Twitris: Socially Influenced Browsing

Ashutosh Jadhav[1], Wenbo Wang[1], Raghava Mutharaju[1], Pramod Anantharam[1], Vinh Nyugen[1], Amit P. Sheth[1], Karthik Gomadam[2], Meenakshi Nagarajan[1], and Ajith Ranabahu[1]

[1] Knoesis Center, Wright State University, Dayton, OH, USA.
{*ashutosh*, *wenbo*, *raghava*, *pramod*, *vinh*, *meena*, *ajith*, *amit*}@knoesis.org
[2] Ming Hsieh Department of Electrical Engineering,
University of Southern California, Los Angeles, CA.
{*gomadam*}@usc.edu

Abstract. In this paper, we present Twitris, a semantic Web application that facilitates browsing for news and information, using social perceptions as the fulcrum. In doing so we address challenges in large scale crawling, processing of real time information, and preserving spatio-temporal-thematic properties central to observations pertaining to real-time events. We extract metadata about events from Twitter and bring related news and Wikipedia articles to the user. In developing Twitris, we have used the DBPedia ontology.

1 Introduction

The emergence of mircoblogging platforms like Twitter, friendfeed etc. have revolutionized how unfiltered, real-time information is disseminated and consumed by citizens. A side effect of this has been the rise of citizen journalism, where humans as sensors are "playing an active role in the process of collecting, reporting, analyzing and disseminating news and information"[3]. An example of this are observations surrounding the health care debate that have recently originated in the United States. The relayed multimodal observations (texts, images and videos) formed a rich backdrop against traditional reports from the news media.

Perhaps, the most interesting phenomenon about such citizen generated data is that it acts as a lens into the social perception of an event in any region, at any point in time. Citizen observations about the same event relayed from the same or different location offer multiple, and often complementary viewpoints or storylines about an event. What is more, these viewpoints evolve over time and with the occurrence of other events, with some perceptions gaining momentum in certain regions after being popular in some others.

Consequently, in addition to what is being said about an event (theme), where (spatial) and when (temporal) it is being said are integral components to the

[3] http://en.wikipedia.org/wiki/Citizen_journalism

analysis of such data. The central thesis behind this work is that citizen sensor observations are inherently multi-dimensional in nature and taking these dimensions into account while processing, aggregating, connecting and visualizing data will provide useful organization and consumption principles.

Twitter has emerged as a pre-eminent medium for sharing citizen-sensor observations. The ability to perform real time search, towards finding opinions and observations in real time, has emerged as an important problem in the area of internet search. In this paper, we present Twitris, an application that leverages semantic Web techniques, to facilitate the browsing for news and information, using social perceptions as the fulcrum. Twitris, a portmanteau of Twitter and Tetris (for arranging activity in space, time and theme), extracts the current trending topics from Twitter. One of the primary objectives of Twitris is to use social signals towards realizing sensemaking. Sensemaking, defined in [1], is the understanding of connections between people, places and events. Awareness of *who, what, when and where* is a critical component in sensemaking. Unlike other similar applications, Twitris also uses the spatio-temporal aspect of these observations. Twitris is not a real time search platform, but is an application that extracts metadata in the form of tags from Twitter posts. Additional resources such as news feeds and wikipedia articles that give information about a current event, are fetched using the extracted metadata. Twitris thus enables users to browse for news and information, using Twitter data as a starting point.

This paper documents our contributions, challenges in developing Twitris and contains only brief information about the algorithms. Readers are requested to read our research paper in WISE 2009, available at: http://tinyurl.com/twitris-wise09.

2 Contributions

In this paper we present Twitris, an application that extracts summaries of social signals or perceptions behind real-world events and facilitates browsing of related news and info using the extracted social signals as the fulcrum. In developing this system, we have made the following contributions:

1. Crawling topically relevant messages from social streams: The first contribution is an approach to crawl social streams (such as Twitter) for messages relevant to a particular topic. To obtain a set of keywords used in crawling, we adopt a hybrid strategy of using a semantic model (DBPedia) and statistical analysis. Our crawling strategy starts by searching for concepts in DBPedia that are pertinent to the event in consideration. For example, we identify DBPedia concepts that are relevant to healthcare and healthcare debate in the United States. Once the concepts are identified, we proceed to identify other concepts that share a relationship (one hop in length in the current implementation, for computational purposes). This set of identified concepts form the semantic keyword cluster. While the use of semantic models enhances the precision of our crawl, users of Twitter often employ words and terms that are casual in nature and not found in semantic models such

as DBPedia. We employ statistical techniques to identify additional seed keywords for the crawl. To do this, we crawl tweets for a fixed time period (a day in the current implementation). The statistical keyword cluster is created using the keywords extracted from the tweets, along with hashtags (user added metadata in tweets identified by #). This set is enhanced by using Google Insights for Search [4], a free service from Google that provides top searched and trending keywords across specific regions, categories, time frames and properties. Crawling is done using the keywords curated from DBPedia and the statistical analysis.

2. Spatio-Temporal-Thematic analysis of social data: Our work is motivated by the need to easily assess local and global social perceptions or **social signals** underlying events over time. Data pertaining to real-world events have unique characteristics because of the event they represent. Certain real-world events naturally have a spatial and temporal bias while some others do not. For example, when observing what people in Oregon (a liberal state) are saying about the public option in the healthcare debate attack, one might wish not to be biased by global and possibly contrasting perceptions from Georgia (a conservative state). In developing Twitris, we have designed algorithms for spatio-temporal clustering of user posts in Twitter. Twitris currently crawls and indexes events with a few days of lag (usually a week).
3. Browsing using social signals as the fulcrum: Real time search has gained a lot of momentum. However, most approaches to real time search such as Twitter search or One Riot, focus on indexing the tweets themselves. In Twitris, we wanted to develop an approach that focuses not so much on search, but rather than on browsing. We present the descriptors extracted for various events on a map (preserving the spatial attribute), grouped by a date. Users can navigate across dates and across various geo-cluster groups. Just as we believe that individual tweets may not be sufficient, the same can be argued in case of descriptors. We enhance Twitris, by integrating news and Wikipedia. We leverage explicit semantic information from DBPedia to identify relevant news and Wikipedia articles. When a user clicks on a particular descriptor, we display the top six current news items as well as related Wikipedia articles. It is our belief that this allows for a richer browsing experience, rather than showing a few hundred thousand messages, each 140 characters long. Twitris as a platform is extensible to include more resources (such as YouTube and Flickr).

3 Challenges

The volume of messages on a popular topic in Twitter coupled with the short nature (140 characters), poses significant challenges for crawling and processing. Identifying topically relevant posts is a significant crawling challenge. As is the case with many social forums, there is a tendency to go off-topic in Twitter. For

[4] http://www.google.com/insights/search/

example, the debate on healthcare can quickly turn into an argument between the conservative and liberal political principles. Often, the off-topic messages tend to carry a few terms that are pertinent to the topic of interest as well. This is exemplified by terms such as *Obama policy*, and *government option*. Being able to identify and minimize the effect of off topic chatter is a key crawling challenge. In Twitris, we employ DBPedia as the semantic model to identify terms of relevance. This has addressed the challenge partially. This is because of the given the casual nature of the language that is used (such as adhoc abbreviations), and the typos.

Another challenge is the lack of categorization in Twitter. Twitter does not already categorize messages and the strongest cue that is there are the hashtags, which are community generated categorization. However, hashtags are not present in all the tweets. Hence the system has to rely on lexical cues to find topically relevant tweets. Given the vocabulary diversity on Twitter, finding a set of keywords that lend optimal coverage is a challenge. Section 4 discusses this challenge in detail.

In the context of text processing, we have identified the following challenges. In developing Twitris, we leveraged on the implicit semantics derived from the messages to address these challenges. In the subsequent versions, we will employ semantic models such as DBPedia in this task, to achieve greater efficiency.

Text processing challenges:

1. Twitter messages often are written using casual text, and contain a lot of abbreviations. This coupled with a **lack of grammar**, necessitates the need to go beyond conventional text processing approaches.
2. The tweets are interspersed with mentions (denoted by @), shortened URL resources, user names, hashtags etc. Each of these have a semantics of their own and the need to identify and process them, adds computational challenges to the already expensive task of text processing.
3. Conversational practices such as *retweeting* and *mentions*, where users repeat a previously posted message are very popular in Twitter. These repetitions tend to create a statistically significant bias in the corpus.
4. Finally, the volume of messages makes it hard to store every tweet that has been crawled for an event. However, purging the tweets changes the corpus almost on a daily basis. To perform statistical computations such as TFIDF computation on a changing corpus, requires fundamental changes in the way these computations are defined and performed. While in Twitris, we have had considerable success, we also realize that we are not completely real-time at this point. As we go closer to realizing the real-time data analysis ambition, we need to enhance our algorithms to deal with a dynamic corpus.

4 Extracting Social Signals

In this section, we briefly discuss our approach to extracting social signals. Our approach is discussed in detail in [2]. Fundamental to the processing of citizen observations is a simple intuition - “depending on what the event is, social

perceptions and experiences reported by citizen sensors might not be the same across spatial and temporal boundaries". One of the goals in the formulation of our algorithm was to preserve these different story-lines that naturally occur in data. The two questions we wish to answer via our work are:
a. For any given spatial location and temporal condition, can we get an idea of what entities or event descriptors are dominating the discussion in citizen observations?
b. If we know dominant descriptors, can we tell what people are saying about them in different parts of the world and over time?
Broadly, our entity-centric approach to summarizing observations in its three-dimensional space consists of the following steps – partitioning available observations into processable sets based on spatial and temporal biases induced by an event, extracting key descriptors and their contexts.

1. Defining Spatio-Temporal Sets

Different events have different spatial and temporal biases that need to be considered while processing observations pertaining to the event. We first partition the volume of tweets into spatio-temporal sets based on two tuneable parameters - the spatial parameter δ_s and the temporal parameter δ_t. Together these two define the granularity at which we are interested in analyzing observations. δ_s for example is defined to cover a spatial region - a continent, a country, city etc. Similarly, δ_t is defined along the time axis of hours, days or weeks.

Depending on the spatial and temporal bias that an event has, the user picks values for δ_s and δ_t. In the Mumbai event for example, there might be interest in looking at *country level* activity on a *daily* basis. For longer running events like the financial crisis, we might be interested in looking at *country level* activity on a *weekly* basis. For events local to a country, a possible split could be by cities.

Using these two parameters, we slice our data into *Spatio-Temporal Sets* $S=\{S_1,S_2...S_n\}$ where n is the number of sets generated by first partitioning using δ_s and next using δ_t. If δ_s = 'country' and δ_t = '24 hour', observations are grouped into separate spatial (country) clusters. In creating these clusters, we employ clustering techniques based on geographical *bounding box*, along with information from the **geo names** ontology. Every spatial set is then divided further into sets that group observations per day, generating n spatio-temporal sets.

Observations are grouped in a spatio-temporal set depending on the values they have for their timestamps and geocode attributes (see Section **??**). A spatio-temporal set can be represented as $S_i=\{T_i,\delta_{si},\delta_{ti}\}$ where $T_i=\{t_1,t_2,..\}$ is a set of tuples where $t_i=\{t_{id},t_c,t_t,t_g\}$ such that $\forall t_i \in T_i$; $t_g \in \delta_{si}$ and $t_t \in \delta_{ti}$. By processing sets in isolation for key descriptors, we ensure that the social signals present in one do not amplify or discount the effect of signals in the other sets.

2. Extracting Strong Event Descriptors

Given a spatio-temporal set definition, we proceed to extract strong descriptors that are local to this set. In other words, extracted descriptors need to preserve the social signals local to a spatio-temporal set. This can trivially be a function of the probability distribution of the descriptors in the corpus T_i defined by the spatio-temporal set. There has been a plethora of work in the area of extracting important keywords in a corpus [3]. In our case, there are additional strong cues in the entity's temporal and spatial contexts that could be exploited. Here, we formalize the interplay between the three dimensions and define functions that extract strong local event descriptors.

Considering each tweet t_i as a sequence of words, we define a descriptor in our work as a vector of n-grams [5]. Each t_i can then be represented as a vector of word tokens $ngrams_i$=$\{w_1,w_2,...\}$ where w_i is the weight of the i^{th} n-gram. w_i is quantified as a function of the n-gram's thematic, spatial and temporal scores computed as follows. Note that the vector representation of each tweet is constructed after removing stop word unigrams, removing all url segments and domain specific stop words like retweet, rt@ etc. Lucene is used as the indexing mechanism. We also discard all hyperlinks and use only the text portion of tweets.

A. Thematic Importance of an event descriptor: We start by calculating the thematic score of an n-gram descriptor, $ngram_i(tfidf)$, as a function of its TFIDF score in addition to using the following heuristics. These are necessary in order to extract meaningful descriptors from volumes of tweets.

1. The descriptor's TFIDF score is calculated from the Lucene index. This score reflects how important a word is to an observation in a collection of observations in the spatio-temporal set.
2. Supporting the intuition that descriptors with nouns in them are stronger indicators of meaningful entities, we parse a tweet using the Stanford natural language parser and amplify (add to) its TFIDF score by the fraction of words that are tagged as nouns.
3. The TFIDF score is also amplified based on the fraction of words that are not stop words.
4. Lower and higher-order n-grams that have overlapping segments ('healthcare' and 'healthcare reform') and the same TFIDF scores are filtered by picking the higher-order n-gram. The n-grams in each observation are sorted by their $ngram_i(tfidf)$ score and the top 5 are picked for further analysis. Picking top 5 is a satisfactory filter given that the length of our observations is at most 140 characters.

Owing to the varied vocabulary used by posters to refer to the same descriptor, region specific dictions and evolving popularity of words, we found that the above thematic score was not representative of a descriptor's importance. Consider this scenario where the phrase 'Obama campaigns' meant to refer to the 'healthcare reform' campaign by the President, was not used as frequently as 'healthcare' or 'healthcare reform'. The presence of contextually relevant words

[5] We set n=3 in all our experiments

should ideally strengthen the score of the descriptor. However, we also need to pay attention to changing viewpoints in citizen observations that may result in descriptors occuring in completely different contexts. If the usage of 'Obama' is not in the context of the 'healthcare reform', i.e. discussions around Obama surround the recent olympic bid of Chicago, its presence should not affect 'healthcare reform's' importance.
5. Finally, we also use DBPedia as an evidence to calculate the weight of extracted terms. We enhance the weights of descriptors that are share a relationship with one or more terms from the semantic keyword cluster. This allows us to pull out descriptors that represent the domain better, but that may not have been used frequently. For example, the descriptor 'medicare prescription bill' on August 19, 2009 from the state of Washington, had a lower thematic importance, but this was enhanced by considering the relevance of medicare to healthcare and healthcare reform.
B. Temporal Importance of an event descriptor: While the thematic scores are good indicators of what is important in a spatio-temporal setting, certain descriptors tend to dominate discussions. In order to allow for less popular, possibly interesting descriptors to surface, we discount the thematic score of a descriptor depending on how popular it has been in the recent past. The temporal discount score for a n-gram, a tuneable factor depending on the nature of the event is calculated over a period of time.
C. Spatial Importance of an event descriptor: We also discount the importance of a descriptor based on its occurrence in other spatio-temporal sets. The intuition is that descriptors that occur all over the world on a given day are not as interesting compared to those that occur only in the spatio-temporal set of interest. We define the spatial discount score for an n-gram as a fraction of spatial sets or partitions (e.g. countries) that had activity surrounding this descriptor.

We use these three intuitions to extract key phrases and words that summarize social perceptions behind tweets.

5 User Interface and Visualization

The primary objective of the Twitris user interface is to integrate the results of the data analysis (extracted descriptors and surrounding discussions) with emerging visualization paradigms to facilitate *sensemaking*. The Twitris user interface facilitates effective browsing of the *when*, *where*, and *what* slices of social perceptions behind an event.

Figure 1 illustrates the theme, time and space components of the interface. To start browsing, users are required to select an event. The event selection widget (not shown due to space considerations) organizes events into categories and subcategories for easier navigation. Once a theme is chosen by the user, the date is set to the earliest date of recorded observations for an event and the map is overlaid with markers indicating the spatial locations from which observations were made on that date (marker 1). We call this the spatio-temporal slice. Users can further explore activity in a particular space by clicking on the

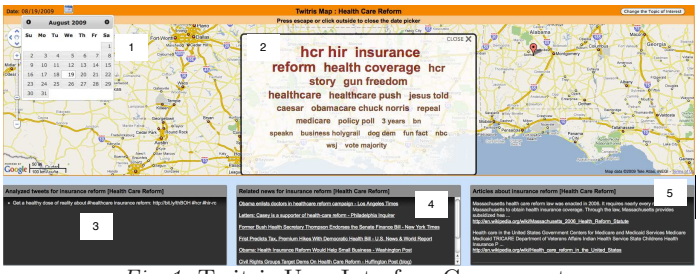

Fig. 1: Twitris User Interface Components

overlay marker. The event descriptors extracted from observations in this spatio-temporal setting are displayed as a tag cloud (marker 2). The current version of Twitris displays the top 15 descriptors weighted by their spatio-temporal-thematic (STT) scores. The STT scores determine the size of the descriptor in the tag cloud, illustrated in Figure 1(b).

When users select a particular tag from the tag cloud, we query the Web to find Wikipedia articles relevant to the selected tag. The purpose of finding Wikipedia articles is to identify a resource in DBPEdia. Once the resource is identified in DBPedia, we use the selected tag, related DBPedia resources along with the current spatial and thematic contexts to fetch relevant news articles. Related resources in DBPedia are identified by exploring pre-determined relationships with a fixed path length of 1.

The alpha version of Twitris can be accessed at http://twitris.dooduh.com.

6 Discussion and Conclusion

This work is a first step in the spatio-temporal-thematic integration of citizen-sensor observations. We presented our system Twitris, one possible approach for processing and presenting crowd-sourced, event related data in its naturally occuring spatio-temporal-thematic contexts. Our entity-driven approach allowed us to cull meaningful units of social perceptions and explore how their discussions varied across space and time. We posit that such crowd-sourced summaries can supplement situation awareness and decision-making applications. In Twitris, we also have the capability to "*mashup*" related content from external resources such as news, video streams and Wikipedia.

Twitris is developed using PHP. We have used Virtuoso SPARQL end point and along MySQL as for data storage and querying. The front is built using JQuery. We have used third party services provided by Twitter, Yahoo! BOSS, Google News and Wikipedia. So far we have processed more than two million tweets, for various events including the healthcare debate, iran election, snow leopard operation system release and the recent tsunami. Twitris currently has the healthcare debate event plugged in, and we are working to integrate more events before the challenge.

Twitris currently uses DBPedia. In our subsequent releases, we intend to incorporate Geonames for geographical clustering, FOAF to identify and highlight

persons (such as sportsmen, senators) in our tag collection, as well as richer domain models created by our hierarchy generation tool described in [4]. We do realize that Twitris is not a perfect system and does contain tags that are not relevant. We are working on improving our text analysis algorithms to create a robust approach to address these challenges[6].

References

1. Klein, G., Moon, B., Hoffman, R.: Making sense of sensemaking 1: alternative perspectives. IEEE Intelligent Systems **21**(4) (2006) 70–73
2. Nagarajan, M., Gomadam, K., Sheth, A.P., Ranabahu, A., Jadhav, A., Mutharaju, R.: Spatio-temporal-thematic analysis of citizen-sensor data - challenges and experiences. In: Web Information Systems Engineering. (2009)
3. Turney, P.: Extraction of keyphrases from text: Evaluation of four algorithms. Technical report, National Research Council, Institute for Information Technology (1997)
4. Thomas, C., Mehra, P., Brooks, R., Sheth, A.P.: Growing fields of interest - using an expand and reduce strategy for domain model extraction. In: Web Intelligence. (2008) 496–502

[6] Twitris is constantly evolving and something will be better everyday. We invite the reader to use the system and share their thoughts with us.