Structured Event Retrieval over Microblog Archives

Donald Metzler, Congxing Cai, Eduard Hovy
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292

Abstract

Microblog streams often contain a considerable amount of information about local, regional, national, and global events. Most existing microblog search capabilities are focused on recent happenings and do not provide the ability to search and explore past events. This paper proposes the problem of structured retrieval of historical event information over microblog archives. Rather than retrieving individual microblog messages in response to an event query, we propose retrieving a ranked list of historical event summaries by distilling high quality event representations using a novel temporal query expansion technique. The results of an exploratory study carried out over a large archive of Twitter messages demonstrates both the value of the microblog event retrieval task and the effectiveness of our proposed search methodologies.

1 Introduction

Real-time user generated content is one of the key driving forces behind the growing popularity of social media-centric communication. The ability to instantly share, often from your mobile phone, your thoughts (via Twitter), your photos (via Facebook), your location (via Foursquare), and a variety of other information is changing the way that information is created, communicated, and consumed.

There has been a substantial amount of research effort devoted to user generated content-related search tasks, including blog search, forum search, and community-based question answering. However, there has been relatively little research on *microblog search*. Microblog services, such as Tumblr and Twitter, provide users with the ability to broadcast short messages in real-time. This is in contrast to traditional blogs that typically have considerably more content that is updated less frequently. By their very nature, microblog streams often contain a considerable amount of information about local, regional, national, and global news and events. A recent study found that over 85% of trending topics on Twitter are news-related (Kwak et al., 2010). Another recent study by Teevan et al. that investigated the differences between microblog and Web search reported similar findings (Teevan et al., 2011). The study also found that microblog search queries are used to find information related to news and events, while Web search queries are more navigational in nature and used to find a variety of information on a specific topic.

It is likely that microblogs have not received much attention because, unlike blog search, there is no well-defined microblog search *task*. Existing microblog search services, such as those offered by Twitter and Google, only provide the ability to retrieve individual microblog posts in response to a query. Unfortunately, this task has limited utility since very few real information needs can be satisfied by a single short piece of text (e.g., the maximum length of a message on Twitter is 140 characters). Hence, novel search tasks defined over microblog streams that go beyond "message retrieval" have the potential to add substantial value to users.

Given the somewhat limited utility of microblog message search and the preponderance of news and event-related material posted on microblogs, this pa-

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 646–655,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

July 16 2010 at 17 UTC, for **11** hours

Summary tweets:

i. *Ok a* 3.6 *“rocks” nothing. But boarding a plane there now, Woodward ho! RT @todayshow:* 3.6 *magnitude #earthquake rocks Washington DC area.*

ii. *RT @fredthompson: 3.6-magnitude earthquake hit DC. President Obama said it was due to 8 years of Bush failing to regulate plate tectonic ...*

iii. *3.6-magnitude earthquake wakes Md. residents: Temblor centered in Gaithersburg felt by as many as 3 million people...* `http://bit.ly/9iMLEk`

Figure 1: Example structured event representation retrieved for the query “earthquake”.

per proposes a novel search task that we call *microblog event retrieval*. Given a query that describes an event, such as *earthquake*, *terrorist bombing*, or *bieber concert*, the goal of the task is to retrieve a ranked list of *structured event representations*, such as the one shown in Figure 1, from a large archive of historical microblog posts.

In this work, structured representations come in the form of a list of *timespans* during which an instance of the event occurred and was actively discussed within the microblog stream. Additionally, for each timespan, a small set of relevant messages are retrieved for the purpose of providing a high-level summary of the event that occurred during the timespan. This task leverages the large amount of real-time, often first-hand information found in microblog archives to deliver a novel form of user generated content-based search results to users. Unlike news search, which finds professionally written articles on a news-related topic, and general-purpose Web search, which is likely to find a large amount of unrelated information, this task is designed to retrieve highly relevant news and event-related information viewed through the lens of users who experienced or discussed the event while it happened (or during its aftermath). Such search functionality would not only be useful for everyday end-users, but also social scientists, historians, journalists, and emergency planners.

This paper has three primary contributions. First, we introduce the microblog event retrieval task, which retrieves a ranked list of structured event representations in response to an event query. By going beyond individual microblog message retrieval, the task adds value to microblog archives and provides users with the ability to find information that was disseminated in real-time about past events, which is not possible with news and Web search engines. Second, we propose an unsupervised methodology for distilling high quality event representations using a novel temporal query expansion technique. The technique synthesizes ideas from pseudo-relevance feedback, term burstiness, and temporal aspects of microblog streams. Third, we perform an exploratory evaluation of 50 event queries over a corpus of 46 million Twitter messages. The results of our evaluation demonstrate both the value of the microblog event retrieval task itself and the effectiveness of our proposed search methodologies, which show improvements of up to 42% compared to a baseline approach.

2 Related Work

There are several directions of microblog research that are related to our proposed work. First, there is a growing body of literature that has focused on the topical content of microblog posts. This research has focused on microblog topic models (Hong and Davison, 2010), event and topic detection and tracking (Sankaranarayanan et al., 2009; Cataldi et al., 2010; Petrović et al., 2010; Lin et al., 2010), predicting flu outbreaks using keyword tracking (Culotta, 2010), and using microblog streams as a source of features for improving recency ranking in Web search (Dong et al., 2010). Most of these approaches analyze content as it arrives in the system. While tracking a small number of topics or keywords is feasible using *online algorithms*, the general problem of topic detection and tracking (Allan et al., 1998) is considerably more challenging given the large number of topics being discussed at any one point. Our work differs in that it does not attempt to track or model topics as they arrive in the system. Instead, given an event query, our system *retrospectively* analyzes the corpus of microblog messages for the purpose of retrieving structured event representations.

There is no shortage of previous work on using pseudo-relevance feedback approaches for query expansion. Relevant research includes classical vector-space approaches (Rocchio, 1971), language modeling approaches (Lavrenko and Croft, 2001; Zhai and Lafferty, 2001; Li and Croft, 2003), among others (Metzler and Croft, 2007; Cao et al., 2008; Lv and Zhai, 2010). The novel aspect of our proposed temporal query expansion approach is the fact that expansion is done over a *temporal stream* of very short, noisy messages.

There has also been recent work on summarizing sets of microblog posts (Sharifi et al., 2010). We chose to make use of a simple approach in favor of a more sophisticated one because summarization is only a minor aspect of our proposed framework.

Finally, there are two previous studies that are the most relevant to our work. First, Massoudi et al. propose a retrieval model that uses query expansion and microblog quality indicators to retrieve individual microblog messages (Massoudi et al., 2011). Their proposed query expansion approach differs from ours in the sense that we utilize *timespans* from the (possibly distant) past when generating expanded queries and focus on event retrieval, rather than individual message retrieval. The other research that is closely related to ours is the work done by Chieu and Lee (Chieu and Lee, 2004). The authors propose an approach for automatically constructing timelines from news articles in response to a query. The novelty of our proposed work derives from our novel temporal query expansion approach, and the fact that our work focuses on microblog streams which are fundamentally different in nature from news articles.

3 Microblog Event Retrieval

The primary goal of this paper is to introduce a new microblog search paradigm that goes beyond retrieving messages individually. We propose a novel task called *microblog event retrieval*, which is defined as follows. Given a query that specifies an event, retrieve a set of relevant *structured event representations* from a large archive of microblog messages. This definition is purposefully general to allow for a broad interpretation of the task.

There is nothing in our proposed retrieval framework that precludes it from producing reasonable results for any type of query, not just those related to events. However, we chose to primarily focus on events in this paper because previous studies have shown that a majority of trending topics within microblog streams are about news and events (Kwak et al., 2010). The information found in microblogs is difficult to find anywhere else, including news and Web archives, thereby making it a valuable resource for a wide variety of users.

3.1 Overview of Framework

Our microblog event retrieval framework takes a query as input and returns a ranked list of structured event representations. To accomplish this, the framework breaks the work into two steps – timespan retrieval and summarization. The timespan retrieval step identifies the timespans when the event happened, while the summarization step retrieves a small set of microblog messages for each timespan that are meant to act as a summary. Figure 1 shows an example result that is returned in response to the query "earthquake". The result consists of a start time that indicates when the event began being discussed, a duration that specifies how long the event was discussed, and a small number of messages posted during the time interval that are meant to summarize what happened. This example corresponds to an earthquake that struck the metropolitan District of Colombia area in the United States. The earthquake was heavily discussed for nearly 11 hours, because it hit a densely populated area that does not typically experience earthquakes.

3.2 Temporal Query Expansion

We assume that queries issued to our retrieval framework are simple keyword queries that consist of a small number of terms. This sparse representation of the user's information need makes finding relevant messages challenging, since microblog messages that are highly related to the query might not contain any of the query keywords. It is common for microblog messages about a given topic to express the topic in a different, possibly shortened or slang, manner. For example, rather than writing "earthquake", users may instead use the word "quake" or simply include a hashtag such as "#eq" in their message. It is impractical to manually identify the full set of related keywords and folksonomy tags (i.e., hashtags) for each query. In information retrieval, this is known as the *vocabulary mismatch* problem.

To address this problem, we propose a novel unsupervised temporal query expansion technique. The approach is unsupervised in the sense that it makes use of a pseudo-relevance feedback-like mechanism when extracting expansion terms. Traditional query expansion approaches typically find terms that commonly co-occur with the query terms in documents (or passages). However, such approaches are not suitable for expanding queries in the microblog setting since microblog messages are very short, yielding unreliable co-occurrence information. Furthermore, microblog messages have an important temporal dimension that should be considered when they are being used to generate expansion terms.

Our proposed approach generates expansion terms based on the temporal co-occurrence of terms. Given keyword query q, we first automatically retrieve a set of N timespans for which the query keywords were most heavily discussed. To do so, we rank timespans according to the proportion of messages posted during the timespan that contain one or more of the query keywords. This is a simple, but highly reliable way of identifying timespans during which a specific topic is being heavily discussed. These timespans are then considered to be *pseudo-relevant*. In our experiments, the microblog stream is divided into hours, with each hour corresponding to an atomic timespan. Although it is possible to define timespans in many different ways, we found that this was a suitable level of granularity for most events that was neither overly broad nor overly specific.

For each pseudo-relevant timespan, a *burstiness score* is computed for all of the terms that occur in messages posted during the timespan. The burstiness score is meant to quantify how trending a term is during the timespan. Thus, if the query is being heavily discussed during the timespan and some term is also trending during the timespan, then the term may be related to the query. For each of the top N time intervals, the burstiness score of each term is computed as follows:

$$burstiness(w, TS_i) = \frac{P(w|TS_i)}{P(w)} \qquad (1)$$

which is the ratio of the term's likelihood of occurring within timespan TS_i versus the likelihood of the term occurring during any timespan. Hence, if a term that generally infrequently occurs within the message stream suddenly occurs many times within a single time interval, then the term will be assigned a high burstiness score. This weighting is similar in nature to that proposed by Ponte for query expansion within the language modeling framework for information retrieval (Ponte, 1998). The following probability estimates are used for the expressions within the burstiness score:

$$P(w|TS_i) = \frac{tf_{w,TS_i} + \mu\frac{tf_w}{N}}{|TS_i| + \mu}, P(w) = \frac{tf_w + K}{N + K|V|}$$

where tf_{w,TS_i} is the number of occurrences of w in timespan TS_i, tf_w is the number of occurrences of w in the entire microblog archive, $|TS_i|$ is the number of terms in timespan TS_i, N is the total number of terms in the microblog archive, V is the vocabulary size, and μ and K are smoothing parameters.

While it is common practice to smooth $P(w|TS_i)$ using Dirichlet (or Bayesian) smoothing (Zhai and Lafferty, 2004), it is less common to smooth the general English language model $P(w)$. However, we found that this was necessary since term distributions in microblog services exhibit unique characteristics. By smoothing $P(w)$, we dampen the effect of overweighting very rare terms. In our experiments, we set the value of μ to 500 and K to 10 after some preliminary exploration. We found that the overall system effectiveness is generally insensitive to the choice of smoothing parameters.

The final step of the query expansion process involves aggregating the burstiness scores across all pseudo-relevant timespans to generate an overall score for each term. To do so, we compute the *geometric mean* of the burstiness scores across the pseudo-relevant timespans. Preliminary experiments showed that the arithmetic mean was susceptible to overweighting terms that had a very large burstiness score in a single timespan. By utilizing the geometric average instead, we ensure that the highest weighted terms are those that have large weights in a large number of the timespans, thereby eliminating spurious terms. Seo and Croft (2010) observed similar results with traditional pseudo-relevance feedback techniques.

The k highest weighted terms are then used as expansion terms. Using this approach, terms that commonly trend during the same timespans that the query terms commonly occur (i.e., the pseudo-relevant timespans) are assigned high weights. Hence, the approach is capable of capturing simple temporal dependencies between terms and query keywords, which is not possible with traditional approaches.

3.3 Timespan Ranking

The end result of the query expansion process just described is an expanded query q' that consists of a set of k terms and their respective weights (denoted as β_w). Our framework uses the expanded query q' to retrieve relevant timespans. We hypothesize that using the expanded version of the query for timespan retrieval will yield significantly better results than using the keyword version.

To retrieve timespans, we first identify the 1000 highest scoring timespans (with respect to q'). We then merge contiguous timespans into a single, longer timespan, where the score of the merged timespan is the maximum score of its component timespans. The final ranked list consists of the merged timespans. Therefore, although our timespans are defined as hour intervals, it is possible for our system to return longer (merged) timespans.

We now describe two scoring functions that can be used to compute the relevance of a timespan with respect to an expanded query representation.

3.3.1 Coverage Scoring Function

The coverage scoring function measures relevance as the (weighted) number of expansion terms that are covered within the timespan. This measure assumes that the expanded query is a faithful representation of the information need and that the more times the highly weighted expansion terms occur, the more relevant the timespan is. Using this definition, the coverage score of a time interval is computed as:

$$s(q', TS) = \sum_{w \in q'} \beta_w \cdot tf_{w,TS}$$

where $tf_{w_i,TS}$ is the term frequency of w_i in timespan TS and β_w is the expansion weight of term w.

3.3.2 Burstiness Scoring Function

Since multiple events may occur at the same time, microblog streams can easily be dominated by the larger of two events. However, less popular events may also exhibit burstiness at the same time. Therefore, another measure of relevance is the burstiness of the event signature during the timespan. If all of the expansion terms exhibit burstiness during the time interval, it strongly suggests the timespan may be relevant to the query.

Therefore, to measure the relevance of the timespan, we first compute the burstiness scores for all of the terms within the time interval. This yields a vector β_{TS} of burstiness scores. The cosine similarity measure is used to compute the similarity between the query burstiness scores and the timespan burstiness scores. Hence, the burstiness scoring function is computed as:

$$s(q', TS) = cos(\beta_{q'}, \beta_{TS})$$

3.4 Timespan Summarization

The final step of the retrieval process is to produce a short query-biased summary for each retrieved time interval. The primary purpose for generating this type of summary is to provide the user with a quick overview of what happened during the retrieved timespans.

We utilize a simple, straightforward approach that generates unexpectedly useful summaries. Given a timespan, we use a relatively simple information retrieval model to retrieve a small set of microblog messages posted during the timespan that are the most relevant to the expanded representation of the original query. These messages are then used as a short summary of the timespan.

This is accomplished by scoring a microblog message M with respect to an expanded query representation q' using a weighted variant of the query likelihood scoring function (Ponte and Croft, 1998):

$$s(q', M) = \sum_{w \in q'} \beta_w \cdot \log P(w|M)$$

where β_w is the burstiness score for expansion term w and $P(w|M)$ is a Dirichlet smoothed language modeling estimate for term w in message M. This scoring function is also equivalent to the cross entropy and KL-divergence scoring functions (Lafferty and Zhai, 2001).

Category	Events
Business	layoffs, bankruptcy, acquisition, merger, hostile takeover
Celebrity	wedding, divorce
Crime	shooting, robbery, assassination, court decision, school shooting
Death	death, suicide, drowned
Energy	blackout, brownout
Entertainment	awards, championship game, world record
Health	recall, pandemic, disease, flu, poisoning
Natural Disaster	hurricane, tornado, earthquake, flood, tsunami, wildfire, fire
Politics	election, riots, protests
Terrorism	hostage, explosion, terrorism, bombing, terrorist attack, suicide bombing, hijacked
Transportation	plane crash, traffic jam, sinks, pileup, road rage, train crash, derailed, capsizes

Table 1: The 50 event types used as queries during our evaluation, divided into categories.

4 Experiments

This section describes our empirical evaluation of the proposed microblog event retrieval task.

4.1 Microblog Corpus

Our microblog message archive consists of data that we collected from Twitter using their Streaming API. The API delivers a continuous 1% random sample of public Twitter messages (also called "tweets"). Our evaluation makes use of data collected between July 16, 2010 and Jan 1st, 2011. After eliminating all non-English tweets, our corpus consists of 46,611,766 English tweets, which corresponds to roughly 10,000 tweets per hour. Although this only represents a 1% sample of all tweets, we believe that the corpus is sizable enough to demonstrate the utility of our proposed approach.

4.2 Event Queries

To evaluate our system, we prepared a list of 50 event types that fall into 11 different categories. The event types and their corresponding categories are listed in Table 1. The different event types can have substantially different characteristics, such as the frequency of occurrence, geographic or demographic interest, popularity, etc. For example, there are more weddings than earthquakes. Public events, such as federal elections involve people across the country. However, a car pileup typically only attracts local attention. Moreover, microbloggers show different amounts of interest to each type of event. For example, Twitter users are more likely to tweet about politics than a business acquisition.

4.3 Methodology

To evaluate the quality of a particular configuration of our framework, we run the *microblog event retrieval* task for the 50 different event type queries described in the previous section. For each query, the top 10 timespans retrieved are manually judged to be relevant or non-relevant. If the summary returned clearly indicated a real event instance occurred, then the timespan was marked as relevant. The primary metric of interest is precision at 10.

In addition to the temporal query expansion approach (denoted TQE), we also ran experiments using relevance-based language models, which is a state-of-the-art query expansion approach (Lavrenko and Croft, 2001). We ran two variants of relevance-based language models. In the first, query expansion was done using the Twitter corpus itself (denoted TwitterRM). This allows us to compare the effectiveness of the TQE approach against a more traditional query expansion approach. In the other variant, query expansion was done using the English Gigaword corpus (denoted NewsRM), which is a rich source of event information created by traditional news media.

For all three query expansion approaches (TQE, TwitterRM, and NewsRM), the two scoring functions, *burstiness* and *coverage*, are used to rank timespans. Hence, we evaluate six specific instances of our framework. As a baseline, we use a simple (unexpanded) keyword retrieval approach that scores timespans according to the relative frequency of event keywords that occur during the timespan.

4.4 Timespan Retrieval Results

Before delving into the details of our quantitative evaluation of effectiveness, we provide an illustrative example of the type of results our system is capable of producing. Table 2 shows the top four re-

July 16 2010 at 17 UTC, for **11** hours *Ok a* 3.6 *“rocks” nothing. But boarding a plane there now, Woodward ho! RT @todayshow:* 3.6 *magnitude #earthquake rocks Washington DC area.*
September 28 2010 at 11 UTC, for **6** hours *RT @Quakeprediction: 2.6 earthquake (possible foreshock) hits E of Los Angeles;* `http://earthquake.usgs.gov/earthquakes/recenteqscanv/Fau`...
September 04 2010 at 01 UTC, for **3** hours *7.0 quake strikes New Zealand - A 7.0-magnitude earthquake has struck near New Zealand’s second largest city. Reside...* `http://ht.ly/18R2rw`
October 27 2010 at 01 UTC, for **5** hours *RT @SURFER_Magazine: Tsunami Strikes Mentawais: Wave Spawned By A 7.5-Magnitude Earthquake Off West Coast Of Indonesia* `http://bit.ly/8Z9Lbv`

Table 2: Top four timespans (with a single summary tweet) retrieved for the query “earthquake”.

sults retrieved using temporal query expansion with the burstiness scoring function for the query “earthquake”. Only a single summary tweet is displayed for each timespan due to space restrictions. As we can see from the tweets, all of the results are relevant to the query, in that they all correspond to times when an earthquake happened and was actively discussed on Twitter. Different from Web and news search results, these types of ranked lists provide a clear temporal picture of relevant events that were actively discussed on Twitter.

The results of our microblog retrieval task are shown in Table 3. The table reports the per-category and overall precision at 10 for the baseline, and the six configurations of our proposed framework. Bolded values represent the best result per category. As the results show, using temporal query expansion with burstiness ranking yields a mean precision at 10 of 61%, making it the best overall system configuration. The approach is 41.9% better than the baseline, which is statistically significant according to a one-sided paired t-test at the $p < 0.01$ level. Interestingly, the relevance model-based expansion techniques exhibit even worse performance, on average, than our simple keyword baseline. For example, the news-based expansion approach was 11.6% worse using the coverage scoring function and 18.6% worse using the burstiness scoring function compared to the baseline. All of the traditional query expansion results are statistically significantly worse than the temporal query expansion-based approaches. Hence, the results suggest that capturing temporal dependencies between terms yields better expanded representations than simply capturing term co-occurrences, as is done in traditional query expansion approaches.

The results also indicate the burstiness scoring function outperforms the coverage scoring function for temporal query expansion. An analysis of the results revealed that in many cases the timespans returned using the coverage scoring function had a small number of frequent terms that matched the expanded query. This happened less often with the burstiness scoring function, which is based on the cosine similarity between the query and timespan’s burstiness scores. The combination of burstiness weighting and l_2 normalization (when computing the cosine similarity) appears to yield a more robust scoring function.

4.5 Event Popularity Effects

It is also interesting to note that the retrieval performance varies substantially across the different event type categories. For example, the performance on queries about “natural disasters” and “politics” is consistently strong. Similar performance can also be achieved for popular events related to celebrities. However, energy-related event queries, such as “blackout”, achieves very poor effectiveness. This observation seems to suggest that the more popular an event is, the better the retrieval performance that can be achieved. This is a reasonable hypothesis since the more people tweet about the event, the easier it is to identify the trend from the background.

To better understand this phenomenon, we compute the correlation between timespan retrieval precision and event (query) popularity, where popularity is measured according to:

$$Popularity(q) = \frac{1}{N}\sum_{i=1}^{N} burstiness(q, TS_i),$$

where q is the event query, $burstiness(q, TS_i)$ is the burstiness score of the event during timespan

Event Category	Baseline	NewsRM		TwitterRM		TQE	
		burst	cover	burst	cover	burst	cover
Business	0.50	0.46	0.30	0.70	0.18	**0.74**	0.64
Celebrity	0.75	0.30	0.40	0.50	0.60	**0.80**	0.45
Crime	0.44	0.28	**0.54**	0.22	0.32	0.46	0.28
Death	0.43	0.20	0.33	0.30	0.30	**0.47**	**0.47**
Energy	0.05	0.10	0.05	**0.20**	0.05	0.15	0.00
Entertainment	0.47	0.53	0.67	0.30	0.53	**0.70**	**0.70**
Health	0.48	0.28	0.36	0.44	0.16	**0.60**	**0.60**
Nat. Disaster	0.50	0.53	0.59	0.66	0.46	**0.87**	0.66
Politics	0.67	0.70	0.53	0.63	0.30	**0.87**	0.60
Terrorism	0.41	0.44	0.39	0.39	0.17	**0.69**	0.51
Transportation	0.21	0.08	0.08	0.08	0.10	**0.31**	0.19
All	0.43	0.35	0.38	0.40	0.26	**0.61**	0.47

Table 3: Per-category and overall (All) precision at 10 for the keyword only approach (Baseline), traditional newswire expansion (NewsRM), traditional pseudo relevance feedback using the Twitter corpus (TwitterRM), and temporal query expansion (TQE). For the expansion-based approaches, results for the burstiness scoring (burst) and the coverage-based scoring (cover) are given. Bold values indicate the best result per category.

	Correlation
Baseline	0.63 $(p < 0.01)$
NewsRM	0.53 $(p < 0.01)$
TwitterRM	0.61 $(p < 0.01)$
TQE	0.50 $(p < 0.01)$

Table 4: Spearman rank correlation between event retrieval precisions and event popularity. All methods use the burstiness scoring function.

TS_i, as defined in Equation 1, and the sum goes over the top N timespans retrieved for the event using our proposed retrieval approach.

Using this measure, we find that Twitter users are more interested in events related to entertainment and politics, and less interested in events related to energy or transportation. Also, we notice that Twitter users actively discuss dramatic crisis-related topics, including natural disasters (e.g., earthquakes, hurricanes, tornado, etc.) and terrorist attacks.

Table 4 shows the correlations between effectiveness and event popularity across different approaches. The correlations indicate a strong correlation with event popularity for the keyword approach. This is expected, since the approach is based on the number of times the keywords are mentioned within the timespan. The correlations are significantly reduced by incorporating query expansion terms. The configurations that use temporal query expansion tend to have lower correlation than the other approaches. Although the correlation is still significant, the lower correlation suggests that temporal query expansion approaches are more robust to popularity effects than simple keywords approaches. Additional work is necessary to better understand the role of popularity in retrieval tasks like this.

5 Conclusions

In this paper, we proposed a novel microblog search task called microblog event retrieval. Unlike previous microblog search tasks that retrieve individual microblog messages, our task involves the retrieval of structured event representations during which an event occurs and is discussed within the microblog community. In this way, users are presented with a ranked list or timeline of event instances in response to a query.

To tackle the microblog search task, we proposed a novel timespan retrieval framework that first constructs an expanded representation of the incoming query, performs timespan retrieval, and then produces a short summary of the timespan. Our experimental evaluation, carried out over a corpus of over 46 million microblog messages collected from Twitter, showed that microblog event retrieval is a feasible, challenging task, and that our proposed timespan retrieval framework is both robust and effective.

References

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic Detection and Tracking Pilot Study. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proc. 31st Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '08, pages 243–250, New York, NY, USA. ACM.

Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA. ACM.

Hai Leong Chieu and Yoong Keok Lee. 2004. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 425–432, New York, NY, USA. ACM.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In *1st Workshop on Social Media Analytics (SOMA'10)*, July.

Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. 2010. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 331–340, New York, NY, USA. ACM.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in twitter. In *1st Workshop on Social Media Analytics (SOMA'10)*, July.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

J. Lafferty and C. Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proc. 24th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 111–119.

V. Lavrenko and W. B. Croft. 2001. Relevance-based language models. In *Proc. 24th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 120–127.

Xiaoyan Li and W. Bruce Croft. 2003. Time-based language models. In *Proc. 12th Intl. Conf. on Information and Knowledge Management*, CIKM '03, pages 469–475, New York, NY, USA. ACM.

Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, and Jiawei Han. 2010. Pet: a statistical model for popular events tracking in social communities. In *Proc. 16th Ann. Intl. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, KDD '10, pages 929–938, New York, NY, USA. ACM.

Yuanhua Lv and ChengXiang Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In *Proc. 33rd Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '10, pages 579–586, New York, NY, USA. ACM.

Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. 2011. Incorporating query expansion and quality indicators in searching microblog posts. In *Proc. 33rd European Conf. on Information Retrieval*, page To appear.

Donald Metzler and W. Bruce Croft. 2007. Latent concept expansion using markov random fields. In *Proc. 30th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '07, pages 311–318, New York, NY, USA. ACM.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. 21st Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 275–281.

Jay Ponte. 1998. *A Language Modeling Approach to Information Retrieval*. Ph.D. thesis, University of Massachusetts, Amherst, MA.

J. J. Rocchio, 1971. *Relevance Feedback in Information Retrieval*, pages 313–323. Prentice-Hall.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, USA. ACM.

Jangwon Seo and W. Bruce Croft. 2010. Geometric representations for multiple documents. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 251–258, New York, NY, USA. ACM.

Beaux Sharifi, Mark-Anthony Hutton, and Jugal K. Kalita. 2010. Experiments in microblog summarization. *Social Computing / IEEE International Confer-*

ence on Privacy, Security, Risk and Trust, 2010 IEEE International Conference on, 0:49–56.

Jaime Teevan, Daniel Ramage, and Meredith Ringel Morris. 2011. #twittersearch: A comparison of microblog search and web search. In *WSDM 2011: Fourth International Conference on Web Search and Data Mining*, Feb.

ChengXiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proc. 10th Intl. Conf. on Information and Knowledge Management*, pages 403–410.

C. Zhai and J. Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214.