Selecting Quality Twitter Content for Events

Hila Becker
Columbia University
hila@cs.columbia.edu

Mor Naaman
Rutgers University
mor@rutgers.edu

Luis Gravano
Columbia University
gravano@cs.columbia.edu

Abstract

Social media sites such as Twitter contain large amounts of user contributed messages for a wide variety of real-world events. While some of these "event messages" might contain interesting and useful information (e.g., event time, location, participants, opinions), others might provide little value (e.g., using heavy slang, incomprehensible language) to people interested in learning about an event. Techniques for effective selection of quality event content may therefore help improve applications such as event browsing and search. In this paper, we explore approaches for finding representative messages among a set of Twitter messages that correspond to the same event, with the goal of identifying high quality, relevant messages that provide useful event information. We evaluate our approaches using a large-scale dataset of Twitter messages, and show that we can automatically select event messages that are both relevant and useful.

1 Introduction

Real-world events often have vast amounts of associated social media content. For example, a search for [Obama inauguration] on YouTube returns over 30,000 videos as of January 2011. A 2010 live broadcast of a U2 concert on YouTube drew over 130,000 posts on Twitter. Even smaller events often feature dozens to hundreds of different content items. At the same time, important applications such as event browsing and search could greatly benefit from social media event content, but need to select and prioritize such content to avoid overwhelming their users with too much information. In this paper, we address the problem of selecting quality Twitter content for an event.

Selecting the most salient social media content for an event is a challenging task, due to the heterogeneity and scale of the data. As one challenge, seemingly related content with good textual quality might not be truly relevant to the event (e.g., "Bill cares about his health" for the United States health care reform bill passage). As another challenge, relevant, high-quality content might not be useful (e.g., "I can't stop thinking about the health care reform bill passage") as it does not provide much information about the event in question. This work examines several approaches for finding high-quality content that is relevant and contains useful information for each event.

Past research has focused on extracting high-quality information from social media (Agichtein et al. 2008; Liu et al. 2008) and summarizing or otherwise presenting Twitter event content (Diakopoulos, Naaman, and Kivran-Swaine 2010; Nagarajan et al. 2009; Shamma, Kennedy, and Churchill 2010). Agichtein et al. (2008) 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, Naaman, and Kivran-Swaine (2010), and Shamma, Kennedy, and Churchill (2010) analyzed Twitter messages corresponding to large-scale media events to improve event reasoning, visualization, and analytics. This research considered measures of relevance and importance to surface content representations such as "top keywords" that are orthogonal to our content selection methods.

In this paper, we focus on the problem of selecting Twitter content for events. We address this problem with two concrete steps. First, we identify each event—and its associated Twitter messages—using an online clustering technique that groups together topically similar Twitter messages (Section 2). Second, for each identified event cluster, we select messages that best represent the event (Section 3). We use centrality-based techniques to select messages that have high textual quality, strong relevance to the event, and, importantly, are useful to people looking for information about the event. We evaluate our proposed content selection techniques using a large-scale dataset of Twitter messages (Section 4) and finally discuss the implications of our findings as well as future work (Section 5).

2 Identifying Event Content

We propose to associate Twitter messages with events using an online clustering framework. Specifically, we elected to use an incremental, online clustering algorithm to effectively cluster a stream of Twitter messages in a scalable fashion, without requiring *a priori* knowledge of the number of clusters. These features of the clustering algorithm are particularly desirable for this domain since Twitter messages are constantly produced and new events are added to the stream over time. We followed the implementation of the algorithm as described in (Becker, Naaman, and Gravano 2010), and used annotators to identify event clusters (see Section 4.1).

Alternative content identification approaches such as topic modeling (Ramage, Dumais, and Liebling 2010) are not well suited for this task as they require a set number of topics to be identified, and often yield coarse, general topics that do not correspond to events. Twitter's own "trending topics" algorithm and similar keyword-based trend detection approaches often result in multiple topics that refer to the same event (e.g., "#egypt," "Cairo," and "Mubarak"), requiring a post processing step to unify these topics and their associated messages. In future work, we plan to experiment with alternative approaches for associating messages with events. However, in this paper we chose to proceed with a clustering approach, which was sufficient for producing a representative set of event clusters on which we apply our content selection techniques, as we describe next.

3 Event Content Selection

Once we have identified events and their associated Twitter messages, we address the problem of selecting a subset of these messages for presentation. We describe our content selection goals and approaches next.

3.1 Content Selection Goals

We select messages for each identified event with three desired attributes: *quality*, *relevance*, and *usefulness*. *Quality* refers to the textual quality of the messages, which reflects how well they can be understood by a human. As previously discussed, the quality of messages on Twitter varies widely. High-quality messages contain crisp, clear, and effective text that is easy to understand (e.g., "The Superbowl is playing on channel 4 right now"). Low-quality messages, on the other hand, contain incomprehensible text, heavy use of short-hand notation, spelling and grammatical errors, and typos (e.g., "obv maj #fail lol"). Interestingly, the quality of a message is largely independent of its associated event.

Relevance in our context refers to how well a Twitter message reflects information related to its associated event. Highly relevant messages clearly refer to or describe their associated event (e.g., "The steelers' touchdown was amazing - I wish they'd show it again"). Messages are not relevant to an event if they do not refer to the event in any way (e.g., "good morning, what are people doing today?"). In between these two extremes are messages that are somewhat relevant to an event, where the event is not the main subject (e.g., "I can't believe I'm stuck at work, I'd rather be watching the superbowl") or messages that are barely relevant and only obscurely refer to the event (e.g., "this game is so boring, but watching the commercials is mildly entertaining").

Usefulness refers to the potential value of a Twitter message for someone who is interested in learning details about an event. Useful messages should provide some insight about the event, beyond simply stating that the event occurred. The level of usefulness of Twitter messages varies. Messages that are clearly useful provide potentially interesting details about the event (e.g., "The Packers and Steelers are playing in this year's Superbowl"). Messages that are clearly not useful provide no context or information about the event (e.g., "super bowl!!! that's all folks"). Other messages may reflect a user's opinion about the event, where somewhat useful event information is directly stated or can be inferred (e.g., "It's the best superbowl game ever").

We use these three attributes as absolute measures of user satisfaction with the selected event content, and as relative measures of the success of our alternative content selection approaches, which we describe next.

3.2 Content Selection Approaches

With our content selection goals in mind, we now propose alternative approaches for selecting a subset of Twitter messages associated with a given event. These approaches rely on the observation that the most topically central messages in a cluster are likely to reflect key aspects of the event better than other, less central cluster messages. This notion of centrality can be defined in a variety of ways:

Centroid: The centroid similarity approach computes the cosine similarity of the *tf-idf* representation (as defined by Kumaran and Allan (2004)) of each message to its associated event cluster *centroid*, where each cluster term is associated with its average weight across all cluster messages. It then selects the messages with the highest similarity value. Since a cluster's centroid highlights important terms used to describe the event (e.g., for Tiger Woods' famous apology speech, centroid terms with high weight might include "tiger," "woods," "apology," and "elin"), messages with high similarity to these key terms are likely to reflect key aspects of the event, as desired by the relevance and usefulness goals. In addition, since centroid term weights are based on frequency across all messages, they tend to be high for quality terms (e.g., without typos or spelling errors), addressing our quality selection goal.

Degree: An alternative view of centrality involves message similarity across all messages in an event cluster. In this alternative approach, we represent each cluster message as a node in a graph, and any pair of nodes whose cosine similarity exceeds a predetermined threshold is connected by an edge. Using this graph formulation, the degree method selects nodes with the highest degree centrality, defined as the degree of each node, weighted by the number of nodes in the graph. Using degree centrality enables us to select messages that contain important terms that may not have been captured by the centroid due to low support in the cluster messages (e.g., a small but highly connected subset of messages might also include the word "mistress" when discussing the Tiger Woods apology). In this method, highly connected messages are also likely to include key event terms, a desirable property for content selection.

The degree centrality method treats each edge as an equal vote for its adjacent nodes' centrality. However, it is often beneficial to associate a weight with each edge, based on the similarity value of the nodes it connects. In fact, this idea has been considered for the task of extractive summarization (Erkan and Radev 2004), a related task where sentences from multiple documents are selected to form a summary. Our third approach, *LexRank*, is based on a state-of-the-art technique by the same name used to select document sentences for summarization (Erkan and Radev 2004).

LexRank: The LexRank approach (Erkan and Radev 2004) defines centrality based on the idea that central nodes are connected to other central nodes. In other words, every node has a centrality value, which it distributes to its neighbors. This idea can be represented using the formula $p(m) = \sum_{n \in adj[m]} (p(n)/deg(n))$, where $p(n)$ is the centrality of node n, $adj[m]$ is the set of nodes adjacent to node m, and $deg(n)$ is the degree of node n. The value of $p(m)$ for each cluster message can be computed using the power method (Erkan and Radev 2004), which estimates the stationary probability distribution resulting from a random walk on the message graph. We select the top messages in the cluster according to their LexRank value.

In addition to these centrality-based approaches, we considered baseline content selection techniques such as selecting the most recent messages added to a cluster or selecting messages from popular users (i.e., users with many followers). Unfortunately, when used in isolation, these techniques suffer from serious drawbacks (e.g., inability to reduce selection of noisy, irrelevant content) so we eliminated them from consideration after running experiments on training data. These potentially useful signals could instead be incorporated with our centrality based approaches in a disciplined way (e.g., using a trained ranking function), a task that we reserve for future work.

4 Experiments

We evaluated our content selection strategies on a large dataset of Twitter data. We describe this dataset and report the experimental settings (Section 4.1), and then turn to the results of our experiments (Section 4.2).

4.1 Experimental Settings

Data: We used the Twitter API to collect over 2,600,000 Twitter messages, or *tweets*, posted during February 2010 by New York City users (i.e., by Twitter users whose location, as entered by the users, is in the New York City area). This dataset was collected as part of a larger initiative for identifying and characterizing event content and is location-centric for this reason. However, we believe that this characteristic of the data does not introduce any bias in our evaluation since our techniques currently do not consider the tweets' location in the selection process.

We cluster our entire dataset in an online fashion as described in Section 2. We used the data from the first week in February to calibrate the parameters of the clustering algorithm, and then used the second week of February for the development of our centrality-based approaches (and to rule out poorly performing alternatives such as time-based selection). Finally, we report our results on test data selected from the latter half of February (i.e., Weeks 3 and 4).

Annotations: To test the content selection approaches, we selected 50 event clusters, with an average of 412 messages per cluster, from our test set (the presence of event content in the cluster was determined by two annotators, with substantial agreement, with Cohen's kappa coefficient κ=0.79). For each event cluster we selected the top-5 messages according to each content selection approach. We used two annotators to label each message according to our desired attributes: quality, relevance, and usefulness. The annotators labeled each message on a scale of 1-4 for each attribute, where a score of 4 signifies high quality, strong relevance, and clear usefulness, and a score of 1 signifies low quality, no relevance, and no usefulness. Agreement between annotators on low (1, 2) and high (3, 4) ratings for each attribute was substantial to high, with kappa coefficient values $\kappa = 0.92, 0.89, 0.61$ for quality, relevance, and usefulness, respectively. In our evaluation, we use the average score for each message to compare the algorithmic results.

Techniques for comparison: We evaluate and compare our three content selection approaches, namely, *Centroid*, *Degree*, and *LexRank*. To compute the degree centrality, we set the similarity threshold for connecting two message nodes to 0.05. For the *LexRank* approach we used the Mead toolkit (Erkan and Radev 2004) with the LexRank feature option, which produces a ranked list of messages according to their LexRank score.

4.2 Experimental Results

We evaluated our three competing approaches according to user-perceived quality, relevance, and usefulness with respect to a specific event. Figure 1 summarizes the average performance of these approaches across all 50 test events. All three approaches received high scores for quality (where a score of 4 implies excellent quality). *Degree* and *Centroid*, on average, selected messages that are either somewhat relevant or highly relevant. However, *Centroid* is the only approach that received a high score for usefulness, indicating that, on average, its selected messages were either somewhat or clearly useful with respect to the associated events.

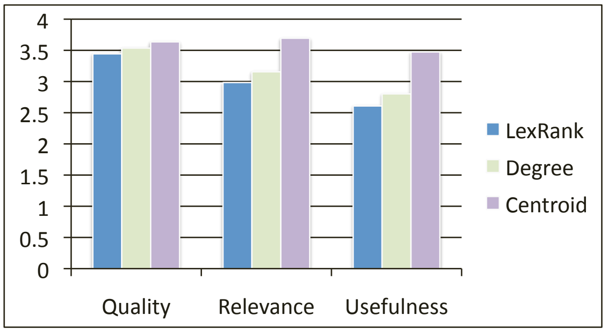


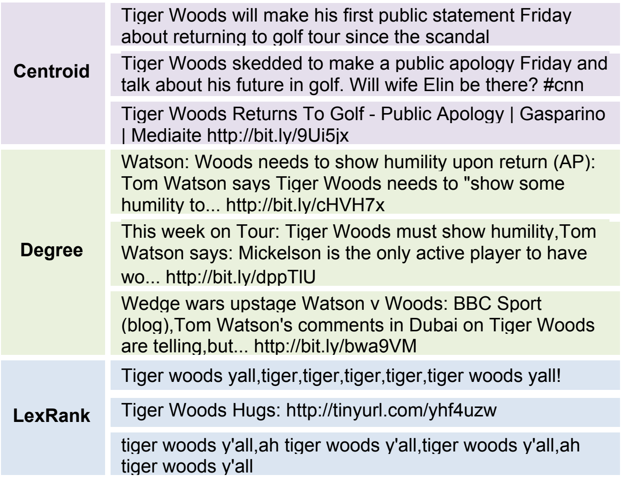
Figure 1: Comparison of content selection techniques.

To test for significant differences between the approaches, we also compared them against each other in terms of the number of events that each approach was preferred for. Table 1 shows the average rank of each approach according to the three desired attributes. We performed a statistical significance analysis based on these ranked preferences using the Friedman test (Demšar 2006), a non-parametric statistical test for comparing a set of alternative models. According to this test, there are significant differences between the approaches ($p < 0.01$) in terms of relevance and usefulness. Post-hoc analysis of our data using the Nemenyi

Method	Quality	Relevance	Usefulness
LexRank	2.23	2.4	2.4
Degree	2.02	2.09	2.08
Centroid	**1.75**	**1.51**	**1.52**

Table 1: Preference rank of content selection approaches, averaged over 50 test events.

test (Demšar 2006) determined that *Centroid* is significantly better than the other approaches in terms of both relevance and usefulness. Significant differences between *Degree* and *LexRank* could not be determined. Additionally, we could not reject the null hypothesis of the Friedman test (i.e., that all approaches have similar performance) in terms of quality.



Centroid	Tiger Woods will make his first public statement Friday about returning to golf tour since the scandal
	Tiger Woods skedded to make a public apology Friday and talk about his future in golf. Will wife Elin be there? #cnn
	Tiger Woods Returns To Golf - Public Apology | Gasparino | Mediaite http://bit.ly/9Ui5jx
Degree	Watson: Woods needs to show humility upon return (AP): Tom Watson says Tiger Woods needs to "show some humility to... http://bit.ly/cHVH7x
	This week on Tour: Tiger Woods must show humility,Tom Watson says: Mickelson is the only active player to have wo... http://bit.ly/dppTlU
	Wedge wars upstage Watson v Woods: BBC Sport (blog),Tom Watson's comments in Dubai on Tiger Woods are telling,but... http://bit.ly/bwa9VM
LexRank	Tiger woods yall,tiger,tiger,tiger,tiger,tiger woods yall!
	Tiger Woods Hugs: http://tinyurl.com/yhf4uzw
	tiger woods y'all,ah tiger woods y'all,tiger woods y'all,ah tiger woods y'all

Figure 2: Sample tweets selected by the different approaches for the "Tiger Woods Apology" event.

5 Discussion

A single event might sometimes attract hundreds or thousands of social media content items, so being able to rank and filter event content is a requirement for a variety of applications that aim to communicate that content effectively. In this paper, we presented Twitter content selection approaches that form a promising initial step towards this goal.

Among three centrality-based approaches, *Centroid* emerged as the preferred way to select tweets given a cluster of messages related to an event. Based on our observation of the data, we believe that the success of this method is related to its inherent assumption that each cluster revolves around one central topic. *LexRank* and *Degree*, on the other hand, tend to select messages that are strongly similar to one another, but may sometimes diverge from the main topic of the cluster (e.g., see Tom Watson's comments on Tiger Woods, selected by *Degree*, in Figure 2).

In addition to the centrality-based approaches described in this paper, we developed a variety of re-ranking techniques that boost the centrality score of tweets with potentially useful features (e.g., URLs, tags). Users can manually adjust these techniques based on their preferences. A preliminary exploration of these re-ranking techniques revealed a disagreement among users on what aspects of a tweet (beyond quality, relevance, and usefulness, as defined in our annotation guidelines) are desirable. Some users tend to prefer tweets with URLs, due to the promise of additional, potentially interesting information, while others see more value in verbose tweets, with self-contained information related to the event. We plan to explore this further in future work.

There are a number of additional interesting directions for future work on content selection. For example, a system could consider the social network of authors posting about an event, in addition to the network of messages connected by similarity. Centrality of authors and content can then be addressed in concert. Other future directions include subevent content and topic analysis, such that multiple views or temporal variations represented in the data can be exposed.

6 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 Univ. reporting requirements, Prof. Gravano acknowledges ownership of Google stock as of the writing of this paper.

References

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *WSDM'08*.

Becker, H.; Naaman, M.; and Gravano, L. 2010. Learning similarity metrics for event identification in social media. In *WSDM'10*.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *JMLR* 7:1–30.

Diakopoulos, N.; Naaman, M.; and Kivran-Swaine, F. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*.

Erkan, G., and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *JAIR* 22(1):457–479.

Kumaran, G., and Allan, J. 2004. Text classification and named entities for new event detection. In *SIGIR'04*.

Liu, L.; Sun, L.; Rui, Y.; Shi, Y.; and Yang, S. 2008. Web video topic discovery and tracking via bipartite graph reinforcement model. In *WWW'08*.

Nagarajan, M.; Gomadam, K.; Sheth, A. P.; Ranabahu, A.; Mutharaju, R.; and Jadhav, A. 2009. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *WISE'09*, 539–553.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *ICWSM'10*.

Shamma, D. A.; Kennedy, L.; and Churchill, E. 2010. Statler: Summarizing media through short-message services. In *CSCW'10*.