Exploiting Social Media to Provide Humanitarian Users with Event Search and Recommendations

John Edmonds
University of Maryland
College Park, MD, USA
jedmond3@umd.edu

Louiqa Raschid
University of Maryland
College Park, MD, USA
louiqa@umiacs.umd.edu

Hassan Sayyadi
University of Maryland
College Park, MD, USA
sayyadi@cs.umd.edu

Shanchan Wu
University of Maryland
College Park, MD, USA
wsc@cs.umd.edu

ABSTRACT

Humanitarian decision makers rely on timely and accurate information for decision-making. Since satisfactory disaster response is key to building public trust and confidence, they need to monitor and track disaster related discourse to gauge public perception and to avert public relations disasters. Social media, e.g., the blogosphere, has empowered citizens to provide content and has increased information diversity. The challenge is to make sense of this diverse and noisy data and interpret results in context. For example, search results can be clustered around an event or occurrence at some geo-location and time. Personalization and recommendations can further filter content and focus on the most relevant and important data. We apply our research on event detection and recommendation to support event based search and apply it to a large blog collection (blog.spinn3r.com).

INTRODUCTION

Humanitarian decision makers rely on timely and accurate information, e.g., a comprehensive situation report, for decision-making. Increasingly, public perception, trust and feedback are important aspects of the success of relief operations, as was demonstrated after hurricanes Katrina and Rita and the Sichuan earthquake. Soon after the disaster, fora and blogs became important sources and also an early warning of the public relations disasters to follow. Humanitarian decision makers increasingly need to monitor and track disaster related discourse, to gauge public perception, and to possibly avert these public relations disasters. The inherent characteristics of a disaster, e.g., being unpredictable, the evolution of available information sources, and the emergence of new trends and patterns, compound the challenge of meeting these information needs. Information sources could rapidly swing from a dearth to a deluge and the decision maker always faces the difficult task of identifying and querying (and re-querying) sources, and analyzing and integrating results to support timely decision-making.

Social media and interactions on the blogosphere has the potential to become a vital source of breaking news as well as knowledge, reflecting the expertise and opinion of crowds. Such sources are of particular importance in evolving disasters where experts may not be known a priori; there is a need for a diversity of information; information evolves over time and the quality can vary. This increases the value of the blogosphere as a valuable source of information. On the other hand, crowd sourcing can also create a massive stream of irrelevant and low quality information. For instance, users of social networking sites (LinkedIn, Twitter) may receive hundreds or thousands of daily blogs, messages, fora, etc., from users with whom they interact, e.g., they belong to a common group. Further, Web search engines now index blog sites so a typical search may return many posts.

The challenge is to make sense of data and interpret results in context. Search can be clustered around an *event occurring at a geo-location and during a time interval.* Recommendation can further filter content and can alert the user to potentially useful current and future content. While commercial search engines and news sites may provide such capabilities, they typically target a wide range of users and their generic information seeking needs. For example, a typical visitor on CNN may only want to read a short description about a disaster but a regional or local government specialist may want to know about all updates to roads and travel conditions in some region affected by a disaster, or may want to be alerted when there is a trend, e.g., a high frequency of

Reviewing Statement: This paper represents work in progress, an issue for discussion, a case study, best practice or other matters of interest and has been reviewed for clarity, relevance and significance.

negative posts complaining about the lack of adequate relief or health services. Our objective is to cluster information around relevant events that are specific to the disaster domain, and then exploit event detection, personalization and recommendation techniques to better meet information gathering needs.

Our motivating examples and additional experiments rely on a large blog collection (blog.spinn3r.com) and human users have validated the results. The dataset from Spinn3r.com is a set of 44 million blog posts crawled between August 1st and October 1st, 2008 [5].

SOCIAL MEDIA AND EVENTS

Our motivating example will focus on providing the user with the capability to retrieve and interpret search results in context. There are many ways to provide context (also called facets) to improve the user search experience. One of the most common ways to provide context, in particular in the humanitarian domain, is to identify events that occur at (or near) some geo-location, during some time interval. One can then cluster documents or posts around such events. As an example, hurricane Katrina is a specific event at some geo-location that occurred during an interval. Clearly an event may be composed of many sub-event(s); each sub-event will be distinguished along the geo-location and/or the time dimension, possibly at a finer level of granularity. For simplicity we only consider higher level atomic events.

We represent an event using a core set of keywords from the KeyGraph. Figure 1 illustrates a KeyGraph. This figure is a small part of a much larger graph. The nodes or keywords in the figure represent noun-phrases that are extracted from a collection of documents (blog) in 2008. The figure illustrates that there are several communities of keywords. Informally, a community is a subgraph where all the nodes in the subgraph are well connected to each other, and where there are very few edges between nodes that occur in different communities. For example, the group of keywords in the upper right of the figure, including “cyclone”, “Myanmar”, “foreign aid workers”, etc., create a community that represents the event “cyclone Nargis in Myanmar”. We expect that presenting such a set of keywords in the context of a KeyGraph can improve the user search experience, in comparison to returning a group of documents to the user.

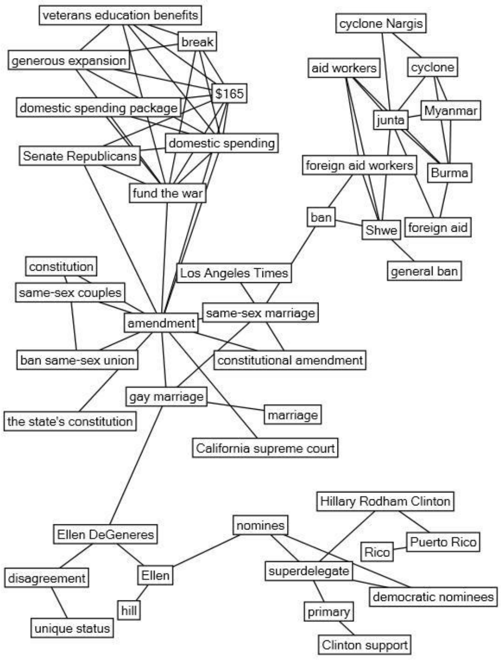

Figure 1. A small sample of the KeyGraph

Next, we illustrate how event based retrieval will further help a user interpret trends that appear in search results. Consider a group of documents (blog posts) that are centered on cyclone Nargis. Figure 2 shows the distribution of the intensity of posts (number of documents) about this event, aggregated on a daily basis, in May 2008. The figure shows that the number of documents increases sharply at the beginning of the event and a peak occurs on May 7 and then number of documents decreases. However, we observe an interesting pattern; there are two peaks in the document intensity. The second peak occurs on May 10. This peak discusses *a related event where* *the Myanmar government refused to accept international aid and did not allow international aid workers to enter the country*. This action created a huge controversy in online communities. Our example both illustrates the importance of keywords in a KeyGraph to identify events, as well as the need to track trends.

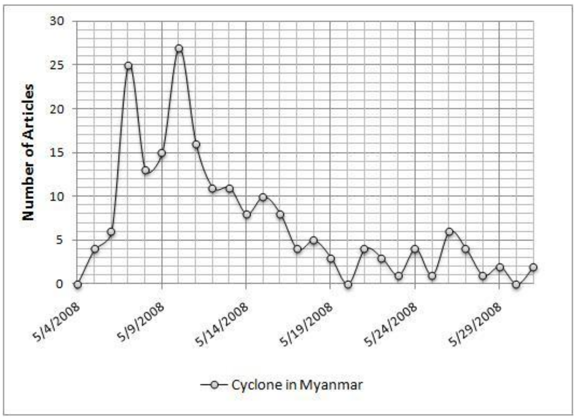

Figure 2. Document frequency of Cyclone Nargis in Myanmar

SOCIAL MEDIA AND PERSONALIZED RECOMMENDATIONS

Next, we discuss recommendations. When a humanitarian specialist reads a document (post) about an event, she may also be interested in recommendations of other bloggers or writers who are most likely to post on this topic in the future. We refer to a stream of posts by a blogger as a blog channel. Our research uses the history of past posts on a blog channel to build a profile and to make recommendations. Consider the event related to the 2008 war between Russia and Georgia or the 2008 South Ossetia War; it is represented by a collection of approximately 40 keywords including the following: *tskhinvali, tbilisi, russia, russia georgia, south ossetia, saakashvili, russian force, troop, tank, rumsfeld, republican presidential candidate john mccain, regist foreign agent, south ossetia abkhazia, etc.* Our blog channel recommendation system identified the following 6 blog channels: http://blog.oneworld.am http://zaxi.livejournal.com http://4international.wordpress.com

http://russiaotherpointsofview.com http://vineyardsaker.blogspot.com and http://darussophile.blogspot.com .

All of these blog channels focus on international affairs. During the period from August 2008 to October 2008, the period in which we analyze the blog channels, they each posted at least one relevant and important post on the event. Further, these channels continuously post on similar events and they are still active. We compare our recommendations with a search on Google where our query included the same keywords above, defining this event. Google returned the following 4 links: http://arisdeslis.blogspot.com/2008_08_01_archive.html

http://noworldsystem.com/category/military-base/ http://digbysblog.blogspot.com/2008_10_01_archive.html

and http://www.cs.cmu.edu/afs/cs/user/lsl/Nice/Urdu-MT/code/Tools/Transliterator/merged.hist.

Our observation is that the links from Google are general links to blog channels that post on many topics; they are not the most relevant or important bloggers to post on this event of interest. Further, the third link is a university web page that is not relevant to our event. We conclude that a commercial search engine may be too general with a focus on the general public, and will not target the needs of a humanitarian specialist.

EVENT DETECTION AND EVENT BASED CLUSTERING OF SEARCH RESULTS

A definition of a topic (event) in the Topic Detection and Tracking (TDT) community is a specific occurrence at a specific time and in a specific place [9]. Past solutions for TDT have typically used clustering algorithms. Documents are treated as database records and words in the document are treated as features. Next, variations of term frequency and inverse document frequency (TF/IDF) [7] are used to compute feature values and cosine similarly is used as a similarity (or distance) measure. Although the gold standard algorithms based on these approaches worked very well for small, clean data collections, they do not scale well to social media and the blogosphere. This is due to the tremendously rapid growth of news sources including community news sources such as bloggers. Social media is also noisy. There is agreement that for each specific topic or event, there is a *core group of identifying words* that will be common to all related documents on that event. This is the inspiration for a new generation of TDT models. Our research is based on the KeyGraph algorithm for event detection based on identifying this core group of words. We briefly summarize the approach and refer the reader to [8] for details. The KeyGraph algorithm can be summarized in the following three steps: (1) Building the KeyGraph; (2) KeyGraph community structure analysis; (3) Document clustering.

Building the KeyGraph: A KeyGraph is built by extracting a set of keywords from the corpus, i.e., the collection of documents to be processed. Each document can be represented as a bag of words. The keywords may be noun phrases, named entities, etc. A node in the KeyGraph is created for each keyword. Then, an edge between two nodes is added if the corresponding two keywords occur in a document. The KeyGraph is a keyword co-occurrence graph. **Community Detection in the KeyGraph:** A *core group of identifying keywords* representing each topic or event participates in meaningful topical relationships with each other. The analogy is to consider the KeyGraph as a social network of relationships between keywords. Community detection algorithms can identify each core group of keywords. From Figure 1, communities of keywords are densely linked, while there are few links between nodes (keywords) from different topical or event communities. Our algorithm applies several heuristics to identify communities. **Document Clustering:** We use cosine similarity to discover document clusters for each community of keywords. We have evaluated our KeyGraph based approach for event detection on the TDT4 dataset [9] and have shown that the method has similar accuracy compared to the gold standard algorithms [4,10]. Further, we show that the performance of the KeyGraph algorithm is bound by the number of documents and so can be applied to on-the-fly event detection. Our approach is also able to handle noisy blog collections.

RECOMMENDING BLOGCHANNELS

Recommendation systems typically use collaborative filtering techniques to identify other users who have similar profiles (similar query keywords, similar history of visited posts, etc.). The recommendations or reviews or ratings of other users are then processed to provide a recommendation to a target user. We model the blogosphere as a set of blog channels where each new post is an event and a blog channel is a (temporal) data stream of events. A humanitarian decision maker who relies on the blogosphere to obtain new information and to keep track of events or trends may have the following information delivery needs: Which blog channel is blogging on Event X? How many are blogging on Event X? Which blog channel is likely to respond to my post on Event X? How frequently or how many posts have appeared on Event X and will continue to appear?

The dynamics of the news cycle has been studied through tracking of topics and memes as they disseminate and evolve over time [6]. The Harvard Media Cloud project [11] examines interesting patterns as news crosses multiple data streams. The BlogScope project [1] has been very successful at online analysis of high volumes of social media. For example, their alert service will inform users when there is a keyword burst for a keyword of interest; this service has superior accuracy to Google Alerts [2].

Given a user who is interested in following Event X, the task is to predict which blog channels are likely to post posts on X within the next few hours or the next news cycle. To address this challenge, one could find authoritative sources with multiple recurrent posts on Event X but these sources may be well known a priori to the humanitarian decision maker. What would be interesting is discovering additional sources to follow. These sources may only post occasionally on Event X, but they may post on the topic before others, or they may respond to posts on an important blog channel. Our blog channel prediction task must consider all these cases.

Figure 4 presents the system architecture. Each blog channel is associated with a history of posts that have appeared on that channel. This history is used as a training set and is used to build two indices. The first is a post index which is similar to a document index. The second is a profile index representing all the posts in the history of the blog channel. The Prediction and Recommendation modules can choose among three methods to make accurate predictions. PROF constructs a topical profile over the documents in a blog channel. It uses a temporal penalty to favor recent posts. VOTE uses topical pairwise similarity between the incoming post and all documents in the blog channel history to choose some candidate posts; voting is used by the candidate posts to select the Top K blog channels. temporalVOTE extends VOTE to use temporal and topical similarity when choosing candidate posts. Details of the methods and evaluation results are in [3].

CONCLUSION

We present the application of two research streams to customize social media for humanitarian users. We focus on the blogosphere to demonstrate our solutions. Our first objective is identification of events and clustering of posts around events. Our second task is providing recommendations of blog data streams (blog channels) to users to watch for future posts.

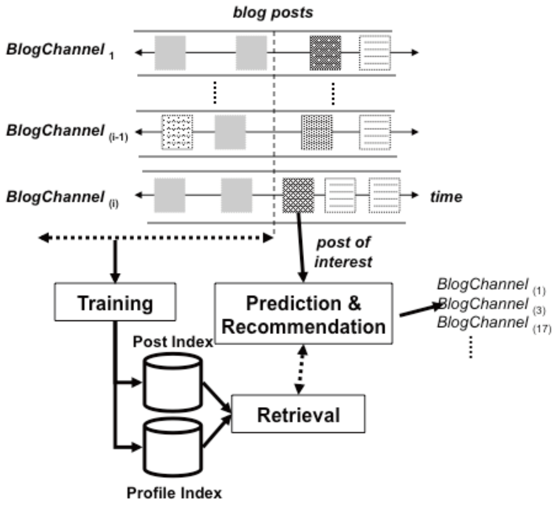

Figure 4. System architecture

ACKNOWLEDGEMENT

This research was partially supported by NSF award CMMI0753124.

REFERENCES

1. Blogpulse. http://www.blogpulse.com/.
2. Google alerts. http://www.google.com/alerts.
3. Which Blog Channel to Watch? Monitoring and Prediction of Top K Blog Channels for Topic Tracking. ANONYMOUS. Under review, 2010.
4. J. Allan, R. Papka, and V. Lavrenko. On-line New Event Detection and Tracking. Proceedings of the ACM International Conference on Research and Development in Information Retrieval, pages 37-45, 1998.
5. K. Burton, A. Java, and I. Soboroff. The ICWSM 2009 Spinn3r Dataset. Proceedings of the Conference on Weblogs and Social Media, 2009.
6. J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. Proceedings of the International Conference on Knowledge Discovery and Data Mining (SigKDD), 2009.
7. G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management, Volume 24, Issue 5, pages 513-523, 1988.
8. H. Sayyadi, M. Hurst, and A. Maykov. Event Detection and Story Tracking in Social Streams. Proceedings of the Conference on Weblogs and Social Media, 2009.
9. Topic Detection and Tracking (TDT) Project. **http://www.nist.gov/speech/tests/tdt/**.
10. Y. Yang, T. Pierce, and J. Carbonell. A Study on Retrospective and On-line Event Detection. Proceedings of the ACM International Conference on Research and Development in Information Retrieval, pages 28-36, 1998.
11. E. Zuckerman. Global Attention Profiles: First Steps Towards a Quantitative Approach to the Study of Media Attention. Working Paper, Berkman Center for Internet and Society, Harvard Law School, 2003.