Identifying Content for Planned Events Across Social Media Sites

Hila Becker*†, Dan Iter†, Mor Naaman‡, Luis Gravano†
† Columbia University, 1214 Amsterdam Avenue, New York, NY 10027, USA
‡ Rutgers University, 4 Huntington St., New Brunswick, NJ 08901, USA

ABSTRACT

User-contributed Web data contains rich and diverse information about a variety of events in the physical world, such as shows, festivals, conferences and more. This information ranges from known event features (e.g., title, time, location) posted on event aggregation platforms (e.g., Last.fm events, EventBrite, Facebook events) to discussions and reactions related to events shared on different social media sites (e.g., Twitter, YouTube, Flickr). In this paper, we focus on the challenge of automatically identifying user-contributed content for events that are planned and, therefore, known in advance, across different social media sites. We mine event aggregation platforms to extract event features, which are often noisy or missing. We use these features to develop query formulation strategies for retrieving content associated with an event on different social media sites. Further, we explore ways in which event content identified on one social media site can be used to retrieve additional relevant event content on other social media sites. We apply our strategies to a large set of user-contributed events, and analyze their effectiveness in retrieving relevant event content from Twitter, YouTube, and Flickr.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Event Identification, Social Media, Cross-site Document Retrieval

1. INTRODUCTION

Event-based information sharing and seeking are common user interaction scenarios on the Web today. The bulk of information from events is contributed by individuals through social media channels: on photo and video-sharing sites (e.g., Flickr, YouTube), as well as on social networking sites (e.g., Facebook, Twitter). This event-related information can appear in many forms, including status updates in anticipation of an event, photos and videos captured before, during, and after the event, and messages containing post-event reflections. Importantly, for known and upcoming events (e.g., concerts, parades, conferences) revealing, structured information (e.g., title, description, time, location) is often explicitly available on user-contributed event aggregation platforms (e.g., Last.fm events, EventBrite, Facebook events). In this paper, we explore approaches for identifying diverse social media content for planned events.

Suppose a user is interested in the "Celebrate Brooklyn!" festival, an arts festival that happens in Brooklyn, New York every summer. This user could obtain information about the various music performances during this year's "Celebrate Brooklyn!" using Last.fm, a popular site that contains information about music events. Fortunately, Last.fm offers useful details about concerts at "Celebrate Brooklyn!," including the time/date, location, title, and description of these concerts. However, since Last.fm only provides basic event information, the user may consider exploring a variety of complementary social media sites (e.g., Twitter, YouTube) to augment this information at different points in time. For instance, before the event the user might be interested in reading Twitter messages, or *tweets*, describing ticket prices and promotions, while after the event the user might want to relive the experience by exploring YouTube videos recorded by attendees. By automatically associating social media content with planned events we can greatly enhance a user's event-based information seeking experience.

Automatically identifying social media content associated with known events is a challenging problem due to the heterogenous and noisy nature of the data. These properties of the data present a double challenge in our setting, where both the known event information and its associated social media content tend to exhibit missing or ambiguous information, and often include short, ungrammatical textual features. In our "Celebrate Brooklyn!" example, event features (e.g., title, description, location) are supplied by a Last.fm user; therefore, these features may consist of generic titles (e.g., "Opening Night Concert"), missing descriptions, or insufficient venue information (e.g., "Prospect Park," with no

[*] Contact author: Hila Becker, hila@cs.columbia.edu. This author is currently at Google Inc.

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.

exact address). Similarly, social media content associated with this event may be ambiguous (e.g., a YouTube video titled "Bird singing at the opening night gala") or not have a clear connection to the event (e.g., a tweet stating "#CB! starts next week, very excited!").

Existing approaches to find and organize social media content associated with known events are limited in the amount and types of event content that they can handle. Most related research relies on known event content in the form of manually selected terms (e.g., "earthquake," "shaking" for an earthquake) to describe the event [21, 24]. These terms are used to identify social media documents, with the assumption that documents containing these select terms will also contain information about the event. Unfortunately, manually selecting terms for any possible planned event is not a scalable approach. Improving on this point, a recent effort [7] used graphical models to label artist and venue terms in Twitter messages, identifying a set of related Twitter messages for concert events. While this work goes a step further in automating the process of associating events with social media documents, it is still tailored to a particular type of event (i.e., concerts) and restricted to a subset of the associated social media documents (i.e., documents containing venue and artist terms). Importantly, these related efforts focus on identifying site-specific event content, often tailoring their approaches to a particular site and its properties.

To address these limitations of the existing approaches, we leverage explicitly provided event features such as title (e.g., "Celebrate Brooklyn! Opening Gala"), description (e.g., "Singer/songwriter Andrew Bird will open the 2011 Celebrate Brooklyn! season"), time/date (e.g., June 10, 2011), location (e.g., Brooklyn, NY), and venue (e.g., "Prospect Park") to *automatically* formulate queries used to retrieve related social media content from *multiple* social media sites. Importantly, we propose a two-step query generation approach: the first step combines known event features into several queries aimed at retrieving *high-precision* results; the second step uses these high-precision results along with text processing techniques such as term extraction and frequency analysis to build additional queries, aimed at improving recall. We experiment with formulating queries for each social media site individually, and also explore ways to use retrieved content from one site to improve the retrieval process on another site. Our contributions are as follows:

- We pose the problem of identifying social media content for known event features as a query generation and retrieval task (Section 3).
- We develop precision-oriented query generation strategies using known event features (Section 4).
- We develop recall-oriented query generation strategies to improve the often low recall of the precision-oriented strategies (Section 5).
- We demonstrate how query generation strategies developed for one social media site can be used to inform the event content retrieval process on other social media sites (Section 6).

We evaluate our proposed query generation techniques on a set of known events from several sources and corresponding social media content from Twitter, Flickr, and YouTube (Section 7). Finally, we conclude with a discussion of our findings and directions for future work (Section 8).

2. RELATED WORK

We describe related work in three areas: quality content extraction in social media, event identification in textual news, and event identification in social media.

Research on extracting high-quality information from social media [1, 16] and on summarizing or otherwise presenting Twitter event content [11, 19, 23] has gathered recent attention. Agichtein et al. [1] examine properties of text and authors to find quality content in Yahoo! Answers, a related effort to ours but over fundamentally different data. In event content presentation, Diakopoulos et al. [11] and Shamma et al. [23] analyzed Twitter messages corresponding to large-scale media events to improve event reasoning, visualization, and analytics. Recently, we presented centrality-based approaches to extract high-quality, relevant, and useful Twitter messages from a given set of messages related to an event [6]. In this paper, we focus on identifying social media documents for known events, so the above approaches complement the work we present here, and can be used as a future extension to select among the social media documents that we collect for each event.

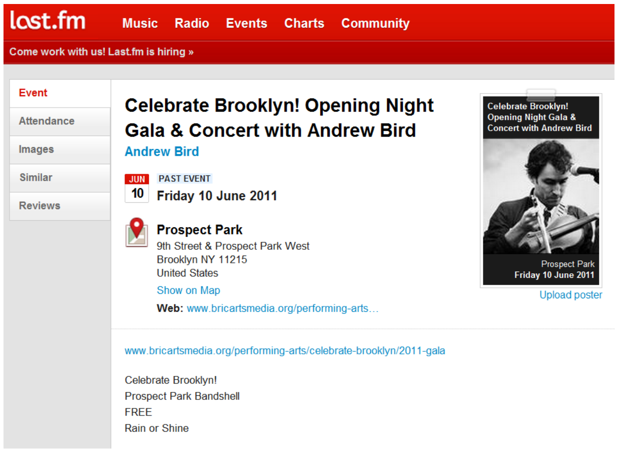
With an abundance of well-formed text, previous work on event identification in textual news (e.g., newswire, radio broadcast) [2, 13, 26] relied on natural language processing techniques to extract linguistically motivated features for identification of news events. Such techniques do not perform well over social media data, where textual content is often very short, and lacks reliable grammatical style and quality. More significantly, this line of research generally assumes that all documents contain event information. To identify events in social media, we have to consider and subsequently eliminate non-event documents when associating content with events.

While event detection in textual news documents has been studied in depth, the identification of events in social media sites is still in its infancy. Several related papers explored the idea of identifying *unknown* events in social media. We proposed an online clustering framework for identifying *unknown* events in Flickr [4]. As part of this framework, we explored the notion of multi-feature similarity for Flickr images and showed that combining a set of feature-driven similarity metrics yields better results for clustering social media documents according to events than using traditional text-based similarity metrics. Sankaranarayanan et al. [22] identified late breaking news events on Twitter using clustering, along with a text-based classifier and a set of news "seeders," which are handpicked users known for publishing news (e.g., news agency feeds). Petrović et al. [20] used locality-sensitive hashing to detect the first tweet associated with an event in a stream of Twitter messages. Finally, we used novel features to separate topically-similar message clusters into event and non-event clusters [5], thus identifying events and their associated social media documents on Twitter. In contrast with these efforts, we focus on identifying *known* events in social media, given a set of descriptive yet often noisy context features for an event.

Several recent efforts proposed techniques for identifying social media content for *known* events. Many of these techniques rely on a set of *manually* selected terms to retrieve event-related documents from a single social media site [21, 24]. Sakaki et al. [21] developed techniques for identifying earthquake events on Twitter by monitoring keyword triggers (e.g., "earthquake" or "shaking"). In their setting, the type of event must be known a priori, and should be easily represented using simple keyword queries. Most related to our work, Benson et al. [7] identified Twitter messages for concert events using statistical models to automatically tag artist and venue terms in Twitter messages. Their approach is novel and fully automatic, but it limits the set of identified messages for concert events to those with explicit artist and venue mentions. Our goal is to automatically retrieve social media documents for any known event, without any assumption about the textual content of the event or its associated documents. Importantly, all of these approaches are tailored to one specific social media site. In this paper we aim to retrieve social media documents across *multiple* sites with varying types of documents (e.g., photos, videos, textual messages).

3. MOTIVATION AND APPROACH

The problem that we address in this paper is how to identify social media documents across sites for a given planned event with known features (e.g., title, description, time/date, location). Records of planned events—including the event features on which we rely—abound on the Web, on platforms such as Last.fm events, EventBrite, and Facebook events. Figure 1 shows a snapshot of such a planned-event record on Last.fm.



Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird
Andrew Bird

JUN 10 PAST EVENT
Friday 10 June 2011

Prospect Park
9th Street & Prospect Park West
Brooklyn NY 11215
United States
Show on Map
Web: www.bricartsmedia.org/performing-arts...

www.bricartsmedia.org/performing-arts/celebrate-brooklyn/2011-gala

Celebrate Brooklyn!
Prospect Park Bandshell
FREE
Rain or Shine

Figure 1: A Last.fm event record for the "Celebrate Brooklyn!" opening night gala and concert.

We regard a social media document (e.g., a photo, a video, a tweet) as relevant to an event if it provides a reflection on the event before, during, or after the event occurs. Consider the "Celebrate Brooklyn!" opening gala concert example (see Figure 1). This event's related documents can reflect anticipation of the event (e.g., a tweet stating "I'm so excited for this year's Celebrate Brooklyn! and the FREE opening concert!"), participation in the event (e.g., a video of Andrew Bird singing at the opening gala), and post-event reflections (e.g., a photo of Prospect Park after the concert titled "Andrew Bird really knows how to put on a show"). All of these documents may be relevant to a user seeking information about this event at different times.

The definition of "event" has received attention across fields, from philosophy [12] to cognitive psychology [25]. In information retrieval, the concept of event has prominently been studied for event detection in news [2]. We borrow from this research to define an event in the context of our work. Specifically, we define an *event* as a real-world occurrence e with (1) an associated time period T_e and (2) a time-ordered stream of social media documents D_e discussing the occurrence and published during time T_e.

Operationally, an event is any record posted to one of the public event planning and aggregation platforms available on the Web (e.g., Last.fm events, EventBrite). Unfortunately, not all user-contributed records on these sites are complete and coherent, and while we expect our approaches to handle some missing data, a small subset of these records lack critical features that would make them difficult to interpret by our system and humans alike. Therefore, we do not include in our analysis records that are potentially noisy and incomplete. Specifically, we ignore:

- Records that are missing both start time/date and end time/date
- Records that do not have any location information
- Records with non-English title or description
- Records for "endogenous" events [8, 18] (i.e., events that do not correspond to any real-world occurrence, such as "profile picture change," a Facebook-specific phenomenon with no real-world counterpart)

Regardless of the platform on which they are posted, user-contributed event records generally share a core set of *context features* that describe the event along different dimensions. These features include (see Figure 1): title, with the name of the event (e.g., "Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird"); description, with a short paragraph outlining specific event details (e.g., "... Celebrate Brooklyn! Prospect Park Bandshell FREE Rain or Shine"); time/date, with the time and date of the event (e.g., Friday 10 June 2011); venue, with the site at which the event is held (e.g., Prospect Park); location, with the address of the event (e.g., Brooklyn, NY). These context features, collectively, can be helpful for constructing queries that can retrieve different types of social media documents associated with the event.

Problem Definition. *Consider any planned-event record posted on an event aggregation platform. Our goal is to retrieve relevant social media documents for this event on multiple social media sites, and identify the top-k such documents from each site, according to given site-specific scoring functions.*

We define the problem of associating social media documents with planned events as a query generation and retrieval task. Specifically, we design query generation strategies using the context features of events on the Web as defined above. For each event we generate a variety of queries, which we use *collectively* to retrieve matching social media documents from multiple sites. Since each event could potentially have many associated social media documents, we further filter the set of documents we present to a user to the top-k most similar documents, using given site-specific scoring functions (e.g., the multi-feature function in [4]). The similarity metrics that we use, and which are not the focus of this paper, might differ slightly across social media sites, since sites vary in their context features (e.g., documents from Flickr and YouTube have titles and descriptions whereas documents from Twitter do not).

Our approach for associating social media documents with planned events consists of two steps. First, we define precision-oriented queries for an event using its known context features (Section 4). These precision-oriented queries aim to collectively retrieve a set of social media documents with high-precision results. Then, to improve the (generally low) recall achieved in the first step, we use term extraction and frequency analysis techniques on the high-precision results to generate recall-oriented queries and retrieve additional documents for the event (Section 5). Figure 2 presents an overview of our query generation approach.

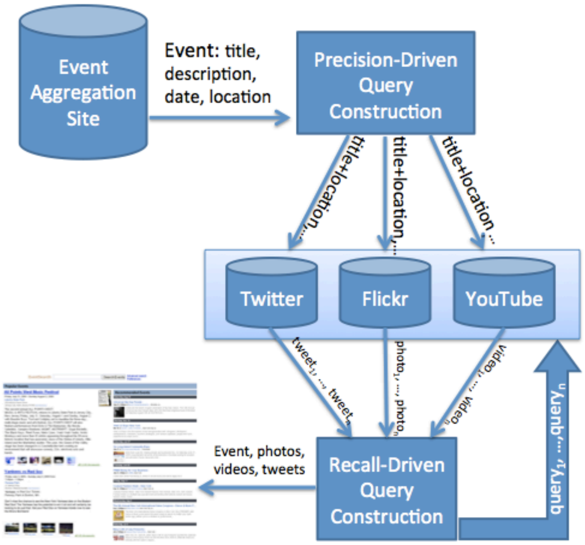


Figure 2: Our query-generation approach.

4. PRECISION-ORIENTED QUERY BUILDING STRATEGIES

Our first step towards retrieving social media documents for planned events consists of simple query generation strategies that are aimed at achieving high-precision results. These strategies form queries that touch on various aspects of an event (e.g., time/date and venue), following the intuition that these highly restrictive queries should only result in messages that relate to the intended event. We consider a variety of query generation strategies for this step, involving different combinations of the context features, namely, title, time/date, and location, of each event.

The precision-oriented queries for an event consist of combinations of one or more event features. One intuitive feature that we include in all strategies is a restriction on the time at which the retrieved social media documents are posted. In a study of trends on Twitter, Kwak et al. [15] discovered that most trends last for one week once they become "active" (i.e., once their associated Twitter messages are generated). Since our (planned) events can be anticipated, unlike the trends in [15], we follow a similar intuition and set the time period T_e that is associated with the event (see Section 3) to start a week prior to the event's start time/date and to end a week after the event's end time/date. For documents that contain digital media items (e.g., photos, videos), we only consider them if their associated media item was created during or after the event's start time. This step, while potentially eliminating a small number of relevant documents, is aimed at improving precision since we do not expect many digital media items associated with the event to be captured prior to the start of the event. We experimented with more restrictive time windows (e.g., one day after the event's end) but observed that relevant documents that contain digital media are generally posted within a week of the event, possibly due to a high barrier to post (e.g., having to upload photos from a camera that does not connect directly to the Internet).

In addition to restricting by time, we always include the title of the event in our precision-oriented strategies, as it often provides a precise notion of the subject of the event. As discussed in Section 3, title values exhibit substantial variations in specificity across event records. Some event titles might be too specific (e.g., "Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird"); for any such specific title, any social media documents matching it exactly will likely be relevant to the corresponding event. If the titles are too specific, however, no matching documents might be available, which motivates the recall-oriented techniques described in the next section. In contrast, other event titles might be too general (e.g., "Opening Night Concert"). To automatically accommodate these variations in title values, we consider different query generation options for the title feature. Specifically, we generate queries with the original title as a phrase, to capture content for events with detailed titles. We also generate queries with the original title as a phrase augmented with (portions of[1]) the event location, to capture content for events with broad titles, for which the location helps narrow down the matching documents. Finally, we consider alternative query generation techniques that include the title keywords as a list of terms—rather than as a phrase—for flexibility, as well as variations of the non-phrase version that eliminate stop words from the queries.

The intuition for the precision-oriented strategies we define is motivated by the informal results of these strategies over planned events from a pilot system. Our system [3] has a customizable interface that allows a user to select among different retrieval strategies. We selected precision-oriented strategies that include three variations of the title (i.e., phrase, list of terms, and list of terms with removed stop words), optionally augmented with either the city or venue portion of the location. We use these precision-oriented strategies to retrieve social media documents for a set of planned events, and verify that they indeed return high-precision results (Section 7). The final set of selected precision-oriented strategies is listed in Table 1.

5. RECALL-ORIENTED QUERY BUILDING STRATEGIES

While the strategies outlined in Section 4 often return high-precision social media documents for an event, the number of these high-precision documents is generally low. To improve recall, we develop several strategies for constructing queries using term-frequency analysis. Specifically, we treat an event's title, description, and any retrieved results from

[1] We observed that social media documents usually mention a single, broad aspect of the event's location, such as city or venue, rather than a full address.

Strategy	Example
[“title”+“city”]	[“Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird” “Brooklyn”]
[title+“city”]	[Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird “Brooklyn”]
[title-stopwords+“city”]	[Celebrate Brooklyn! Opening Night Gala Concert Andrew Bird “Brooklyn”]
[“title”+“venue”]	[“Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird” “Prospect Park”]
[title+“venue”]	[Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird “Prospect Park”]
[“title”]	[“Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird”]
[title]	[Celebrate Brooklyn! Opening Night Gala & Concert with Andrew Bird]
[title-stopwords]	[Celebrate Brooklyn! Opening Night Gala Concert Andrew Bird]

Table 1: Our selected precision-oriented strategies.

the precision-oriented techniques as “ground-truth” data for the event. We consider using the precision-oriented results from each social media site individually, and also from all social media sites collectively (Section 6).

Using the ground-truth data for each event, we design query formulation techniques to capture terms that uniquely identify each event. These terms should ideally appear in any social media document associated with the event but also be broad enough to match a larger set of documents than possible with the precision-oriented queries. We select these recall-oriented queries in two steps. First, we generate a large set of candidate queries for each event using two different *term analysis and extraction techniques*. Then, to select the most promising queries out of a potentially large set of candidates, we explore a variety of *query ranking strategies* and identify the top queries according to each strategy.

Frequency Analysis: The first query candidate generation technique aims to extract the most frequently used terms, while weighing down terms that are naturally common in the English language. The idea is based on the traditional term-frequency, inverse-document-frequency approach [17] commonly used in information retrieval. To select these terms, we compute term frequencies over the ground-truth data for word unigrams, bigrams, and trigrams. We then eliminate stop words and remove infrequent n-grams (determined automatically based on the size of the ground-truth corpus). We also eliminate any term that appears in the top 100,000 most frequent words indexed by Microsoft’s Bing search engine as of April 2010[2], with the assumption that any of these queries would be too general to describe any event.

To normalize the n-gram term frequency scores, we use a language model built from a large corpus of Web documents (see Section 7). With this language model, we compute log probability values for any candidate n-gram term. The probability of a term in the language model provides an indication of its frequency on the Web and is used to normalize the term’s computed frequency. We sort the n-grams extracted for each event according to their normalized term frequency values, and select the top 100 n-grams as candidate queries for the event.

Term Extraction: The second query candidate generation technique aims to identify meaningful event-related concepts in the ground-truth data using an external reference corpus. For this, we use a Web-based term extractor over our available textual event data [14]. This term extractor leverages a large collection of Web documents and query logs to construct an entity dictionary, and uses it along with statistical and linguistic analysis methodologies to find a list of significant terms. The extracted terms for each event serve as additional recall-oriented query candidates, along with the term-frequency query candidates described above.

Each of the techniques we describe could potentially generate a large set of candidate queries. However, many of these queries could be noisy (e.g., [@birdfan], with the name of a user that posts many updates about the event), too general (e.g., [concert tonight]), or describing a specific or non-central aspect of the event (e.g., [Fitz and the Dizzyspells], the name of an Andrew Bird song from the concert). Issuing hundreds of queries for each event is not scalable and could potentially introduce substantial noise, so we need to further reduce the set of queries to the most promising candidates. We explore a variety of strategies for selecting the top candidate queries out of all possible queries that we construct for each event. We consider two important criteria for ordering the event queries: specificity and temporal profile.

Specificity: Specificity assures that we rank long, detailed queries higher than broad, general ones. Since we use conjunctive query semantics, longer queries consisting of multiple terms (e.g., [a,b]), are more restrictive than shorter queries consisting of fewer terms (e.g., [a]). Particularly, since we use term n-gram shingles with n=1, 2, and 3 to construct the recall-oriented queries, our set of candidate queries often includes bigram queries that are subsets of trigram queries (e.g., [bird concert] and [andrew bird concert]). If both such candidates are present in the set, we favor the longer, more detailed version, as we observed that this level of specificity generally helps improve precision and yet is not restrictive enough to hurt recall.

Temporal Profile: The historical temporal profile of a query is another criterion we use to select among the candidate queries for an event. A local spike in document frequency around the time of the event might serve as an indication that the query is indeed associated with the event. We keep a record of the number of documents retrieved by each query during the week before and the week after the event, and compare this number to the query’s document volume during shorter time periods (one or two days) around the event’s time span. We used a similar signal successfully in our prior work [5] as an indicative feature for identifying events in textual streams of Twitter messages.

For example, Figure 3 shows a document volume histogram over Twitter documents for two recall-oriented queries retrieved around the week of Andrew Bird’s concert at “Celebrate Brooklyn!” We can see that the volume of a general query such as [state farm insurance] is consistent over time, whereas the volume of [andrew bird concert], while lower, increases around the time of the event. While this temporal analysis is promising for some social media sites

[2] http://web-ngram.research.microsoft.com/info/

(e.g., Twitter) where the time of the messages generally coincides with the time of the event, it may be problematic for other sites (e.g., YouTube, Flickr) that tend to exhibit a delay between an event's time and upload time of the associated digital media documents for the event, because of the nature of these sites. Therefore, for sites containing digital media, we use the content creation time rather than the upload time, if possible. (This feature is, unfortunately, often noisy or missing, especially for YouTube videos.)

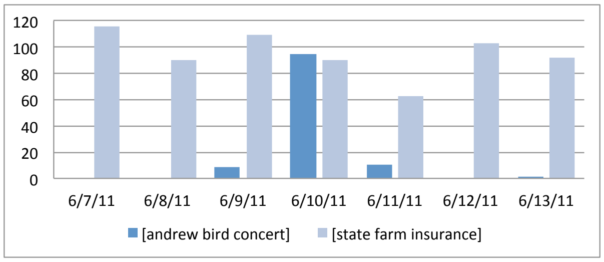


Figure 3: Histogram of Twitter document volume over time for two queries.

We consider using each of these query selection strategies individually, and also explore ways of combining them, to identify the top candidate queries for any given event. With these queries, we can retrieve associated social media documents from a variety of social media sites. Interestingly, content retrieved from the various sites can generate complementary signals for the recall-oriented query generation and retrieval process, as we will see next.

6. LEVERAGING CROSS-SITE CONTENT

Querying for event content from multiple social media sites can provide a holistic perspective on the event, complete with digital media and a variety of user perspectives. For the "Celebrate Brooklyn!" opening concert, for instance, we can use these different social media sites to learn about the event (e.g., via a Twitter message "Celebrate Brooklyn kicks off TONIGHT with Andrew Bird concert in Prospect Park!"), watch a video of a song performed at the event (e.g., "Andrew Bird - Effigy (Live) - Prospect Park - Brooklyn, NY" on YouTube), and see up-close photos of Andrew Bird on stage during the event (e.g., "Andrew Bird: Prospect Park Bandshell" photo set on Flickr).

Moreover, we can leverage event content from one social media site to help retrieve event documents from another social media site in different ways, following the query generation strategies proposed in the previous sections. One simple way is, of course, to generate recall-oriented queries for each site individually and use these queries across sites. Specifically, we can use the high-precision results obtained from an individual site to formulate recall-oriented queries as described in Section 5. We can then use these site-specific recall-oriented queries to obtain additional results from other social media sites. This is especially useful when the precision-oriented strategies do not retrieve results from all sites. This is the case for our "Celebrate Brooklyn!" example: since the event title is too specific, the precision-oriented queries fail to retrieve any documents from YouTube, and hence we cannot generate recall-oriented queries for the site. Fortunately, as is often the case, Twitter has a wealth of results for the precision-oriented queries for the event, and the resulting recall-oriented queries (e.g., [andrew bird concert], [brooklyn celebrate]) retrieve relevant videos from YouTube. In short, we manage to extract useful YouTube content through queries derived based on Twitter content.

An alternative way to leverage multi-site social media content is to generate recall-oriented queries using the high-precision results returned from all social media sites collectively. Whenever we obtain precision-oriented results from multiple sites, this approach yields a larger "ground-truth" corpus for the recall-oriented query generation than the ones obtained from each site individually, which may be helpful for identifying salient event terms that appear frequently across sites. At the same time, the results may be dominated by content from one site, possibly obscuring useful content from another site. This approach may also introduce noise or irrelevant content that is often present in some sites and not others (e.g., content-free titles of Flickr photos, Twitter username mentions).

Although content from different social media sites provides promising opportunities, it also presents challenges for our techniques. First, site-specific notations and conventions often introduce noise or inhibit recall. For example, Flickr users often tag photos with their camera settings (e.g., "canoneos5dmarkii"), which may be mistakenly identified as an important event term by the term frequency analysis, especially if the ground-truth corpus for the event is small. In addition, each Flickr tag must consist of a single term, so users often resort to very long multi-word tags (e.g., "greatcanadiancheesefestival"). In contrast, YouTube tags may each consist of several terms, so querying for such long multi-word tags on YouTube rarely yields results. We experimentally evaluate the merits of these alternative multi-site approaches in the next section.

7. EXPERIMENTS

We evaluated our query selection and retrieval techniques using a large dataset of real-world events from several event aggregation sites. For each event, we used our query generation strategies to collect related documents from popular social media sites. We performed three different sets of experiments:

- Comparison of the automatically generated queries against human-produced queries for the events
- Evaluation by human judges of the automatically generated queries
- Evaluation of the quality of the documents retrieved by the automatically generated queries

We report on the dataset and experimental settings, then turn to the results of our experiments.

7.1 Experimental Settings

Planned Event Dataset: We assembled a dataset of event records posted between May 13, 2011 and June 11, 2011 on four different event aggregation platforms: Last.fm events, EventBrite, LinkedIn events, and Facebook events. We used the Last.fm API with a location parameter set to "United States"[3] to collect musical performance events. To

[3] This was the only way to retrieve a set of events from Last.fm without issuing specific queries.

collect events from EventBrite, we used its API with the date parameters set to our specified date range. For LinkedIn events, where an API was not available, we retrieved and parsed event search pages in HTML format, using HTTP GET parameters to specify the date range.

Facebook events deserve special attention due to the difficulty of collecting such data via the site's API. Facebook events can only be retrieved in response to a specific search query or event id. To search for events, we used the most common event terms found in event titles collected by our event tracking system [3]. This list includes terms that describe specific types of events (e.g., [concert]) and also general terms commonly found in event titles (e.g., [national], [international]). We removed any returned event records that had no location or time information, and events that listed a virtual location (e.g., "everywhere") in their location or venue fields. Unfortunately, after filtering for these required fields we were left with very few events that matched our criteria. Still, we included these events in our experiments as they add diversity to our dataset.

To ensure that we collected events that would potentially have associated social media documents, we filtered out obscure events by requiring a minimum number of event attendees. We tuned this minimum threshold for each site given the observed distribution of attendees over all collected events. At the end of the process, we collected a total of 393 events, with 90 events from Last.fm, 94 events from EventBrite, 130 events from LinkedIn, and 25 events from Facebook. The above events constitute the test set over which we report our results. For training and tuning, we used a separate set of 329 event records, collected (in a similar fashion) between April 26 and May 11, 2011.

Social Media Documents: We collected social media documents for the events in our dataset from three social media sites: Twitter, YouTube, and Flickr. Specifically, we used each site's respective search API to issue precision-oriented (Section 4) and recall-oriented (Section 5) queries. From the retrieved results, we further eliminated any document that did not exactly match the search query since some site search engines (e.g., for Twitter) search for the query in any content that is linked from the document, and return matching documents as relevant results.

Note that part of our evaluation considers the quality of the top-k documents retrieved by the automatically generated queries (see problem definition in Section 3). The ranking of *documents* for an event is not the focus of this paper. For our evaluation, we rank the documents retrieved for an event by computing their similarity to the event record using (an adaptation of) the multiple-feature similarity function in [4]. As one additional component of the similarity, not present in [4], we consider the percentage of queries that retrieve a given document when we compute the score for the document and an event. Intuitively, we have observed that documents that are retrieved by several of our queries for an event should be preferred over documents that are retrieved by one such query.

Precision-Oriented Query Generation: For each event, we generated precision-oriented queries as defined in Section 4 using the event's context features, namely, title, time/date, city and venue. As an exception, we did not generate queries using the three title-only strategies for Last.fm events since we observed that many of the event titles on Last.fm consist of the name of a performer without any other context. Even though we restrict the social media documents that we retrieve to a specific time period around the event, it is often difficult in the Last.fm case to distinguish between two events held by the same performer in close time proximity. By forcing the location (i.e., city or venue) as part of the query for such events, we ensure that our precision-oriented queries produce results from the intended performance. For event records from the rest of the sites we use all precision-oriented queries from Section 4.

Recall-Oriented Query Generation: For each event, we generate recall-oriented queries as described in Section 5. To perform the frequency analysis, we index the documents using Lucene[4], with term n-grams, for n=1, 2, and 3. To normalize n-gram term frequency scores, we use the Microsoft Web n-gram Service[5], which provides n-gram log probability values. This service returns the joint probability of n-gram terms using a language model created from documents indexed by Microsoft's Bing search engine.

We extract meaningful queries from the high-precision results using the Yahoo! Term Extraction Web Service[6], which returns a list of significant terms or phrases given a segment of text. This term extractor leverages a large collection of documents and query logs to construct an entity dictionary and uses it along with a statistical and linguistic analysis [14] to process the given textual event data. This term extraction service has shown promising results on preliminary experiments with training data, to complement the first term frequency analysis technique above. It has also been successfully used in prior work for similar tasks [10, 14].

Query Generation and Ranking Techniques: Our experiments consider a subset of the (potentially many) queries generated using the precision- and recall-oriented strategies above. Different techniques will vary on how these subsets are selected. We consider two basic options to rank the queries for selection, namely, using (1) the "specificity" of the queries, as determined by the n-gram score on the Microsoft Web document corpus, or (2) variations of a "temporal" profile of the queries, determined by analyzing the volume of matching documents for the queries over time. Each alternative technique selects the top-10 queries according to the associated ranking criterion, as follows:

- MS n-gram Score (MS): n-gram score of the query from the Microsoft Web n-gram Service
- Time Ratio (TR): ratio of the number of documents created in the 48 hours before and after the event to the number of documents created in the week before and after the event
- Restricted Time Ratio (RTR): ratio of the number of documents created in the 24 hours before and after the event to the number of documents created in the week before and after the event
- MS n-gram Score and Time Ratio (MS-TR): MS score multiplied by TR score
- MS n-gram Score and Restricted Time Ratio (MS-RTR): MS score multiplied by RTR score

We apply the above techniques to documents from Twitter, YouTube, and Flickr individually, and also to documents

[4] http://lucene.apache.org/

[5] http://research.microsoft.com/web-ngram

[6] http://developer.yahoo.com/search/content/V1/termExtraction.html

from all three sites collectively. We use the site's name or "All," along with the strategy name (e.g., Twitter-MS, All-TR), to distinguish among these alternatives. We also compare the above techniques, which include both precision- and recall-oriented queries, against a technique that selects all precision-oriented queries. We refer to this technique as Precision.

Evaluation and Metrics: To evaluate our strategies, we collected annotations for a random sample of 60 events in our dataset. For each event, we used two annotators for three different tasks: comparison against human-produced queries, human evaluation of generated queries, and evaluation of document retrieval results. To compare our automatically generated queries against human-produced queries, we asked each annotator to provide 5 different queries that would be useful for collecting social media documents for each event. We use the Jaccard coefficient to measure the similarity of the set of automatically generated queries G to the set of human-produced queries H for each event. Specifically, for each query $q_g \in G$ and each query $q_h \in H$ we compute $J(q_g, q_h) = |q_g \cap q_h|/|q_g \cup q_h|$, with set operations performed over query terms. The Jaccard value that we report for G is then $\sum_{q_g \in G} max_{q_h \in H}(J(q_g, q_h))/|G|$. In other words, for each event, we computed the sum of maximum similarities between each automated query and the best-matching human-produced query.

For the human evaluation of the automatically generated queries, we asked two annotators to label 2,037 queries selected by our strategies for each event on a scale of 1-5, based on their relevance to the event. Here, we aim to gauge the potential of each query to retrieve results related to the event. For our "Celebrate Brooklyn!" example, the queries [celebrate], [celebrate brooklyn], and [andrew bird celebrate brooklyn] would receive scores of 1, 3, and 5, respectively. In cases of disagreement between annotators, we use the average rating. For two events in this set, our annotators were unable to provide queries due to ambiguous content (e.g., "ready film" as the title, without description), and content in a foreign language (e.g., queries in Italian for "FashionCamp," despite setting our API parameters for English-only content). These events were removed from the analysis.

Finally, for the evaluation of the quality of the documents retrieved by the automatically generated queries, we used Amazon's Mechanical Turk[7] to collect relevance judgments for the top-20 documents retrieved from Twitter, YouTube, and Flickr for each of our query selection techniques above. We collected two binary relevance judgments for each document. Agreement between our annotators was substantial, with Cohen's kappa coefficient values $\kappa = 0.85, 0.69$, and 0.93, for Twitter, YouTube, and Flickr documents, respectively. In cases of annotator disagreement, we collected a third judgment. To evaluate the retrieved documents, we use a standard metric, namely, normalized discounted cumulative gain, or *NDCG* [9], which captures the quality of ranked lists with focus on the top results. We use the binary version of NDCG [9], to measure how well our approach ranks the top documents relative to their ideal ranking.

7.2 Experimental Results

We begin by comparing the similarity of the automatically generated queries and human-produced queries for our events. Table 2 shows the results of our query generation methods using documents from Twitter, Flickr, and YouTube separately, and documents from all sites collectively. Across all strategies, queries generated using Flickr or YouTube documents were less similar to the human-produced queries compared to queries generated using Twitter documents. For Flickr, this result can be explained by the common use of long multi-word tags, which were often selected as the top queries by our strategies (e.g., [20110603musichallofwilliamsburgbrooklynny]). While these queries may not reflect human behavior, they could still be useful for retrieving event content, as we will see. In contrast, Precision had the highest Jaccard value at 0.705, indicating that the human-produced queries were most similar to the precision-oriented queries we defined in Section 4. Interestingly, using documents from all sites collectively did not improve the similarity, possibly due to the presence of Flickr tags among the selected queries for this strategy.

Strategy	Twitter	Flickr	YouTube	All
MS	0.571	0.216	0.181	0.272
TR	0.524	0.254	0.097	0.277
RTR	0.517	0.253	0.094	0.317
MS-TR	0.531	0.209	0.141	0.244
MS-RTR	0.523	0.209	0.141	0.263

Table 2: Jaccard coefficient for automatically generated queries and human-produced queries.

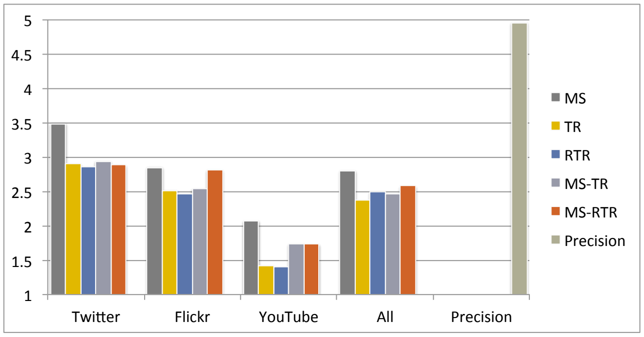


Figure 4: Average annotator rating of our automatically generated queries.

For the next step in our analysis, Figure 4 shows the average annotator rating for our alternative query generation approaches. Not surprisingly, Precision achieved the best average rating since, by design, it produced very detailed queries that are expected to return relevant results for their associated events. The query generation techniques that used Twitter documents, especially Twitter-MS, were again the most successful set of techniques. Based on our annotation guidelines, the score of Twitter-MS indicates that, on average, queries generated by this strategy are expected to retrieve some results for their associated event. The query generation techniques that used YouTube documents received the lowest scores in this evaluation. One possible explanation is that the query-generation strategies may not be effective when formulated using YouTube data alone, which may be related to the lack of reliable temporal information for YouTube documents, as we discussed in Section 5.

For our third set of experiments, we examined the rele-

[7] https://www.mturk.com

Strategy	5 Docs	10 Docs	15 Docs	20 Docs
Twitter-MS	0.759	0.724	0.690	0.690
Twitter-RTR	0.828	0.793	0.759	0.759
Precision	0.414	0.293	0.241	0.224

Table 3: Percentage of events with Twitter results at different recall levels for alternative query strategies.

vance of documents retrieved by our query generation strategies to their associated events. Figure 5 shows the NDCG scores for the top $5, 10, 15$, and 20 Twitter documents retrieved by Precision, Twitter-MS, and Twitter-RTR (Twitter-TR, Twitter-MS-TR, and Twitter-MS-RTR produced similar results to Twitter-MS and Twitter-RTR, and were therefore omitted). Validating our earlier observation (Section 4), Precision retrieved highly relevant results. Both Twitter-MS and Twitter-RTR also produced good results, demonstrating their effectiveness at retrieving Twitter documents for known events.

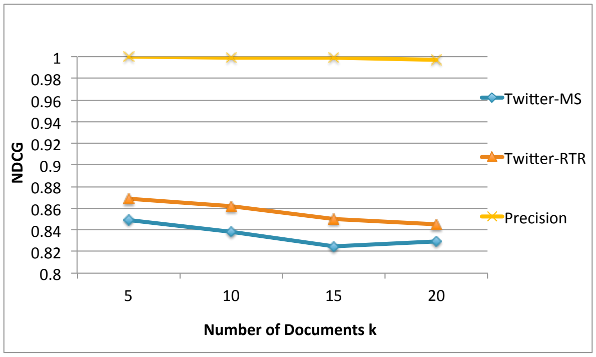


Figure 5: NDCG scores for top-k Twitter documents retrieved by our query strategies.

Note that the NDCG scores at each level of recall were averaged over the set of events that had some returned results for each strategy. Table 3 reports the percentage of events in our dataset for which each strategy returned results at various levels of recall. As expected, Precision returned results for a small fraction of the events. Interestingly, Twitter-RTR returned results for a larger proportion of the events than Twitter-MS. This can be explained by the way these alternative strategies select their top queries. Specifically, all queries selected for Twitter-RTR must have some matching documents, since we consider each query's document volume over time as the selection criterion. In contrast, Twitter-MS is biased towards rare terms (i.e., terms with lower probability scores), making it the second most precise among the strategies, following Precision.

Our next set of results examines the effectiveness of our approaches for retrieving event documents across social media sites. Given our observations from the query-based evaluations, we evaluated the relevance of documents retrieved by the best performing query generation approach, namely, Twitter-MS, from both YouTube and Flickr. Figure 6 shows the NDCG scores of Precision, Twitter-MS, and YouTube-MS for the top-k YouTube documents, averaged over all events. In addition, the size of each point reflects the number of events that had at least k documents retrieved by the strategy. As we can see, Twitter-MS performed better and retrieved results for more events than YouTube-MS, indicating that Twitter documents can be potentially used to improve both precision and recall of YouTube documents for planned events.

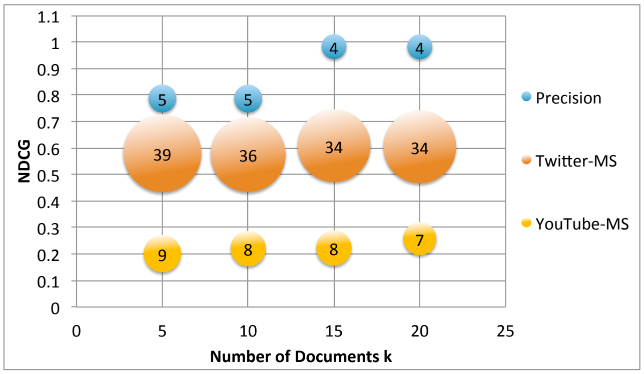


Figure 6: NDCG scores for top-k YouTube documents retrieved by our query strategies.

We performed a similar evaluation over documents from Flickr, using Precision, Twitter-MS, and Flickr-MS. Precision, expectedly, retrieved relevant results for a small number of events. Interestingly, unlike YouTube-MS, Flickr-MS achieved higher NDCG scores than Twitter-MS. However, the number of events covered by Flickr-MS is smaller than the number of events covered by Twitter-MS, showing that Twitter-MS can still retrieve relevant Flickr documents and can be particularly useful in cases where Flickr-MS returns no results.

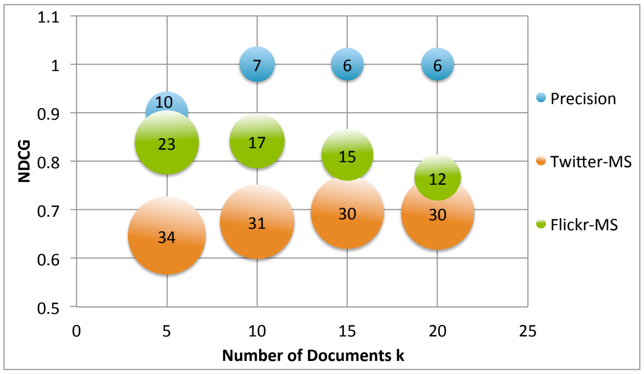


Figure 7: NDCG scores for top-k Flickr documents retrieved by our query strategies.

Overall, our evaluation showed that our query generation approaches can effectively retrieve relevant social media documents for planned events on multiple social media sites. In addition, we demonstrated that we can leverage social media documents on Twitter to generate a query strategy (i.e., Twitter-MS) that can retrieve relevant event documents on YouTube and Flickr.

8. CONCLUSIONS

In this paper, we presented a query-oriented solution for retrieving social media documents for planned events across different social media sites. This work provides an essential step in the process of organizing social media documents for events, towards improved browsing and search for event media. Using a combination of precision-oriented and recall-oriented query generation techniques, we showed how to automatically and effectively associate social media documents with planned events from various sources. Importantly, we demonstrated how social media documents from one social media site can be used to enhance document retrieval on another social media site, thus contributing to the diversity of information that we can collect for planned events. Many opportunities for future work remain to improve our general approach, such as learning how to exploit the content of pages linked from event-related documents and how to leverage direct links between such documents across social media sites, in an iterative cross-site retrieval process. Other future directions include sub-event content and topic analysis, so that multiple views or temporal variations represented in the data can be exposed. Overall, our techniques help unveil important information from, and about, planned events as they are reflected through the eyes of hundreds of millions of users of social media sites.

9. ACKNOWLEDGMENTS

This material is based on work supported by NSF Grants IIS-0811038, IIS-1017845, and IIS-1017389, and by two Google Research Awards. In accordance with Columbia University reporting requirements, Professor Gravano acknowledges ownership of Google stock as of the writing of this paper. We thank Feiyang Chen for his invaluable help with the precision-oriented strategies of Section 4, and our annotators Brandy Hartley and Emma Moore.

10. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the First ACM International Conference on Web Search and Data Mining (WSDM '08)*, 2008.

[2] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization.* Kluwer Academic Publisher, 2002.

[3] H. Becker, F. Chen, D. Iter, M. Naaman, and L. Gravano. Automatic identification and presentation of Twitter content for planned events. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, 2011.

[4] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)*, 2010.

[5] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, 2011.

[6] H. Becker, M. Naaman, and L. Gravano. Selecting quality Twitter content for events. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, 2011.

[7] E. Benson, A. Haghighi, and R. Barzilay. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11)*, 2011.

[8] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[9] W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice.* Addison-Wesley Publishing Company, 2009.

[10] W. Dakka and P. G. Ipeirotis. Automatic extraction of useful facet hierarchies from text databases. In *Proceedings of the IEEE 24th International Conference on Data Engineering (ICDE '08)*, 2008.

[11] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST '10)*, 2010.

[12] Events, 2002. In Stanford Encyclopedia of Philosophy. Retrieved June 2nd, 2010 from http://plato.stanford.edu/entries/events/.

[13] V. Hatzivassiloglou, L. Gravano, and A. Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd ACM International Conference on Research and Development in Information Retrieval (SIGIR '00)*, 2000.

[14] R. Kraft, F. Maghoul, and C. C. Chang. Y!Q: Contextual search at the point of inspiration. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*, 2005.

[15] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, 2010.

[16] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang. Web video topic discovery and tracking via bipartite graph reinforcement model. In *Proceedings of the 17th International World Wide Web Conference (WWW '08)*, 2008.

[17] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge Univ. Press, 2008.

[18] M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

[19] M. Nagarajan, K. Gomadam, A. P. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *Proceedings of the 10th International Conference on Web Information Systems Engineering (WISE '09)*, 2009.

[20] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '10)*, 2010.

[21] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, 2010.

[22] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: News in tweets. In *Proceedings of the 17th ACM International Conference on Advances in Geographic Information Systems (GIS '09)*, 2009.

[23] D. A. Shamma, L. Kennedy, and E. Churchill. Statler: Summarizing media through short-message services. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*, 2010.

[24] S. Yardi and d. boyd. Tweeting from the town square: Measuring geographic local networks. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM '10)*, 2010.

[25] J. M. Zacks and B. Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127, 2001.

[26] K. Zhang, J. Zi, and L. G. Wu. New event detection based on indexing-tree and named entity. In *Proceedings of the 30th ACM International Conference on Research and Development in Information Retrieval (SIGIR '07)*, 2007.