An Empirical Assessment of Contemporary Online Media in Ad-Hoc Corpus Creation for Social Events

Kanika Narang, Seema Nagar, Sameep Mehta, L V Subramaniam, Kuntal Dey
IBM Research Labs, India
{kaninara,senagar3,sameepmehta,lvsubram,kuntadey}@in.ibm.com

Abstract

Social networking sites such as Facebook and Twitter have become favorite portals for users to discuss and express opinions. Research shows that topical discussions around events tend to evolve socially on microblogs. However, sources like Twitter have no explicit discussion thread which will link semantically similar posts. Moreover, the discussion may be evolving in multiple different threads (like Facebook). Researchers have proposed the use of online contemporary documents to act as external corpus to connect pairs of contextually related semantic topics. This motivates the question: *given a significant social event, what is a good choice of external corpus to identify evolution of discussion topics around the event's context?* In this work, we compare the effectiveness of contemporary blog posts, online news media and forum discussions in creating ad-hoc external corpus. Using social propensity of evolution of topical discussions on Twitter to assess the goodness of the creation, we find online news media as most effective. We evaluate on three large real-life Twitter datasets to affirm our findings.

1 Introduction

Social media has become a hotbed of user generated content. Multiple online platforms have emerged for users to participate, interact and discuss. Social contact and activity networks like Facebook, video sharing networks like Youtube, photo/image sharing platforms like Pinterest, social bookmarking platforms like Digg, and microblogging platforms like Twitter have become prominent online social media platforms.

Research suggests that social microblogging platforms, like Twitter, diffuse information in a manner similar to news media (Kwak and Lee, 2010). In a world of millions of people (Twitter subscribers) with inherent entropy, in absence of explicit discussion threads (unlike online forums, for example), conversations around any event are expected to move towards different directions over time. Contradictory to this apparent expectation, research suggests that these discussions tend to temporally grow and evolve along social relationships of people engaging in these discussions, much more strongly, compared to random evolution (Narang and Nagar, 2013).

Interestingly, trending events on unstructured microblogs often get built around non-traditional, temporary and contemporary factors, entities and relationships. Because of the sheer number of diverse contemporary events, event types and associated documents, it is impossible to prepare well-defined, validated and clean corpus for each and every event. For instance, political turmoils have existed for ages; however, one may not expect a dedicated corpus to preexist for the Libya 2011 turmoil associating its places and locations, contemporary leaders, and all the other global political factors. Hence, there is a strong need of using contemporary online media for ad-hoc corpus creation in such a setting.

The use of external corpus has been shown to improve performance in language tasks such as question-answering, machine translation, and information retrieval (Kilgarriff, 2003; Clarke and Cormack, 2002; Dumaisl and Banko, 2002; Metzler and Diaz, 2005; Xu and Croft, 2000; Diaz and Metzler, 2006). But questions relating to the rele-

International Joint Conference on Natural Language Processing, pages 1375–1382,
Nagoya, Japan, 14-18 October 2013.

vance of the corpus have not been studied as much.

The finding of (Narang and Nagar, 2013) that topical discussions on microblogs tend to evolve socially is interesting. However, it simply uses contemporary online news media as the only source of external corpus, to establish *extended semantic relationships* across topic clusters. It does not attempt to use any other source of relevant semantic data for creating external corpus; nor does it assess the goodness of contemporary online news portals for this purpose.

In this study, we propose evaluating the goodness of different sources of external data for constructing ad-hoc corpus to connect topic clusters using extended semantic edges. In addition to the contemporary online news media corpus, we use two other independent external corpus for constructing extended semantic edges, namely contemporary online forum discussions and contemporary blog posts. To the best of our knowledge, ours is the first study of its kind.

We use three large scale real-life Twitter datasets, namely Libya 2011 political turmoil, Egypt 2011 political turmoil and London 2012 Olympics, having thousands of users and up to millions of tweets, to conduct experiments. For all datasets, we find online news media to best capture the evolution of discussion topics along social relationships, measured by the normalized mutual information or *NMI* (Coombs and Dawes, 1970) of the social discussion threads to discussion sequences. We believe this insight to be both novel and interesting. We further observe that, for most cases, online discussion forums perform better compared to blogs.

In summary, the main contributions of our work are the following.

- We empirically evaluate different contemporary relevant external documents, to establish extended semantic relationships across topical clusters formed around events.
- We assess the goodness of contemporary online discussion forums, blogs, online news media and Random search results, in forming extended semantic relationships, and find online news media to be most effective in creating ad-hoc corpus around given events.
- We demonstrate our findings on microblogging data using three real events.

2 Problem settings and our approach

Problem settings

We observe the following:

1. There exist several concepts that are connected in a given context, but are not connected by any widely accepted relationship such as synonyms, antonyms, hypernymns, hyponymns *etc.* when taken in isolation. As an example, *damage* and *relief* are intuitively connected concepts in semantic clusters containing {*damage, fire, death, toll*} and {*fire, relief, spray, water*}. Yet, none of the traditional semantic relationships will connect these when considered in absence of the larger context, namely an event of fire. Practically, discussions on microblogs about damage caused by fire stands a realistic chance to evolve towards discussions about relief.

2. Events on microblog networks form around non-traditional, temporary and contemporary factors, entities and relationships. It is impractical to expect well-defined corpus to exist a priori.

Clearly, creating corpus applicable for a given event, to be able to connect concepts that are related in context of the event, is a research problem to solve. It is also important to assess the quality of the corpus created in the process.

Algorithm

In absence of traditional a priori corpus, we attempt to construct ad-hoc corpus applicable to the context of the event. We follow the approach of (Narang and Nagar, 2013) to construct our graphs, conducting our experiments and measuring the goodness of our results. We use Twitter as our testbed. We attempt to use four independent types of contemporary external documents to be able to connect concepts related contextually, namely online forum discussions, blog posts, news media and Random search documents, to derive the *extended semantic relationships*. Our approach consists of the following steps.

2.1 Topic-based cluster creation

We collect tweets belonging to an event from Twitter for our experiments. The whole tweet corpus is divided into clusters of tweets which are semantically related . Event topic cluster detection not being the focus of our work, we use an existing online clustering algorithm (Weng and Lee, 2011) to create clusters of topics related semantically. A semantic event cluster E^i is represented

as $\{K^i, [T_s^i, T_e^i]\}$, where K^i denotes the set of keywords extracted from the tweets which form the event E^i and T^i is time period of the event. We use existing established methods for computing K and T. K contains *idf* vector and proper nouns (extracted by PoS tagging) from the tweets, and uses Standford's NLP Toolkit and the associated Named Entity Recognizer. T is simply the time of first and last tweet in the event cluster.

2.2 Extracting the relationships

Essentially, we generate an event topic graph $\mathcal{G} = \{\mathcal{E}, \{R\}\}$, in which $\mathcal{E}$ represents the event topics (topical clusters), and act as the vertices of the graph. The set $\{R\}$ represents the relationship edges between the clusters, and are formed from each of contextual semantic, temporal and social perspectives.We, hereby, elaborate on the algorithm used for extracting these relationships.

Extended Semantic Relationships: This relationship is extremely useful but challenging to establish. Lets us motivate the need for such relationship by a simple example. Consider two events with associated keywords E_1 = {damage, earthquake, dead, toll} and E_2 = {earthquake, relief, shelter}. Now, lets pick one work from each set *damage* and *relief*. One cannot establish any of the widely accepted relationships like synonym, antonyms, hypernyms, hyposnyms etc when the words are taken in isolation. However, coupled with prior knowledge about the larger event *earthquake*, the words can be semantically related. In essence (with abuse of notation and terminology), damage and relief are independent variables without extra information, however, they are related given *earthquake*. Therefore, we would like to add the semantic edge between these events. We use external corpus to extract and quantify such semantic relationships. (Narang and Nagar, 2013) used only Google News for creation of the external corpus. But, due to the increased presence of users on Internet, these global and prominent topics are bound to be discussed in blogs and online discussions forums. The natural question which arise is then, which corpus is deemed to be best for the purpose.In this paper, we use different data sources as ad-hoc corpus, namely contemporary online discussions, blogs , online news and Random Search documents, to form four different graphs extended semantic per event.The novelty of our work lies in empirically determining the goodness of each of these four different data source types in forming ad-hoc corpus.

Extended semantic relationship extraction
We establish weighted extended semantic relationships across event clusters by the following steps. The input to the extended semantic relationship extraction algorithm for two events E^i and E^j is keyword list K^i and K^j.

Step 1: Generating Pairs and Pruning Mechanism- We generate $|K^i| \times |K^j|$ pairs of keywords which need to be evaluated for extended semantic relationship. Such large number of pairs would pose computational issues. To handle this, we prune pairs which are related semantically (synonyms, antonyms, hypernymns and hyponynms).We look at the similarity scores of K^i and K^j in Wordnet. We use the well-established Lin's method (Lin, 1998) to compute similarity scores of K^i and K^j using the feature vector built into the Wordnet lexical database.For sake of completeness, please note that Lin's measure of similarity between pair of words w1 and w2 is defined as:

$$sim(w1; w2) = \frac{2I(F(w1) \cap F(w2))}{I(F(w1)) + I(F(w2))}$$

, where $F(w)$ is the set of features of a word w, and $I(S)$ is the amount of information contained in a set of features S. Assuming that features are independent of one another, $I(S)$ $=-\sum_{f \in S} Plog(P(f))$, where $P(f)$ is the probability of feature f.

We retain a pair of words if the similarity score S_{ij} is lesser than a desirable similarity threshold S, and discard the pair otherwise. Since POS tagging is done on the tweets in the event, we also remove pairs where one of the word is Proper Noun or Active Verbs.

Step 2: Document Corpus Generation and Searching- We use the keywords used for filtering Twitter Public API to search for news stories for the same time period on contemporary external documents.The retrieved documents act as our external corpus. We create an inverted index for this corpus, where for each word we store the document ids as well as the frequency of the word in the documents. Given the pair of words (K_l^i, K_m^j) (we will omit subscript l and m, when there is no ambiguity), we find the intersection of corresponding document lists. Therefore, at the end of this step we have list of documents (denoted by D_{lm}) in which both the

words co-occur along with their frequency in the documents.

Step 3: Pairwise Score Computation - For each of the selected document, we compute the coupling of the pair of words. Assume, $C(K_l^i, D_t)$ gives the *tf-idf* score of word K^i in document D_t. The pairwise coupling can be computed as minimum $(C(K_l^i, D_t), C(K_m^j, D_t))$. The overall coupling is calculated as average of coupling over all documents.

Step 4: Overall Score Computation - This process is repeated for all pair of words in E^i and E^j. Finally, for a given pair of event clusters E^i and E^j, if w_{ij} words were discarded and the rest were retained, then

$$overallscore = \frac{\sum_{K^i,K^j} Coupling}{(|K^i| \times |K^j| - w_{ij})}.$$

The final scores are ranked in descending order and top K% are selected based on user preference or can be pruned based on threshold.

Social Relationships: Direct social connections are the core constituent elements of social relationships. Higher order social relationships can be established by exploring the social network structure. Well-defined structures such as communities with maximum modularity (Girvan and Newman, 2002; Clauset and Newman, 2004) can be extracted using efficient modularity maximization algorithms such as BGLL (Blondel and Guillaume, 2008).

Social relationship extraction

We construct social linkage graphs between pairs of events using social connections of event cluster members to construct edges. Each event associates a number of microblog posts (tweets) from a set of members of the microblog network (Twitter).

A person P is said to belong to an event cluster E^i iff $(\exists M)$, a microblog post, made by P, such that $M \in E^i$. Please note that with this definition, a person can potentially belong to multiple event clusters at the same time.

These connections are established by participation of direct social neighbors of individuals across multiple events. We draw an edge across a given pair of events if there is at least one direct (one-hop) neighbor in each event belonging to the pair of events. The weight of an edge between event cluster E^i and E^j is determined by the total number of one-hop neighbors existing between these two clusters. So if E^i has P^i memberships, E^j has P^j memberships, the average number of neighbors in E^j of a member belonging to E^i is a_{ij} and the average number of neighbors in E^i of a member belonging to E^j is a_{ji} then the strength of the social edge between E^i and E^j is $(P^i.a_{ij} + P^j.a_{ji})$.

Temporal relationship The third kind of relationship we extracted is temporal relationship. We look at two kinds of temporal relationships. (a) We draw a temporal edge from event E^i to event E^j if E^i ended before E^j started and the timespan between the two events has to be less than or equal to 2 days. This follows from the assumption that on microblogging services like Twitter, a discussion thread will not last longer than this.This thresholding also prevents the occurrence of spurious edges across different clusters.It captures the *meets* and *disjoint* relationships described by (Allen and J.F., 1983). We call this a T_1 temporal relationship. (b) We draw a temporal edge from event E^i to event E^j if E^i started before E^j started, and ended after the start but before the end of E^j. This captures the *overlaps* relationship described by (Allen and J.F., 1983). We call this a T_2 temporal relationship. Please note that unlike the undirected semantic and social relationship edges, a temporal relationship edge is always directed. The source of a temporal relationship edge is the event with the earlier starting time, and the sink is the one with the later starting time.

2.3 Identifying and characterizing discussions

Finally, after establishing the relationships, we identify Discussion and Social discussion sequences in the same manner as described by(Narang and Nagar, 2013).

Identifying discussion sequences: A discussion sequence graph is defined as, a directed acyclic graph (DAG) of topics that are related using the semantic edges obtained by our earlier semantic relationship extraction process, where the relationships are established over time. Intuitively, a discussion sequence captures the topical evolution of discussions over time. We identify discussion sequences using the logical intersection (AND) of the relationship set of the undirected semantic and the directed temporal graphs, with the directions of the latter preserved in the output.

So, the discussion sequence DAG $\mathcal{G}_{DS}$ is formed as: $\mathcal{G}_{\mathcal{DS}} = \{\mathcal{E}, \{\{R_T\} \cap \{R_S\}\}\}$, where the set $\{R_T\}$ represents the edge set of the directed temporal graph and the set $\{R_S\}$ represents the edge set of the undirected semantic graph.

Identifying Social discussion sequences: We take the above graph, and take an edge set intersection with the social graph. This results in retaining the discussion sequences that are socially connected and eliminating the discussion sequences that are socially disconnected. The retained discussion sequences show the social evolution of discussion topics around events on microblogs. Hence, these socially connected discussion sequences are identified as social discussion threads.

2.4 Evaluation

In order to measure the goodness of the approach, we find the BGLL (Blondel and Guillaume, 2008) communities for the discussion sequence graphs and social discussion threads, and compute the normalized mutual information (NMI) (Coombs and Dawes, 1970) for each of these intersections. Please note that NMI values range between 0 and 1, and higher NMI values indicate higher overlaps of the two inputs.
(Narang and Nagar, 2013) showed that Discussion threads tend to evolve socially and as a result, the NMI values between communities formed on Discussion Sequences and Social Discussion thread is higher than in between BGLL communities formed on purely Social and Semantic Graph. In this paper, we will compare the NMI values between Discussion threads and Social Discussion threads with taking different extended semantic graphs for their construction.The corpus which results in highest NMI value between the two graphs has most relevant retrieved documents for the event.

3 Results

We collect Twitter data from three events that had created significant impact on social media - Libya 2011 political turmoil (collected 4 - 24 Mar'11), Egypt 2011 political turmoil (collected 1 - 4 Mar'11) and the London 2012 Olympics (collected 27 Jun - 13 Aug'12). We use Google News (http://news.google.com) with custom date ranges to collect the contemporary online news data, and Google blog and discussions search options on Google's portal (http://w.google.com) with custom date ranges to collect the blog and forum discussions data respectively.We also used Google Search (http://google.com) to collect random search results for the same events which will be a mixture of all the data sources to act as a baseline.We gave the same keywords over the same time range while collecting documents from Google which were used for collection of tweets in the Twitter. Table 1 shows the basic statistics of the datasets.

Following the approach outlined in Section 2, we form semantic topic clusters from the tweets following the algorithm of (Weng and Lee, 2011). We now establish extended semantic, social and temporal relationships. For extended semantic relationships, we form four graphs, one each for online news media, discussions, blog and Random documents. For temporal relationships, we form two graphs, one each for the *follows* and *overlaps* relationships. Thus overall, for each dataset (Libya, Egypt and London), we construct 8 different graphs, constructing a total of 24 graphs for experimentation.

We now identify the discussion sequences by taking a logical intersection of the extended semantic and temporal graphs, and the social discussion threads by taking a logical intersection of the discussion sequences with the social relationship graph. We find the NMI (Coombs and Dawes, 1970) across these two graphs using the BGLL (Blondel and Guillaume, 2008) communities formed around these two graphs. We retain the top 10%, 20%, 30%, 40% and 50% of the graph edges and repeat our experiments to observe the overall trend. Figure 1 captures our findings for the temporal *follows* relationships.

The results clearly indicate that in each case, contemporary online news yields the best results (maximum NMI values). In most cases (except Egypt), online discussion forums give better results compared to blogs. This trend becomes more yet prominent as we retain higher fraction of the relationships.Random search results generally behave the worst, except in London which is a little surprising and interesting.

Table 2 shows the corresponding results for the temporal *overlaps* relationship for Libya, which also prominently shows a similar trend. We observe similar trends for other temporal *over-*

Table 1: Keywords used to collect the Twitter datasets, dates of data collection, number of users, tweets collected and clusters, and number of contemporary external news, forum and blogs documents collected

Dataset	Keywords	NumUsers	Tweets	Clusters	#News	#Forum	#Blogs
Libya	Libya, Gaddafi	83,177	1,011,716	1,344	3,266	280	263
Olympics	London, Olympic	1,313,578	2,319,519	299	1,186	516	307
Egypt	Egypt, Protest	37,961	60,948	141	1,753	513	285

Table 2: NMI values for temporal *overlaps* based graphs of Libya

Source	10%	20%	30%	40%	50%
Blogs	0.01	0.04	0.09	0.09	0.13
Forums	0.01	0.06	0.06	0.11	0.16
News	0.02	0.09	0.14	0.09	0.17
Random	0.02	0.04	0.00	0.08	0.11

laps graphs also, with the London Olympics data shows a few exceptions (omitted for space constraints).

To eliminate the possible bias due to the number of documents received for each type of corpus, we also repeated the experiment with taking top 200 documents from each sources namely, Online news, blogs, discussions and Random documents. We, then evaluated the performance of these corpora on the Libya dataset.

The figure 2 shows the NMI graph for the Libya dataset with taking only top 200 articles from each of the data sources. The contemporary news article consistently give best results even in this case which corresponds to the finding in the above experiment. Although, Random results also give an equivalent performance but this can be attributed to the fact that the initial results in Google search mostly contains Google News results which will be in prominence because of the low number of documents selected in this experiment.

4 Related Work

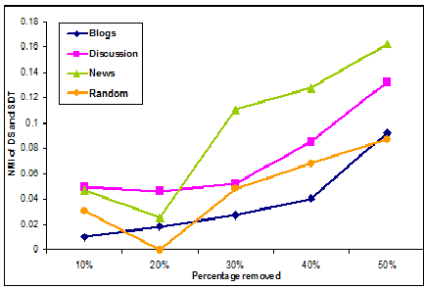
Significant research has been conducted on content analysis of information discussed on social media sites (Kwak and Lee, 2010). Grinev et al. (Grinev and Grineva, 2009) demonstrate TweetSieve, a system that obtains news on any given subject by sifting through the Twitter stream. Along similar lines, Twinner by Abrol et al. Abrol and Khan (Abrol and Khan, 2010) identify news content of a query by taking into account the geographic location and the time of query. Nagar et al. (Nagar and Seth, 2013) demonstrate how content flow occurs during natural disasters.

Several ways to cluster social content have been studied. There has been work on clustering based on links between the users by doing agglomerative clustering, min-cut based graph partitioning, centrality based and Clique percolation methods ((Porter and Onnela, 2009), (Fortunato, 2007)). Other approaches consider only the semantic content of the social interactions for the clustering (Zhou, 2006). More recently there has been work on combining both the links and the content for doing the clustering ((Pathak and Delong, 2008), (Sachan and Contractor, 2012)). In (Narang and Nagar, 2013) relationships between clusters are determined based on semantic, social and temporal information but did not study the impact of different corpus on their results.

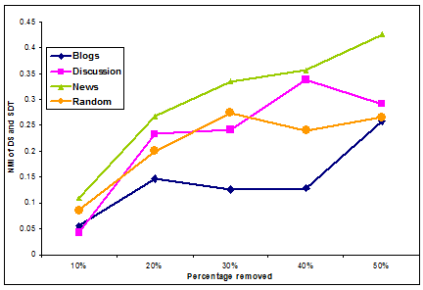
External corpora have been used by researchers to create knowledge base in various fields like for question-answering (Clarke and Cormack, 2002; Dumaisl and Banko, 2002) models such as Chatbots etc, helping machine to translate documents like expanding queries (Kilgarriff, 2003; Metzler and Diaz, 2005) and also for improving Information retrieval using external information (Xu and Croft, 2000; Diaz and Metzler, 2006). They use generic corpora and to the best of our knowledge, there is no study which analyses the relevance of different corpora for the given problem.

5 Conclusions

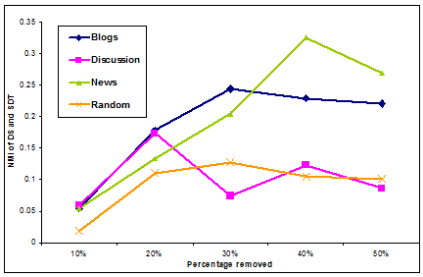
In this work, we studied different contemporary online external data sources for constructing adhoc corpus to connect event topic clusters. We explored the content of contemporary online discussion forums, blogs, online news media and Mixture of different corpus, and evaluated their effectiveness in establishing semantic relationships across topical clusters. Exploiting the social propensity of evolution of such discussions, we assessed the goodness of these diverse data sources



(a) NMI for Libya turmoil



(b) NMI for London Olympics



(c) NMI for Egypt turmoil

Figure 1: NMI of social discussion threads (SDT) with respect to discussion sequences (DS): temporal *follows* relationship

using Twitter as a microblogging platform, and eventual NMI values as a qualitative indicator of the goodness of the extended semantic relationships established.

We found contemporary online news media to be the most effective type of external data source for creating ad-hoc corpus, using three large real-life Twitter datasets collected around major events. Further, we found contemporary online discussion forums to be usually, but not always, more effective compared to contemporary blogs.We also found using Mixture of all documents to be mostly give the worst performance.

Our work will be useful to studies and applications that require capturing the evolution of topical discussions on microblogs like Twitter. As future work, we propose evaluating other external

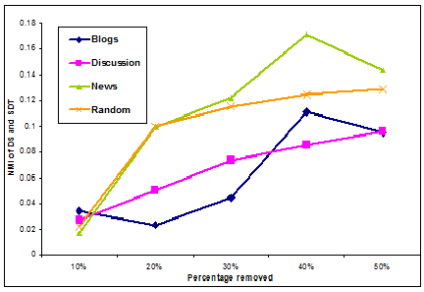


Figure 2: NMI of social discussion threads (SDT) with respect to discussion sequences (DS): temporal *follows* relationship with retaining only top 200 articles

sources of semantic data, and also apply on other microblogging platforms and data sets, for a more comprehensive and complete study.

References

S. Abrol and L. Khan. Twinner: Understanding news queries with geo-content using twitter. In *Proceedings of the GIS*, 2010.

Allen J. F.: *Maintaining Knowledge about Temporal Intervals.* In: Communications of the ACM (1983).

Blondel V.D., Guillaume J.L., Lambiotte R., Lefebvre E.: *Fast unfolding of communities in large networks.* In: J. Stat. Mech. P10008 (2008).

Clarke C. L. A., Cormack G. V., Laszlo M., Lynam T. R., Terra E. L.: *The impact of corpus size on question answering performance.* In: SIGIR 02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (2002).

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. In *Proc. Ntl. Acad. Sci, USA*, 99(7821),2002.

Clauset A., Newman M. E. J., Moore C.: *Finding community structure in very large networks.* In: Phys. Rev. E. 70(066111) (2004).

Coombs C. H., Dawes R. M., Tversky A.: *Mathematical psychology: An elementary introduction* In: Englewood Cliffs, NJ: Prentice-Hall (1970).

Diaz F., Metzler D.: *Improving the estimation of relevance models using large external corpora.* In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

Dumais S., Banko M., Brill E., Lin J., Ng. A. : *Web question answering: is more always better?* In SIGIR 02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (2002).

S. Fortunato and M. Barthelemy. Resolution limit in community detection. In *Proceedings of the National Academy of Sciences*, 104(1):3641, 2007.

M. Grinev, M. Grineva, A. Boldakov, L. Novak, A. Syssoev and D. Lizorkin. Tweetsieve: Sifting microblogging stream for events of user interest. In *Proceedings of the SIGIR*, 2009.

Kilgarriff A., Grefenstette. G.: *Introduction to the special issue on the web as corpus.* In: Computational linguistics 29.3 (2003): 333-347.

Kwak H., Lee C., Park H., Moon S.: *What is Twitter, a Social Media or a News Media.* In: Proceedings of the WWW (2010). In: Proceedings of the WWW (2010).

Lin D.: *An information-theoretic definition of similarity.* In: Proceedings of the International Conference on Machine Learning (1998).

Metzler D., Diaz F., Strohman T., Croft W. B.: *Umass at robust 2005: Using mixtures of relevance models for query expansion.* In: The Fourteenth Text REtrieval Conference (TREC 2005) Notebook (2005).

Narang K., Nagar S., Mehta S., Subramaniam L.V., Dey K.: *Discovery and analysis of evolving topical social discussions on unstructured microblogs.* In: European Conference on Information Retrieval (2013).

S. Nagar, A. Seth and A. Joshi. Characterization of Social Media Response to Natural Disasters. In *Proceedings of the WWW*, 2012.

N. Pathak, C. DeLong, A. Banerjee and K. Erickson. Social topics models for community extraction. In *Proceedings of the 2nd SNA-KDD Workshop*, 2008.

M. A. Porter, J.-P. Onnela and P. J. Mucha. Communities in networks. In *Notices of the American Mathematical Society*, 56(9):1082-1097, 2009.

M. Sachan, D. Contractor, T. A. Faruquie and L. V. Subramaniam. Using Content and Interactions for Discovering Communities in Social Networks. In *Proceedings of the WWW*, 2012.

Weng J., Lee B.S: *Event detection in Twitter.* In: ICSWM - Proceedings of the AAAI conference on weblogs and social media (2011).

Xu J., Croft. W. B.: *Improving the effectiveness of information retrieval with local context analysis.* In: ACM Trans. Inf. Syst., 18(1):79-112 (2000).

D. Zhou, E. Manavoglu, J. Li, C. L. Giles and H. Zha. Probabilistic models for discovering e-communities. In *Proceedings of the WWW*, 2006.