Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences

Meenakshi Nagarajan, Karthik Gomadam, Amit P. Sheth, Ajith Ranabahu, Raghava Mutharaju, Ashutosh Jadhav

Knoesis Center, Wright State University, Dayton, OH, USA.
{*meena*, *karthik*, *amit*, *ajith*, *raghava*, *ashutosh*}@knoesis.org

Abstract. We present work in the spatio-temporal-thematic analysis of citizen-sensor observations pertaining to real-world events. Using Twitter as a platform for obtaining crowd-sourced observations, we explore the interplay between these 3 dimensions in extracting insightful summaries of social perceptions behind events. We present our experiences in building a web mashup application, *Twitris*[1] that extracts and facilitates the spatio-temporal-thematic exploration of event descriptor summaries.

1 Introduction

The emergence of mircoblogging platforms like Twitter, friendfeed etc. have revolutionized how unfiltered, real-time information is disseminated and consumed by citizens. An important 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" [1].A significant portion of information generated and consumed by this interconnected network of participatory citizens is *experiential* in nature [2], i.e., contains first-hand observations, experiences, opinions made in the form of texts, images, audio or video about *real-world events*. In the recent past, such experiential attributes of an event have proved valuable for crowdsourced situational awareness applications.An example of this are observations that originated from Mumbai during the 2008 terrorist attacks. The relayed multimodal observations in the form of 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 occurence 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 analysis of such data. The central thesis behind this work is that citizen sensor

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

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.

Such an n-dimensional analysis is analogous to the processing of social stream, newswire or blog data where thematic, temporal, spatial and social aspects of the data have been taken into account in the past. In [3], the goal was to extract events from social text streams taking content, social, and temporal aspects in account. An event in their work is a set of text pieces (topically clustered) conditioned by social actors that talk about the same topic over a certain time interval with similar information flow patterns. Work in [4] attempts to identify spatiotemporal theme patterns in blog data. They extract common themes (semantically coherent topics defined over a text collection) from weblogs and subsequently generate theme life cycles for each given location and theme snapshot for each given time period. [5] use a graph-theoretic approach to discover storylines or latent themes among the top search results for a query.

In our work, we do not attempt to identify available latent themes, storylines or events in a given corpus of text. We start with a corpus of observations pertinent to an event and attempt to extract meaningful units that are good descriptors of the underlying event. We take an *entity-driven approach*, as opposed to a document collection approach in past efforts, to summarize social perceptions in citizen observations. We also do not concern ourselves with the social aspect or attributes of the poster, since our goal is to facilitate summaries for situation awareness applications that care more about "knowing what is going on so you can figure out what to do" [6].

1.1 Contributions

Our work is motivated by the need to easily assess local and global social perceptions behind 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. When observing what India is saying about the Mumbai attack, one might wish to not be biased by global and possibily contrasting perceptions from Pakistan, as an example. The larger goal of our ongoing work is to perform a spatial, temporal and thematic integration of citizen sensor observations. In this paper, we present the first step in this direction - analyzing data in these three dimensions to study what constitutes good spatial, temporal and thematic slices of observations underlying events.

Using Twitter as our platform for observations, we find that the confluence of space, time and theme in analyzing tweets allows us to extract insightful summaries of citizen perceptions behind events. Of the many analysis that are possible over Twitter data and available metadata for extracting social perceptions, we conduct the following investigations in our work:
1. What is a region paying attention to today? Our first goal is to extract meaningful descriptors or entities, i.e. *key words and phrases*, from mass citizen observations pertaining to an event for any spatial and temporal setting. Selecting discriminatory keywords has been a problem of historical importance with prob-

ability distribution methods like TFIDF being the most popular [7]. In our work, cues for a descriptor's importance are found in a corpus, in space and time. Consider this scenario where two descriptors 'mumbai attacks' and 'hawala funding' that occured in the Tweets[2] pertaining to the *Mumbai Terror Attack* event on the same day in the US. 'Mumbai attacks' occurred every day the last week while 'hawala funding' is a new descriptor for today. Users are more likely to be interested in perspectives and experiences that are different from that of yesterday's. Looking at spatial contexts, we also find that 'hawala funding' did not appear in any other country on the same day, while 'mumbai attacks' occured in almost all countries that day. This suggests that the discussion around 'hawala funding' is a perspective shared by citizens local to this spatial setting while 'mumbai attacks' is a weaker descriptor in terms of uniqueness to the local region. Our algorithm exploits this *interplay between space, theme and time* in order to cull out words and phrases that best summarize citizen observations.
2. What are they saying about the entity or descriptor? Since the social perception of an event may vary within and between spatial regions and temporal settings, there is a need to group and understand the context of discussion or storylines surrounding a descriptor. Using well understood principles of information theory, we extract an entity's strong thematic context i.e. strongly associated descriptors, while taking into account its spatial and temporal settings. Figure 4(a) shows an example of discussions surrounding two event descriptors, in different countries on the same day that we were able to extract.

Our approach to presenting extracted descriptors and surrounding adopts the interface design paradigm of *experience design*[3]. One of the goals of experience design is to consider the multiple contexts surrounding the use of an application and create unified user interaction models across all contexts. Our challenge was to create a visualization model that allows users to browse thematic descriptors of events in their spatio-temporal contexts.

We present our approach for extracting and visualizing event descriptors as an implemented system, Twitris [1] (a portmanteau of Twitter and Tetris, for arranging activity in space, time and theme) that allows users to browse extracted summaries of citizen-sensor activity. We use citizen sensor observations made via Twitter during three different events.Ideally, evaluating our system would involve measuring the efficacy of our algorithms in extracting event descriptors and the effectiveness of our interface in summarizing user activity. Owing to space restrictions, we limit our discussion in this paper to only the description of the Twitris system. Evaluations will be made available in an extended version of this paper at [1].

In the rest of this paper, we present our challenges and experiences in obtaining close to real-time citizen observations from Twitter (section 2), processing them in space, theme and time (section 3) and presenting the extracted summaries within their multi-dimensional contexts (section 4).

[2] 140 character long messages posted by users on Twitter
[3] http://en.wikipedia.org/wiki/Experience_design

2 System Overview

Twitris is currently designed to
- Collect user posted tweets pertaining to an event from Twitter
- Process obtained tweets to extract key descriptors and surrounding discussions
- Present extracted summaries to users

The duration and intervals of data collection and processing are configured based on the event being analyzed. Figure 1 illustrates the various steps and services involved in data collection, analysis and visualization.

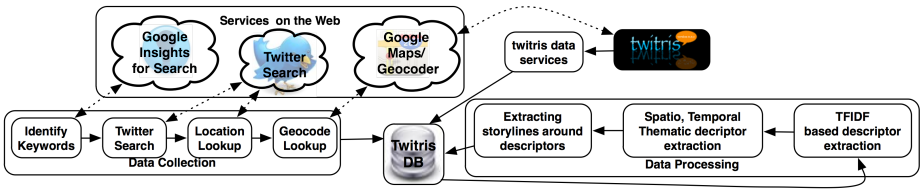

Fig. 1: Data Collection, analysis and visualizing in Twitris

Gathering Topically Relevant Data

The process of obtaining citizen observations from Twitter deserves some explanation since Twitter does not explicitly categorize user messages into topics. However, there is a search API[4] to extract tweets. A recent trend in Twitter has been the community-driven convention of adding additional context and metadata to tweets via *hashtags*, that can also be used to retrieve relevant tweets. Hashtags are similar to tags on Flickr, except they are added inline to a tweet. They are created simply by prefixing a word with a hash symbol, for example, users would tag a tweet about Madonna using the hashtag #madonna.

Our strategy for obtaining posts relevant to an event uses a set of seed keywords, their corresponding hashtags and the Twitter search API. Seed keywords are obtained via a semi-automatic process using Google Insights for Search[5], a free service from Google that provides top searched and trending keywords across specific regions, categories, time frames and properties. The intuition is that keywords with high search volumes indicate a greater level of social interest and therefore more likely to be used by posters on Twitter.

We start with a search term that is highly pertinent to an event and get top X keywords during a time period from Google Insights. For the g20 summit event for example, one could use the keyword g20 to obtain seed keywords. These keywords are manually verified for sufficient coverage for posts using the Twitter Search API, placed in set $\hat{K}$, and used to kick-start the data collection process. Past this step, the system automatically collects data every few hours. The list of keywords $\hat{K}$ is also continually updated using two heuristics:
1. The first uses Google Insights to periodically obtain new keywords using keywords in $\hat{K}$ as the starting query.
2. The second uses the corpus of tweets collected so far to detect popular key-

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

words that were not previously used for crawling. A keyword is considered to be a good data extractor if it has a high TFIDF score [7] and high collocation scores with the keywords in $\hat{K}$. The keyword with the highest score is periodically added to the set $\hat{K}$.

In this work, we collected data for three events - one long-running global financial crisis event and two short-lived events, the g20 summit and the Mumbai terror attack event. The nature of an event determines the strategy for data collection. For long-running events, data is collected on a regular basis but in longer intervals. Shorter events demand more frequent data collection and continuous update of keywords.

Spatial, Temporal and Thematic Attributes of Twitter Posts: The content of a Twitter post is the thematic component of a citizen observation. In this work, we ignore urls and links posted by users in a tweet and only use the textual component.

Spatial attributes for Twitter data can be of two types - location where the data originated from, and the location mentioned *in* the content. We do not concern ourselves with the latter since our goal is to study the social signals *originating* from a location in response to an event. There are two ways to obtain the spatial information associated with a tweet. The first method is to provide a location as a parameter to the search API. The other is to use the poster's location as an approximation for the origination of the tweet. We adopt the second alternative, as our crawl needs to be location independent.

The location information for an author either has geocoordinates (in cases where GPS enabled devices were used in accessing Twitter) or has a location descriptor (city, state or country) free-text information provided by posters in their profiles. In case of the former, we use the coordinate information as is, while in the latter, we make use of Google GeoCoding API [6] to identify the coordinates. We realize the limitations of this approach (for example, an author might have posted a tweet from Boston, but updated his location later), but given the lack of geocoding information in the tweets, we consider this approach as a sufficient approximation.

For this work, we collected nearly 310,000 tweets for the financial crisis event starting from Nov 22 2008, out of which we could get location information for nearly 160,000 tweets. Nearly 76% of this data was contributed by users in the US. We collected approximately 75,000 tweets for the g20 event between March 9, 2009 and April 10, 2009; 50,000 of which had location information. Majority of these tweets originated from the US or the UK (57% and 21% respectively). For the Mumbai terror attack, data was collected between November 29, 2008 and February 28th, 2009. We collected around 10,000 tweets, 6000 of which had location information. Over 70% of these tweets originated from the US and India (38% and 34% respectively).

The temporal information for each tweet is obtained from the time the tweet was posted (available via the API). Since we are interested in social signals over

[6] http://code.google.com/apis/maps/documentation/geocoding/index.html

time, we do not concern ourselves with identifying temporal information that might be available *in* the content of a tweet.

We model a tweet t as a 4-tuple; $t = \{t_{id}, t_c, t_t, t_g\}$ where t_{id} is a unique alpha-numeric identifier, t_c is the textual content, t_t and t_g are the time and geographical coordinates obtained for the tweet. $t_g = \{lat, lng\}$ where lat is the latitude and lng is the longitude of the geographical coordinates of t_g.

3 Processing Citizen Observations

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. 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 2). 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 [8]. 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 [7]. 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 ('general' and 'general motors') 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 'Big 3' meant to refer to the three car giants 'GM', 'Ford' and 'Chrysler' was not used as frequently as the three words together or vice versa. The presence of contextually relevant words should ideally

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

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 'Ford' is not in the context of the 'Big Three', i.e. discussions around Ford surround its new 'Ford Focus' model, its presence should not affect 'Big Three's' importance.

Contextually Enhanced Thematic Score: Here, we describe how the thematic score of an extracted descriptor, 'Big 3' in the above example, is amplified as a function of the importance of its strong associations - 'General Motors', 'Ford' and 'Chrysler' and the association strengths between the descriptor and the associations. For sake of brevity, let us call the $ngram_i$ descriptor whose thematic score we are interested in affecting as the focus word fw and its strong associations as C_{fw}=$\{aw_i$,$aw_2...\}$. The thematic score of the focus word is then enhanced as:

$$fw(th)=fw(tfidf)+\sum assoc_{str}(fw,aw_i)*aw_i(tfidf) \quad (1)$$

where $fw(tfidf)$ and $aw_i(tfidf)$ are the TFIDF scores of the focus and associated word as per Step 3 in the previous section; $assoc_{str}(fw,aw_i)$ is the association strength between the focus word and the associated word. Here we describe how we find strong associations for a focus word and compute $assoc_{str}$ scores. Our algorithm begins by first gathering all possible associations for fw and places it in C_{fw}. We define *associations* or the *context of a word* as thematically strong descriptors (in the top 5 n-grams of an observation) that co-occur with the focus word in the given spatio-temporal corpus. The goal is to amplify the score of the focus word only with the strongly associated words in C_{fw}. One way to measure strength of associations is to use word co-occurence frequencies in language [9]. Borrowing from past success in this area, we measure the association strength between the focus word and the associated words $assoc_{str}(fw,aw_i)$ using the notion of point-wise mutual information in terms of co-occurrence statistics. We measure $assoc_{str}$ scores as a function of the point-wise mutual information between the focus word and the *context of* aw_i. This is done to ensure that the association strengths are determined *in the contexts* that the descriptors occur in. Let us call the contexts for aw_i as Caw_i=$\{caw_1$,$caw_2..\}$, where caw_k's are thematically strong descriptors that collocate with aw_i. $assoc_{str}(fw,aw_i)$ is computed as:

$$assoc_{str}(fw,aw_i)=\frac{\sum_k(pmi(fw,caw_k))}{|Caw_i|},\forall caw_k \in Caw_i$$

where the point-wise mutual information between fw and caw_k (the context of aw_i), is calculated as:

$$pmi(fw,caw_k)=log\frac{p(fw,caw_k)}{p(fw)p(caw_k)}=log\frac{p(caw_k|fw)}{p(caw_k)} \quad (2)$$

where $p(fw)=\frac{n(fw)}{N}$;$p(caw_k|fw)=\frac{n(caw_k,fw)}{n(fw)}$; $n(fw)$ is the frequency of the focus word; $n(caw_k,fw)$ is the co-occurrence count of words caw_k and fw; and N is the number of tokens. All statistics are computed with respect to the corpus defined by the spatio-temporal setting. As we can see, this score is not symmetric and if the context of aw_i is poorly associated with fw, $assoc_{str}(fw,aw_i)$ is a low score.

At the end of evaluating all associations in C_{fw}, we pick those descriptors whose association scores are greater than the average association scores of the

mumbai	1.4553	pakistan pres promised	1.0065
photographers capture images	1.3998	mumbai attacks	0.9594
images of mumbai	1.2792	foreign relations	0.9490
foreign relations perspective	1.2165	rejected evidence	0.8741
attacks in mumbai	1.1261	evidence provided	0.8741
photographers capture	1.0986	uk indicating	0.8741
capture images	1.0986	mumbai attacks in	0.7927
india prime minister	1.0839	rejected evidence provided	0.7916
country of india	1.0280		

Event descriptors sorted by their TFIDF scores

foreign relations perspective	1.7185	photographers capture images	1.3028
india prime minister	1.5853	rejected evidence provided	1.2933
country of india	1.5295	mumbai attacks	1.2048
pakistan pres promised	1.5080	images of mumbai	1.1822
foreign relations	1.4510	mumbai	1.1083
rejected evidence	1.3758	mumbai attacks in	1.0797
evidence provided	1.3758	photographers capture	1.0017
uk indicating	1.3758	capture images	1.0017
attacks in mumbai	1.3293		

Event descriptors sorted by their enhanced spatio-temporal-thematic scores

(a)

Day1	Day2	Day3	Day4	Day5
world india blasts	november 2008 donation	liking	mumbai outfit bjp	foreign relations perspective
blasts pakistani denial	manmohan singh	terror outfit	terror backlash	india prime minister
india blasts pakistani	solution india	work	india network	country of india
world india	2008 donation page	teachings	paki gov	pakistan pres promised
attack mahal	donation page provided	assisting terror outfits	punjabi taliban	foreign relations
denial on terrorism	newspakistan seeks	mafia assisting terror	obama aide	rejected evidence
hotels scarred	closure	statesman terror mafia	backlash	evidence provided
month attack mahal	rtt news voice	india expert	cripple	uk indicating
terror hotels	long term damage	terrorist attacks awaken	earthquake dia plane	attacks in mumbai
nri family	economy manmohan singh	teachings exposed	strife china earthquake	photographers capture images
alleged terror attacker	attacks in mumbai	attacks awaken bollywood	terrorism pakistan	rejected evidence provided
a month attack	quickfix solution india	film stars dictate	thing terrorism	mumbai attacks
defiant terror inches	seeks closure	terror mafia assisting	war on terror	images of mumbai
terror attacker seeks	terror attack	revision of history	a mumbai outfit	mumbai
newsnri family drawn	manmohan singh rtt	efforts asks china	attack punjabi taliban	mumbai attacks in

(b)

Fig. 2: (a) Extracted descriptors sorted by TFIDF vs. spatio-temporal-thematic scores (b) Top 15 extracted descriptors in the US for Mumbai attack event across 5 days

focus word and all associations in C_{fw}. The thematic weights of these associations along with their strengths are plugged into Eqn 1 to compute the enhanced thematic score $ngram_i(th)$, of the n-gram descriptor.

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 as:

$$ngram_i(te) = temporal_{bias} * \sum_{d=1}^{D} \frac{ngram_i(th)^d}{d}$$

where $ngram_i(th)^d$ is the enhanced thematic score of the descriptor on day d, D is the duration for which we wish to apply the dampening factor, for example, the recent week. However, this temporal discount might not be relevant for all applications. For this reason, we also apply a $temporal_{bias}$ weight ranging from 0 to 1 - a weight closer to 1 gives more importance, while a weight closer to 0 gives lesser importance to past activity.

C. Spatial Importance of an event descriptor: We also discount the importance of a descriptor based on its occurence 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.

$$ngram_i(sp) = \frac{k}{|spatio-temporal sets|} * (1 - spatial_{bias})$$

where k = number of spatio-temporal sets the n-gram occured in. Similar to the temporal bias, we also introduce a $spatial_{bias}$ that gives importance to local vs. global activity for the descriptor on a scale of 0 to 1. A weight closer to 1 does not give importance to the global spatial discount while a weight closer to 0 gives a lot of importance to the global presence of the descriptor.

Depending on the event of interest, both these discounting factors can also vary for different spatio-temporal sets. For example, when processing tweets from India for the Mumbai attack setting the $spatial_{bias}$ to 1 eliminates the influence of global social signals. While processing tweets from the US, one might want a stronger global bias given that the event did not originate there. Both these parameters are set before we begin the processing of observations.

Finally, the spatial and temporal effects are discounted from the final score, making the final spatio-temporal-thematic (STT) weight of the n-gram as

$$w_i = ngram_i(th) - ngram_i(te) - ngram_i(sp) \tag{3}$$

Figure 2(a) illustrates the effect of our enhanced STT weights for extracted event descriptors pertaining to the Mumbai terror attack event, in the US on a particular day. We used a temporal bias of 1 suggesting that past activity was important and a spatial bias of 0 giving importance to the global presence of the descriptor. As we see, descriptors generic to other spatial and temporal settings (e.g., mumbai and mumbai attacks) get weighted lower, allowing the more interesting ones to surface higher.

Figure 2(b) shows top 15 extracted descriptors in the US across five days (days that had atleast three citizen observations). As we see, the descriptors extracted by our system offer a good indication of what is being talked about on those days. In an ongoing user study, we are showing users tweets on any given day and investigating how useful descriptors extracted by our system are compared to those generated using the TFIDF baseline. Results of the same will be made available at [1].

3. Discussions around Event Descriptors

While it is useful to know what entities people are talking about, there might be different storylines surrounding these entities that could offer an insight into the social perceptions of an event. The goal here is to thematically group discussions surrounding event descriptors, while also allowing users to observe how these discussions change over time and space. We take a simple clustering approach to this problem, forming k clusters, each representing a viewpoint or storyline within a spatio-temporal setting. While this is similar in spirit to clustering of documents to reveal storylines as presented in [5], we use a mutual information based approach.

Let us call the n-gram of interest as the *focus word* fw. The steps involved in identifying storylines surrounding fw are the following (see Figure 3):

1. As in our previous algorithm, we find all associations for a focus word; $C_{fw}=\{aw_i, aw_2 ...\}$, i.e. thematically strong descriptors that collocate with the fw in the given spatio-temporal corpus.
2. In order to pick cues for complementary viewpoints, we pick n associations

from C_{fw} such that $n<|C_{fw}|$ and all n associations are weakly associated with each other (lending support for separate threads of discussions). Weak associations are indicated by negative pmi scores (see computed association strengths in the earlier section). As before, association strengths are computed only from the underlying corpus of tweets in a spatio-temporal setting in order to preserve observed signals. Figure 3 shows an example where 'Pakistan' is the focus word and the 2 associations offering cues for separate storylines are 'declared terror state' and 'pm gilani'.

3. For each of the n associations, we create a cluster populated with a pair of words - the focus word and the association (see bullet 2 in Figure 3). The association is further removed from C_{fw}. The idea is to expand each cluster progressively by adding strongly associated descriptors from C_{fw}. Descriptors are added to a cluster if they result in a positive change in the *Information Content* of the cluster [9], i.e. increase the amount of information that was present in the cluster. Creating word clusters using association strengths have been used in the past for assigning words to syntactic and semantic categories, learning language models and so on.

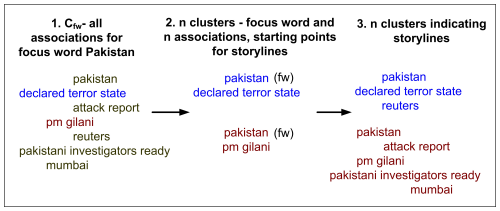

Fig. 3: Extracting discussions around descriptors

The next step is to expand each of the n clusters. Let us refer to the cluster n_i with the focus word fw and one association word as $C1$ and the associations for fw, C_{fw} as $C2$. The idea is to gradually expand $C1$ by adding keywords from $C2$ that are strongly associated with $C1$. At every iteration, the algorithm measures the change in Information Content (IC) of $C1$, $IC(C1,k_i)_\delta$, before and after adding every descriptor k_i from $C2$ to $C1$ as:

$$IC(C1,k_i)_\delta = IC(C1,k_i) - IC(C1) \tag{4}$$

where $IC(C1,k_i)$ is the information content of $C1$ after adding keyword k_i from $C2$. $IC(C1,k_i)_\delta$ is *positive* when k_i is *strongly associated* with words in $C1$ and *negative* when k_i is *unrelated* to words in $C1$. $IC(C1)$ is the strength of the semantic associations between words in the cluster and is defined as the average pairwise Mutual Information (MI) of the words.

$$IC(C1) = MI(C1)\binom{|C1|}{2} \tag{5}$$

where $|C1|$ denotes the cardinality of the cluster $C1$ and $\binom{|C1|}{2}$ is the number of word pairs in the cluster $C1$, normalizing for clusters of different sizes. $MI(C1)$ is the Mutual Information of cluster $C1$, defined as the sum of pairwise mutual information of words within the cluster.

$$MI(C1) = \sum_{w_i,w_j \in C1, i \neq j} PMI(w_i,w_j) \tag{6}$$

where $PMI(w_i,w_j)$ is reused from 2. The descriptor k_i from $C2$ that results in a positive and minimum $IC(C1,k_i)_\delta$ score is added to $C1$ and removed from $C2$. Additionally, keywords resulting in negative $IC(C1,k_i)_\delta$ scores are discarded as weak associations. The algorithm terminates when all keywords in $C2$ have been evaluated or when no more keywords in $C2$ have positive $IC(C1,k_i)_\delta$ scores (no strong associations with $C1$). All co-occurence statistics are obtained only from the underlying corpus of tweets in a spatio-temporal setting in order to preserve observed signals.

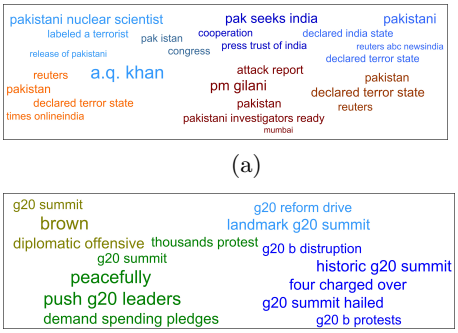

Fig. 4: (a) Discussions surrounding focus word "Pakistan" in the US (shades of blue), India (orange) and Pakistan (shades of red) on a particular day (b) Discussions surrounding focus word "g20" in Denmark across 4 days shown in different colors

The reasoning behind picking the descriptor that offers a minimum delta as opposed to the maximum delta in Information Content is as follows. A keyword k_i occuring in specific contexts with words in $C1$ will increase the Information Content of the $C1$ relatively less than a keyword that occurs in generic contexts. This strategy has the tendency of adding specific to general keywords from $C2$ to $C1$. At the end of this process n clusters are populated with strongly associated descriptors from C_{fw}, with each cluster representing a viewpoint in terms of cohesive descriptors (see Figure 3). We note here that a descriptor can belong to more than one cluster. Figures 4(a) and 4(b) also provide examples showing viewpoints varying over space and time for two focus words, "Pakistan" and "g20". This view is different from what is available on Twitris and has been altered for presentation purposes.

Thematic Integration - Discussion: In this work, we do not attempt to reconcile descriptors that refer to the same real-world entity of interest i.e. we do not reconcile that 'pak istan' and 'pakistan' in Figure 4(a) are the same or that entity 'general motors' and 'gm' are the same. In our ongoing efforts we are using domain models culled from DBPedia [10] in addition to word-sense disambiguation techniques to disambiguate and annotate entity references. This will also allow us to thematically integrate citizen sensor observations.

4 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*. Sensemaking, defined in [11], is the understanding of connections between people, places and events. Awareness of *who, what, when and where* is a critical component in sensemaking. Attributes of *who* posted a tweet does not play a role in this work. The

Twitris user interface facilitates effective browsing of the *when*, *where*, and *what* slices of social perceptions behind an event. To achieve this, Twitris is built as a *mashup* Web application. Figure 5(a) illustrates the theme, time and space

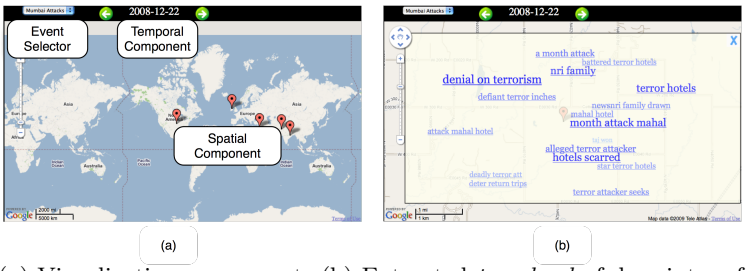

Fig. 5: (a) Visualization components (b) Extracted *tag cloud* of descriptors for USA

components of the interface. To start browsing, users are required to select an event from the start screen (not shown due to space considerations). Users have the option of changing the event from the dropdown, illustrated in the top left corner in figure 5(a). Once a theme is chosen, the time menu is set to the earliest date of recorded observations for an event and the map is overlayed with markers indicating the spatial locations from which observations were made on that date. We call this the spatio-temporal slice. Users can further explore activity in a particular space by clicking on the overlay marker. The event descriptors extracted from observations in this spatio-temporal setting are displayed as a tag cloud. 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 5(b).

At this stage, the descriptors serve as the focal point for further browsing and exploring of discusssions or storylines. On clicking a descriptor of interest, the user is shown discussions surrounding the descriptor on that day from all spatial regions (see sample in Figure 4(a)). We show all storylines on the same screen to allow users to contrast and compare complimentary discussions. The descriptors for these storylines are weighted by their STT scores. Subsequent interaction with any keyword leads the user to discussions surrounding the selected keyword. At any point in time, the user has the option of exiting this view and going back to the current spatio-temporal slice.

The alpha version of Twitris can be accessed at http://twitris.dooduh.com. Demos are available for two events, G20 and Mumbai Terror event attack. We employ a spatial parameter δ_s of country and temporal parameter δ_t of a date for processing these observations (see section 3).

Limitations and Improvements

While Twitris presents a new paradigm in browsing citizen sensor observations, the current version of the system has a few limitations, due to system pre-configuration. The first such limitation is the events for which browsing is currently supported. The second is the lack of social interaction centered around Twitris. We intend to address these in the next version (due late summer) by supporting user configuration of events. Users will be able to personalize events they

want to monitor based on keywords that kick-start the crawling and processing phases. We will also provide personalization, sharing and social interaction with Twitris using Facebook Connect. We also intend to enable retweet and reply actions via Twitris. The next version will also incorporate a calender for better navigation across dates and search capabilities over extracted descriptors.

5 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.

Few other prototypes similar to ours are available today. VoteIndiaReport[8] is one such example of a collaborative citizen-driven election monitoring platform for the 2009 Indian general elections. Besides situation awareness applications where people can track what a crowd is saying, other possible applications of our work include, search over real-time event related data; monitoring of citizen opinions and sentiments across spatial distributions; studying patterns in evolution of citizen perceptions behind events etc. There are several other exciting avenues for future work. Some of immediate interest to us include the semantic annotation of extracted descriptors in order to facilitate integration of citizen-sensor observations.

References

1. : Twitris: Twitter through space, time and theme. http://twitris.dooduh.com
2. Jain, R.: Experiential computing. Commun. ACM **46**(7) (2003) 48–55
3. Zhao, Q., Mitra, P., Chen, B.: Temporal and information flow based event detection from social text streams. In: AAAI. (2007) 1501–1506
4. Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and on-line event detection. In: SIGIR '98, New York, NY, USA, ACM (1998) 28–36
5. Kumar, R., Mahadevan, U., Sivakumar, D.: A graph-theoretic approach to extract storylines from search results. In: KDD. (2004) 216–225
6. Adam, E.: Fighter cockpits of the future. (Oct 1993) 318–323
7. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24**(5) (1988) 513–523
8. Turney, P.: Extraction of keyphrases from text: Evaluation of four algorithms. Technical report, National Research Council, Institute for Information Technology (1997)
9. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. In: Proceedings of the 27th annual meeting on ACL. (1989)
10. 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
11. Klein, G., Moon, B., Hoffman, R.: Making sense of sensemaking 1: alternative perspectives. IEEE Intelligent Systems **21**(4) (2006) 70–73

[8] http://votereport.in/