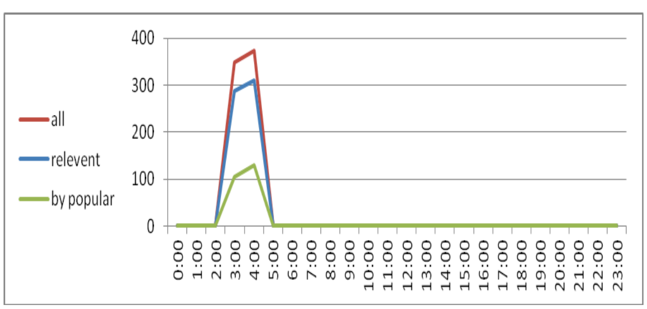

Figure 6. Here, we consider the tweets retrieved using #twitter as a query term, in the same color scheme as used in Figure 1. These tweets were published on October 20, 2011.

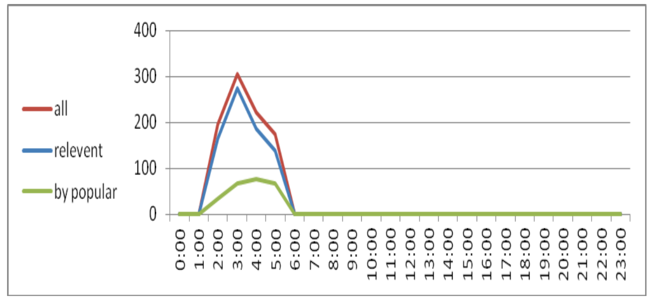

Figure 7. Here, we consider the tweets retrieved using #google as a query term, in the same color scheme as used in Figure 1. These tweets were published on October 19, 2011.

From the simple frequency distribution (red line) of the tweets that correspond to our query #microsoft (Figure 1) we could observe a zigzag pattern of documents matching the query. The constant change in the matching pattern implies that a popular query like #microsoft is continually discussed by the public. We consider this behavior from the public as noise that might hide some more important information. In such circumstances, it is hard to identify the important time merely from the simple frequency distribution. In contrast, the frequency distribution of the documents that are published by popular users tends to be more stable with few spikes. This reflects the nature of the news related to a topic; when no events are occurring related to that topic, the rate at which news stories about the topic are published is low and somewhat evenly distributed. The sudden change in the matching pattern indicates that the time when such change occurs must be significant to the query of interest. A closer look at Figure 1 shows that the counts of the *relevant* tweets (blue line) at 16:00 and 19:00 were nearly the same. The baseline mechanism favors the time slot 16:00 as it has more tweets, disregarding the fact that the difference resulted from the *irrelevant* tweets.

Moving from #microsoft to @apple, looking more closely at Figure 2, we see that each mechanism chooses a different most relevant time slot. The baseline chooses 22:00, but we observe that this is heavily influenced by a larger number of irrelevant tweets. Using the relevant tweets, 20:00 is chosen, which would be ideal. But recall that the relevance information is not generally available. Our mechanism suggests a slot (21:00), which while not as good as 20:00, does have fewer irrelevant tweets than the baseline's slot. We argue that this is a positive result. Should the documents published at 22:00 be chosen, for example, to perform query expansion, the results would be polluted by those irrelevant (spam) tweets ultimately generating misleading search results. For example the following tweets, corresponding to the query @apple, were published at the time slot suggested as a peak time in the baseline (22:00):

- "Find more the best Angel Christmas Tree O http://t.co/d5G1rY5d@Starfrit @Apple @Peeler"
- "Find more the best Cyst On Knee deal.Check price of best Cyst On Knee and offers now. http://t.co/qzYnNuvI@Automatic @Apple"
- "Find more the best Maytag Washer Control Board deal.Check pric http://t.co/SNdkAg3e@Best @Apple @Peeler."

Using such *irrelevant* tweets, e.g., in a query expansion, will produce misleading results. In Figure 3, we see that our mechanism suggests a slot that is 4th best according to the baseline. However, both slots, suggested by the baseline and our mechanism, give an equal percentage of the irrelevant tweets. The rest of the histograms in Figures 4-5 show that the results of our technique and baseline are almost proportional to each other.

It is worth mentioning that on October 18, Apple reported fourth quarter results, and on October 17, Apple announced that sales of the iPhone surpassed 4 million. Also, October 16 was declared in the state of California to be "Steve Jobs day[4]". Due to these events occurring at that time, both the baseline and our technique show continuous change in the matching pattern.

As we mentioned previously, in the dataset that is used in this study, all tweets corresponding to #twitter (Figure 6) and #google (Figure 7) were retrieved within a very short time frame. The results are included for completeness (they are part of our corpus) but we do not draw any conclusions from them.

The following table compares our technique to identify the peak time, frequency distribution of tweets published by popular users, against the baseline. Here we use a one-hour time slot. For each search term we report the peak time and the percentage of irrelevant tweets at that time in both techniques. In fact, our procedure either matches the baseline estimation or suggests a different time with a lower percentage of irrelevant tweets. This will inevitably reflect on the efficiency of the temporal retrieval process as we rely on time slots that have more concentration of accurately relevant documents.

Table 2. Comparison of the peak time (by hour) and the percentage of irrelevant tweets within the suggested peak time in both the baseline and our (by popular) approaches.

Query/Day	Simple frequency distribution		Published by popular distribution	
	Peak Time	% of irrelevant	Peak Time	% of irrelevant
#microsoft 19 Oct	16:00	6.85%	19:00	0.00%
@apple 15 Oct	22:00	15.79%	21:00	11.76%
@apple 16 Oct	18:00	0.00%	02:00	0.00%
@apple 17 Oct	18:00	3.33%	18:00	3.33%
@apple 18 Oct	15:00	0.00%	15:00	0.00%
#twitter 20 Oct	04:00	16.89%	04:00	16.89%
#google 19 Oct	03:00	11%	04:00	16%

Using the same frequency distribution mechanisms, we repeat the previous comparison for all tweets corresponding to the query @apple. But, this time we use a one-day time slot, as shown in Figure 8.

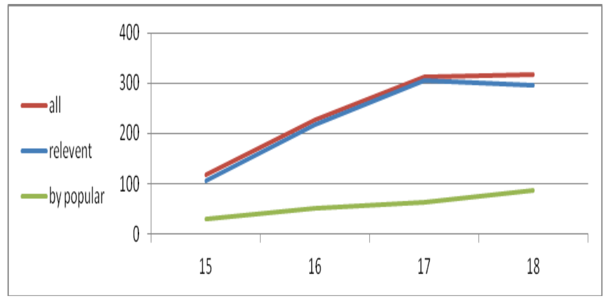

Figure 8. Here, we consider the tweets retrieved using @apple as a query term, in the same color scheme as used in Figure 1. These tweets were published from October 15, 2011 to October 18, 2011.

Table 3. Comparison of the peak time (by day) and the percentage of tweets that are irrelevant to the query @apple within the peak time.

Days	Simple frequency distribution		Published by popular distribution	
	Peak Time	% of irrelevant	Peak Time	% of irrelevant
15-19 Oct, 2011	18 Oct	6.00%	18 Oct	6.00%

We make several observations about Figure 8. It is clear that both techniques agree on the same peak time, which is October 18. However, the matching pattern in the simple frequency distribution (baseline) increases radically until October 17. Then, it remains at an almost constant level, which might suggest that there is nothing new after October 17. However, we pointed out previously that there is relevant news regarding Apple on each of those days. With the baseline technique, there is very little difference between October 17 and October 18, so we cannot be confident that the peak is really October 18 as opposed to October 17. In contrast, the matching pattern resulting from our technique keeps increasing. Such change in the matching pattern over the time reflects the fact that there is something significant going on. Sometimes, it happens that both techniques, the baseline and ours, suggest the same peak time. It is also noticeable that there is a small difference between the distribution of the full set (red line) and the relevant set (blue line), which means that the baseline does its job fairly well. So why bother with the new mechanism? Figure 8 shows that our mechanism better reveals some information (beyond just picking the most relevant time) that might remain hidden when using the baseline technique due to the noise resulting from considering all micro-bloggers.

6. COMPUTING TEMPORAL RELEVANCE

Results presented here have focused on finding the single most relevant time period for a given query term. In the previous sections, we showed that the popularity-based mechanism is an effective way to achieve this. We now introduce our temporal query model to estimate how relevant each time frame is to a query, which we define formally as $P(q|t_i)$. Let q be the query of interest and let t_i be a single time frame (e.g., a single hour or day) within a time interval of interest, $T = \bigcup_{i=1}^{k} t_i$. If $n_{pop}(t,q)$ is the number of tweets published by popular users that match query q during the time frame t_i, and N(T,q) is the total number of tweets published that match query q over our time of interest T, we define:

$$P(q|t_i)= n_{pop}(t_i,q)/N(T,q).$$

Normalizing by the total number of matching tweets rather than the total number of matching tweets published by popular users better estimates the actual importance of the time frame. Had two popular users published tweets in a time frame, this would not necessarily signal the importance of this time frame.

We propose this as an alternative to estimation mechanisms suggested by other studies, e.g. [1] and [2]. Our popularity-based values could be incorporated into the retrieval process in similar ways, e.g. as in [1].

7. CONCLUSION AND FUTURE WORK

Concentrating on the tweets from popular users when determining important time periods for a given query term produces better matches to real news events and reveals information that might otherwise be hidden by the background noise of ongoing discussion. It also might help in identifying relevant time periods with smaller percentages of irrelevant tweets.

Our outcomes could be used to enhance ranking retrieval systems and extend the traditional query-likelihood language model [3]. The temporal information is also incorporated into the query expansion process, which is called temporal query-expansion, e.g. [7]. This approach scores candidate expansion terms according to the temporal feedback. More generally, our outcomes are also of potential value to others who are concerned with identifying important time intervals. We consider this a potential direction for the future.

For our next steps, we would like to apply our mechanism to more and larger datasets. Also, we are interested in incorporating our mechanism in complete micro-blogging retrieval context. This would also allow us to evaluate the temporal relevance measure introduced in Section 6, which was beyond the scope of our current effort, as we focused on identifying the relevant time periods. Then we plan to consider two dimensions related to publisher's popularity: *(i)* the publisher's topical popularity, or in other words the publisher's popularity for a given topic, and *(ii)* the rise and fall of a user's popularity over time. Then, we will examine how far those two dimensions might enhance the peak time estimation. Finally, we would like to experiment with different measurements of a user's popularity beyond a simple threshold of the number of followers, for comparison purposes.

8. ACKNOWLEDGMENT

We would like to thank Dr. Sharon Gower Small, Director of the Siena College Institute for Artificial Intelligence, for her helpful suggestions and comments, which have greatly improved our work.

9. REFERENCES

[1] Dakka, W., Gravano, L., and Ipeirotis, P.G. 2008. Answering general time sensitive queries. In *Proceedings of the 17th ACM conference on Information and knowledge management* (CIKM '08). ACM, New York, NY, USA, 1437-1438. DOI= http://doi.acm.org/10.1145/1458082.1458320

[2] Jones, R. and Diaz, F. 2007. Temporal profiles of queries. *ACM Trans. Inf. Syst.* 25, 3, Article 14 (July 2007). DOI=http://doi.acm.org/10.1145/1247715.1247720

[3] Kumar, N. and Carterette, B. 2013. Time based feedback and query expansion for twitter search. In *Proceedings of the 35th European conference on Advances in Information Retrieval* (ECIR'13), Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, and Stefan Rüger (Eds.). Springer-Verlag, Berlin, Heidelberg, 734-737. DOI=http://dx.doi.org/10.1007/978-3-642-36973-5_72

[4] Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 591-600. DOI= http://doi.acm.org/10.1145/1772690.1772751

[5] Li, X. and Croft, W.B. 2003. Time-based language models. In *Proceedings of the 12th international conference on Information and knowledge management* (CIKM '03). ACM, New York, NY, USA, 469-475. DOI= http://doi.acm.org/10.1145/956863.956951

[6] Teevan, J., Ramage, D., and Morris, M.R. 2011. #TwitterSearch: a comparison of microblog search and web search. In *Proceedings of the 4th ACM international conference on Web search and data mining* (WSDM '11). ACM, New York, NY, USA, 35-44. DOI= http://doi.acm.org/10.1145/1935826.1935842

[7] Willis, C., Medlin, R. and Arguello, J. 2012. Incorporating Temporal Information in Microblog Retrieval. In *Proceedings of the 21st Text REtrieval Conference* (TREC'12), National Institute of Standards and Technology, special publication.

[1] http://www.statisticbrain.com/twitter-statistics/

[2] http://www.sananalytics.com

[3] https://code.google.com/p/language-detection/

[4] http://www.cultofmac.com/196211/october-16th-is-steve-jobs-day-how-will-you-be-celebrating/