Spatio-Temporal-Thematic Analysis of Citizen Sensor Data

Challenges and Experiences

Meenakshi Nagarajan, Karthik Gomadam, Amit Sheth, Ajith Ranabahu, Raghava Mutharaju and Ashutosh Jadhav
Kno.e.sis, Wright State University

Presented at the conference by Pablo Mendes
Text at: http://knoesis.org/library/resource.php?id=00559

- Micro-blogging platforms – Twitter, Friendfeed..

- Revolutionizing how unaltered, real-time information is disseminated and consumed

- Significant portion of data is *Experiential* in nature

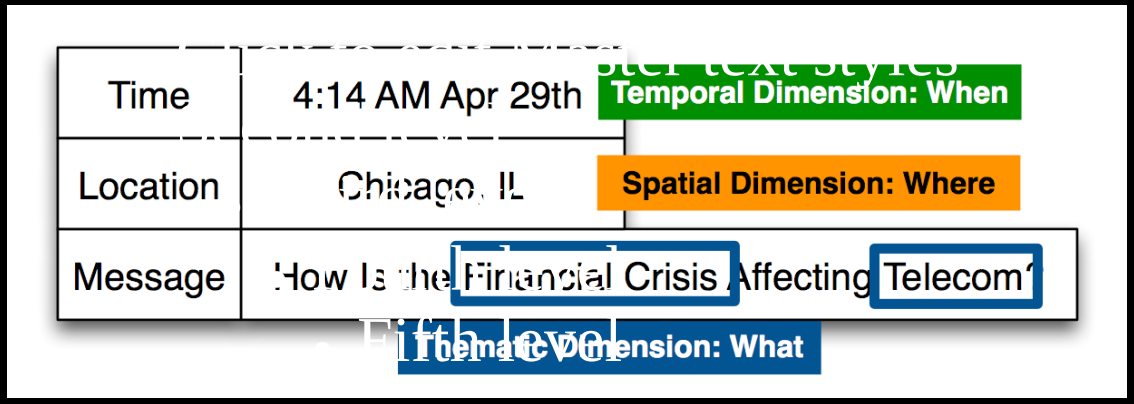
- First-hand observations, experiences, opinions via texts, images, audio, video (Citizens as sensors)

Citizen Sensor Observations

Are a lens into the social perception of an *event* in any *region* at any point in *time*

Mumbai Terror Attacks, Iran Elections, Obama's Health Care Reform…

They present complementary, sometimes lagged viewpoints that evolve over time and with other external stimuli



what is being said about an event (the theme),
IS AS IMPORTANT AS
where (spatial) and when (time) it is being said

Contribution, Presentation Focus

A Web MashUp that

Processes textual citizen sensor observations pertaining to real-world events

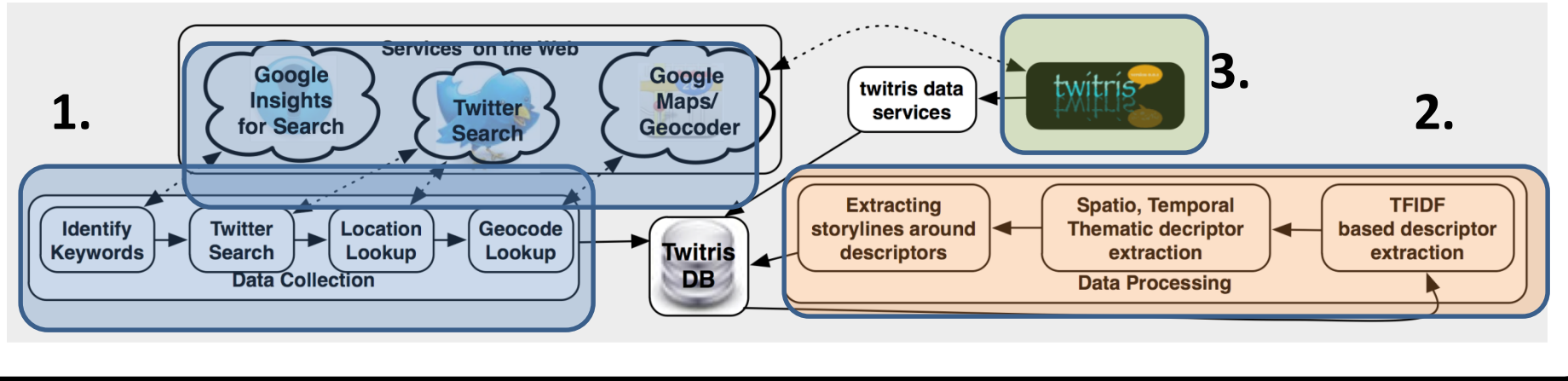
Takes three dimensions of space, time and theme into consideration

Extracts local and global social signals/ perceptions over time

Crawling, Processing, Visualization

TWITRIS – System overview

Twitris - System Overview



1. Obtain Topically Relevant Tweets, Extract Location, Time stamp Information, Store in DB

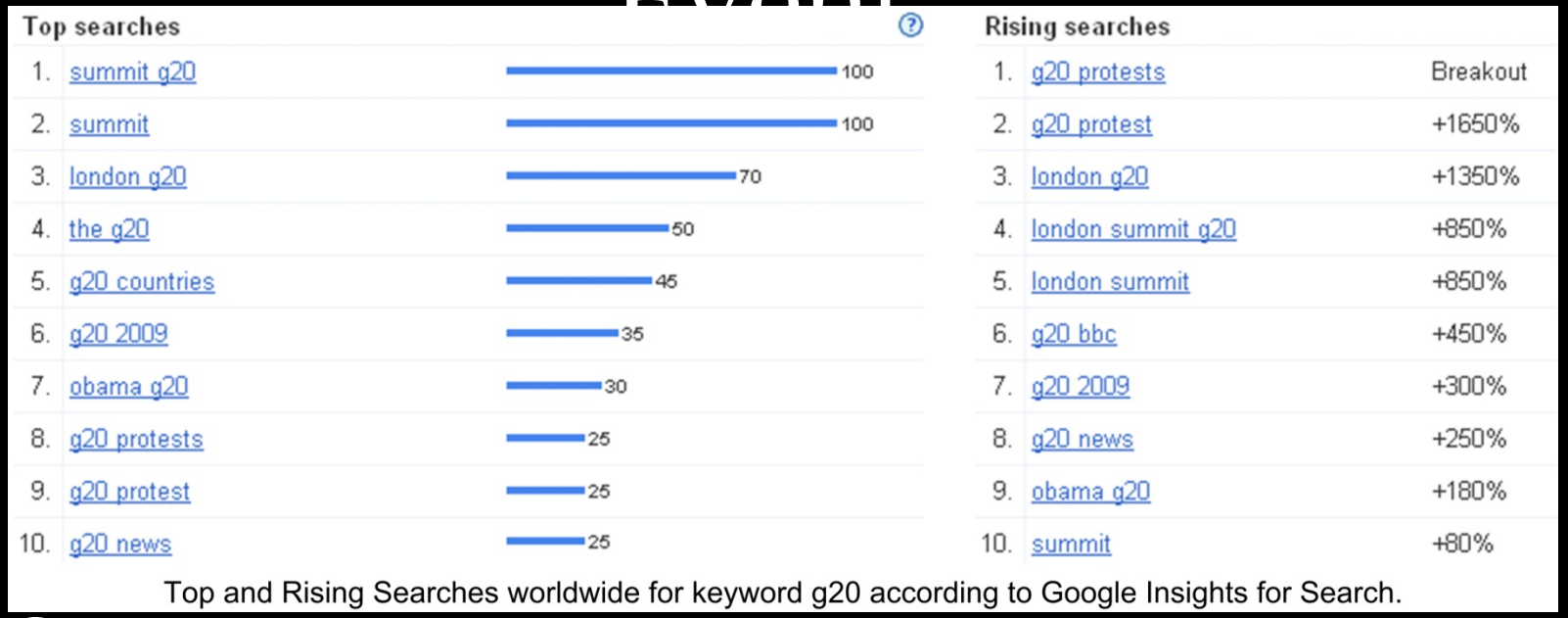
2. Process Tweet Contents, Store extracted metadata in DB

3. User Visualization talks to the DB

Obtaining Citizen-Sensor Observations

1. Gathering topically relevant data, extracting Location, time stamps

Crawling Tweets Relevant to Event

- 

Top searches

1.	summit g20	100
2.	summit	100
3.	london g20	70
4.	the g20	50
5.	g20 countries	45
6.	g20 2009	35
7.	obama g20	30
8.	g20 protests	25
9.	g20 protest	25
10.	g20 news	25

Rising searches

1.	g20 protests	Breakout
2.	g20 protest	+1650%
3.	london g20	+1350%
4.	london summit g20	+850%
5.	london summit	+850%
6.	g20 bbc	+450%
7.	g20 2009	+300%
8.	g20 news	+250%
9.	obama g20	+180%
10.	summit	+80%

Top and Rising Searches worldwide for keyword g20 according to Google Insights for Search.

- Strategy
  - Start with manually selected keywords (seed)
  - Obtain additional hints from Google Insights
  - Crawl using keywords, hashtags

Crawling Tweets Relevant to Event

- Events change, Topics of discussions change
- Periodically update keywords used for crawl
  - Process crawled tweets, extract top 1 TFIDF keyword, obtain Google Insights Suggestions
- Continue crawl

Challenges, Limitations of Crawl

Volume and Rapidity of Change, Quality of data is key

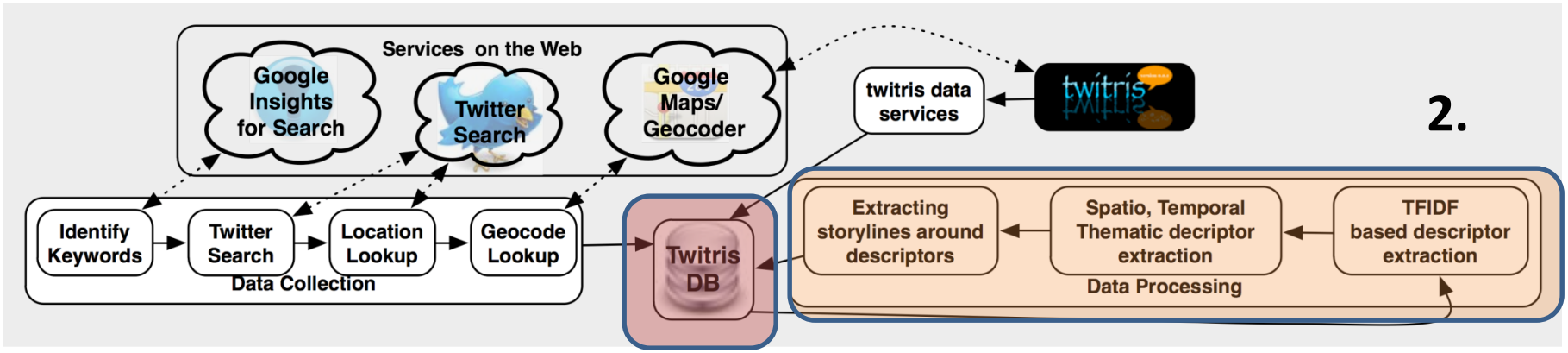
- Keyword gathering, Crawl requires supervision

Twitter API restrictions

- Hourly / Daily access limits
- Severe limitations on extracting past data
- Can go back only a few weeks!

Large scale geo-code conversions

Twitris - System Overview



Extracted time stamps, geo-coordinates stored in the DB

2. Process Tweet Contents, Store extracted metadata in DB

Spatio-Temporal Sets of Tweets

Intuition behind processing of tweets

Events have inherent spatial temporal biases associated with them

Bias dictates granularity of data processing

- Mumbai terror attack: country level activity everyday
- Health care reform: US state level activity per week

Spatial Temporal Sets

Group observations using spatial, temporal bias cues

E.g., for Mumbai Terror Attack, create X sets of tweets per day, each cluster represents activity in one country

Thematic processing over each set

Ensures local, temporal social signals are preserved

Processing Tweets

Extracting important event **descriptors**

Key words or phrases (n-grams)

What is a region paying attention to today?

Objective: from volumes of tweets to key descriptors

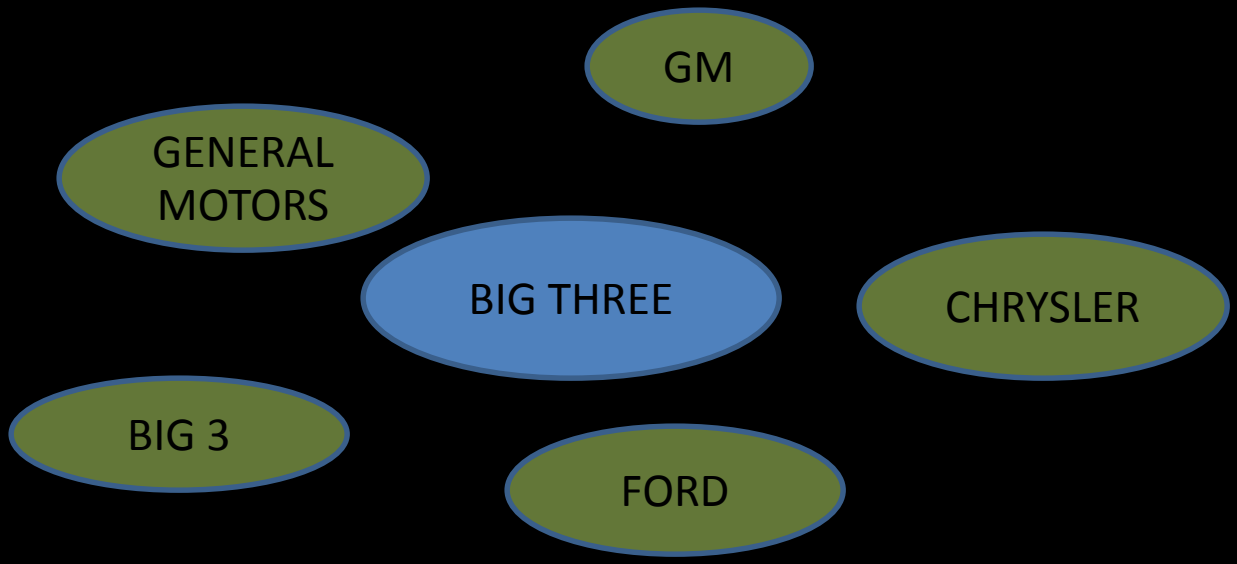
Using three attributes: Thematic, Temporal and Spatial importance of a descriptor

Descriptor: *Thematic* Importance

TFIDF weighted 3-grams

Amplified if nouns, no stop words

Amplified by presence of contextual evidence



Descriptor: Thematic-*Temporal* Importance

Certain descriptors always dominate observations
Terrorism in the Mumbai Terror Attack
Healthcare in the Health Care Reform discussions

To allow less popular, interesting descriptors to surface, we discount thematic score proportional to recent popularity

Certain descriptors always dominate observations

Terrorism in the Mumbai Terror Attack

Healthcare in the Health Care Reform discussions

To allow less popular, interesting descriptors to surface, we discount thematic score proportional to recent popularity

Descriptor: Thematic-Temporal-*Spatial* Importance

Descriptors that occur all over the world not as interesting as those local to a region

Discount thematic-temporal score proportional to number of spatial sets (not local) that mention the descriptor

Descriptors that occur all over the world not as interesting as those local to a region

Discount thematic-temporal score proportional to number of spatial sets (not local) that mention the descriptor

TFIDF vs. Spatio-Temporal-Thematic (STT) Scores of Descriptors

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

Interesting descriptors surface up!

Other examples in the paper

Discussions around Descriptors

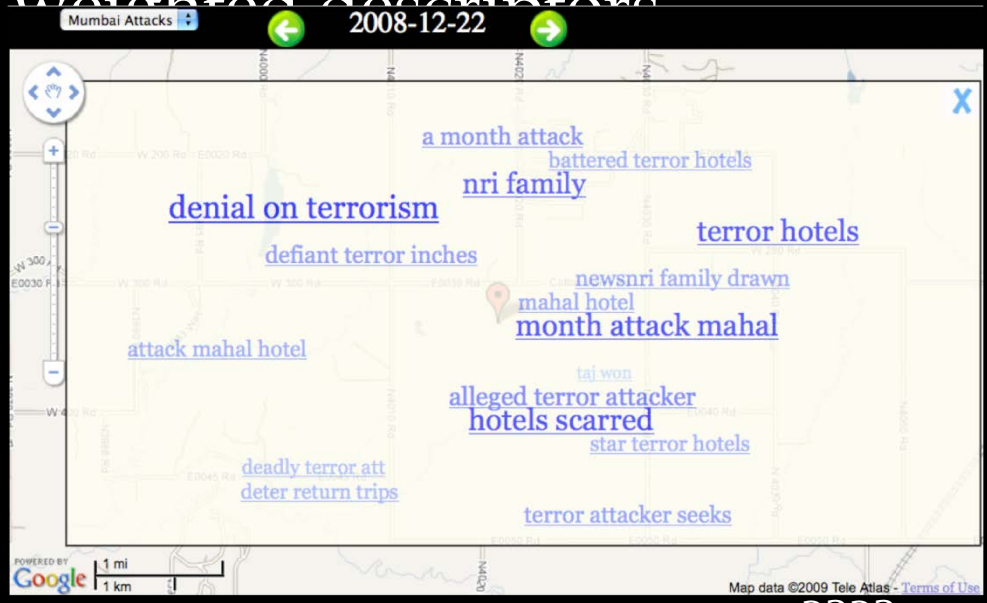
For some context : Extracting chatter surrounding a descriptor of interest

Using a clustering approach

Figure showing top X STT weighted descriptors

What are people saying about a descriptor?

(user click driven)



Clustering Algorithm Overview

For a descriptor of interest

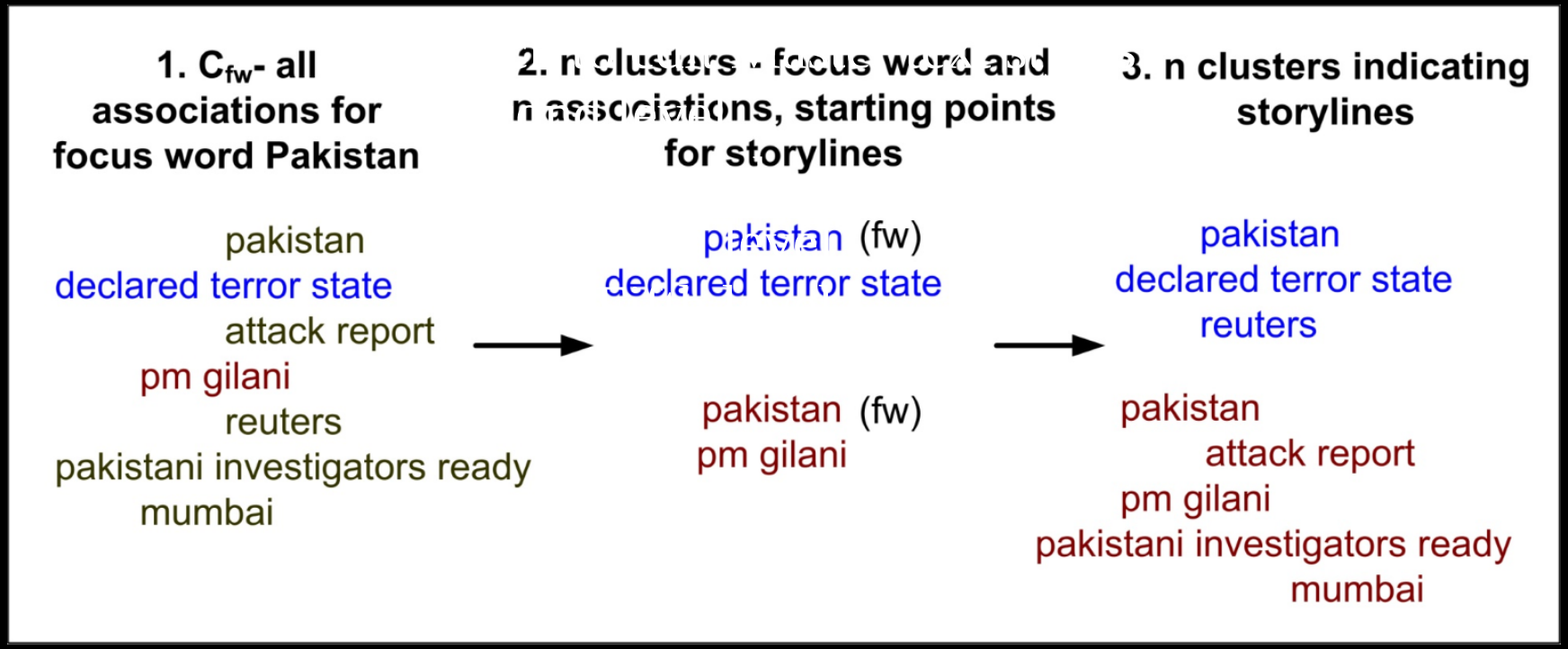
We generate complementary viewpoints expressed in the data

Using a Information Content based Clustering Algorithm

Basic Intuition

Among descriptor associations, select complementary viewpoint hints

Algorithm Overview – Example for focus word ‘Pakistan’



Discussions around Descriptors - Example

Around Pakistan on a particular day

US (shades of blue), India (orange), Pakistan (red)

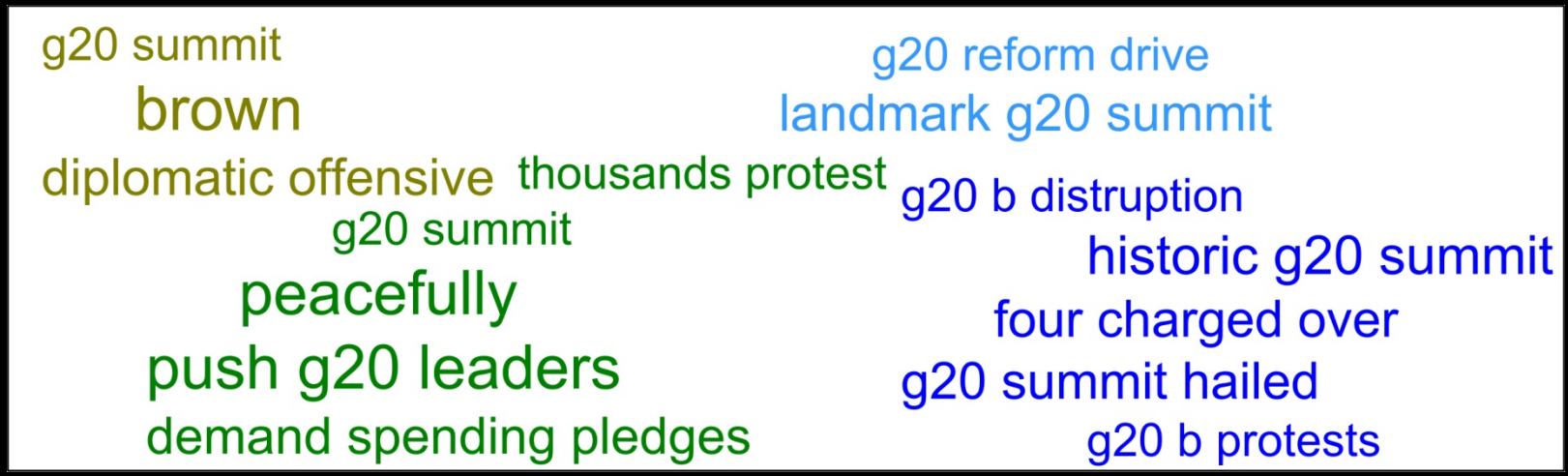
Size indicates STT score

This summarized viz only for presentation

Discussions around Descriptors - Example

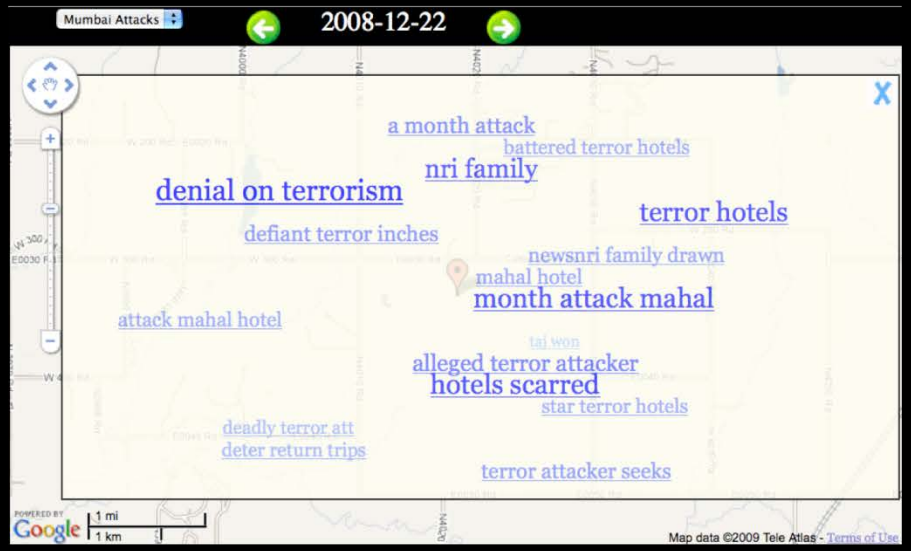
Around G20 in Denmark across 4 days (color)

Size indicates STT scores



User Interface and Visualizations

Browsing the *when*, *what* and *where* slices of social perceptions behind events



Events in Twitris

WISE 2009

Mumbai Terror Attack, G20

ISWC Challenge 2009

Health Care Reform, Iran Election

New features

Integration with news, Wikipedia, Tweets mentioning descriptors

Current Explorations, Investigations :

For more informatic
meena@knoesis.org
karthik@knoesis.org
amit@knoesis.org
ajith@knoesis.org

For more information

Try it on-line: http://twitris.knoesis.org
http://knoesis.org/research/semweb/projects/socialmedia/
meena@knoesis.org

karthik@knoesis.org

amit@knoesis.org

ajith@knoesis.org

Try it on-line: http://twitris.knoesis.org

http://knoesis.org/research/semweb/projects/socialmedia/

Development Team