

# SocialSensor: Sensing User Generated Input for Improved Media Discovery and Experience

Topic discovery in tweeter streams

**Dr. Yiannis Kompatsiaris, Project Coordinator**

International Workshop on Search Computing, Brussels, 25 September 2012

# Overview

- Motivation
- Objectives
- Architecture
- Use Cases and Requirements
  - News
  - Infotainment
- Research Activities and Results
  - Topic Discovery in Tweeter Streams
- Conclusions

# What is SocialSensor?

- 3-year FP7 European Integrated Project
  - <http://www.socialsensor.eu>
- Members: CERTH, ATC (Greece), Deutsche Welle, University Koblenz, Research Center for Artificial Intelligence (Germany), The City University London, Alcatel – Lucent Bell Labs, JCP Consult (France), University of Klagenfurt (Austria), IBM Israel, Yahoo Iberia
- 1 year into the project (Development of user requirements, use case scenarios, architecture study and implementation and first R&D prototypes)

# Motivation: Social Networks as Sensors

- Social Networks is a data source with an extremely dynamic nature that reflects events and the evolution of community focus (user's interests)
- Transform **individually rare** but **collectively frequent** media to meaningful topics, events, points of interest, emotional states and social connections
- Mine the data and their relations and exploit them in the right context
- Scalable mining and indexing approaches taking into account the content and social context of social networks



# Relevant Applications

Xin Jin, Andrew Gallagher, Liangliang Cao, Jiebo Luo, and Jiawei Han. *The wisdom of social multimedia: using flickr for prediction and forecast*, International conference on Multimedia (MM '10). ACM.

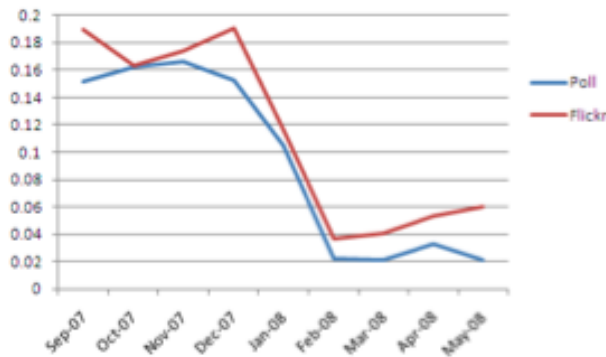
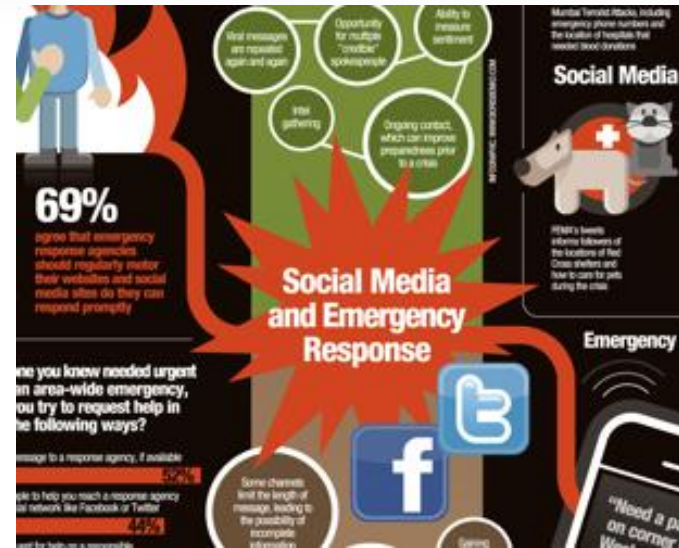
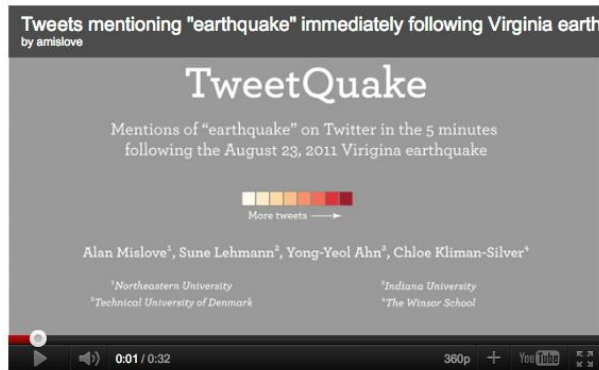


Figure 7: Reuters/Zogby Poll v.s. Flickr. Y-axis denotes the percentage of popularity for candidate Edwards.



Federal Emergency Management Agency *plans to engage the public* more in disaster response by sharing data and leveraging reports *from mobile phones and social media*

“...if you're more than 100 km away from the epicenter [of an earthquake] you can read about the quake on twitter before it hits you...”

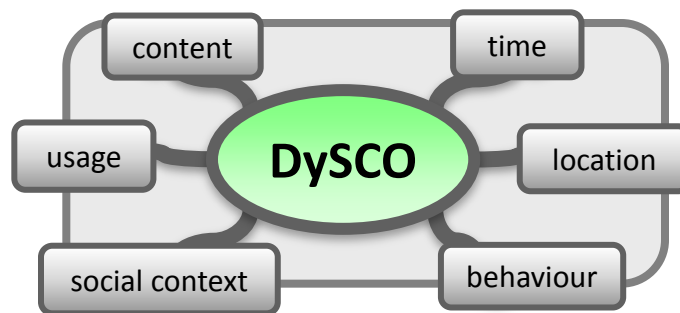
# Objective

SocialSensor quickly surfaces trusted and relevant material from social media – with context.



Massive social media and unstructured web

Social media mining  
Aggregation & indexing



Personalised access  
Ad-hoc P2P networks



News - Infotainment

# The SocialSensor Vision

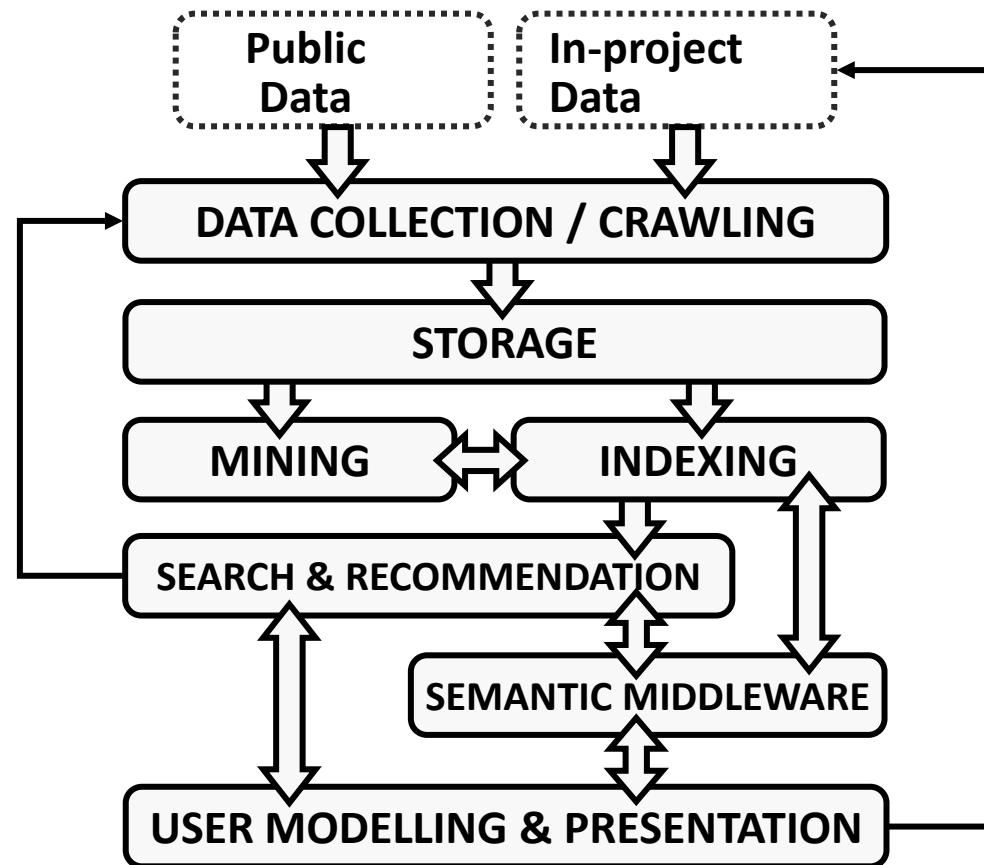
**SocialSensor quickly surfaces trusted and relevant material from social media – with context.**

- “quickly”: in real time
- “surfaces”: automatically discovers, clusters and searches
- “trusted”: automatic support in verification process
- “relevant”: to the users, personalized
- “material”: any material (text, image, audio, video = multimedia), aggregated with other sources (e.g. web)
- “social media”: across all relevant social media platforms
- “with context”: location, time, sentiment, influence



# Conceptual Architecture and Main components

- Real time dynamic topic and event clustering
- Trend, popularity and sentiment analysis
- Calculate trust/influence scores around people
- Personalized search, access & presentation based on social network interactions
- Semantic enrichment and discovery of services





*“Social media is transforming the way we do journalism”*  
(New York Times)


*“Social media is the key place for emerging stories – internationally, nationally, locally”* (BBC)

*“It has changed the way we do news”*(MSN)



Source: picture alliance / dpa





**“It’s really hard to find the nuggets of useful stuff in an ocean of content” (BBC)**

**“Things that aren’t relevant crowd out the content you are looking for” (MSN)**

**“The filters aren’t configurable enough” (CNN)**

Source: Getty



# Verification was simpler in the past...



Source: Frank Grätz

# An example: BBC Verification Procedure: Arab Spring Coverage

- Referencing locations against maps and existing images from, in particular, geo-located ones.
- Working with our colleagues in BBC Arabic and BBC Monitoring to ascertain that accents and language are correct for the location.
- Searching for the original source of the upload/sequences as an indicator of date.
- Examining weather reports and shadows to confirm that the conditions shown fit with the claimed date and time.
- Maintaining lists of previously verified material to act as reference for colleagues covering the stories.
- Checking scenery, weaponry, vehicles and licence plates against those known for the given country.

# Infotainment

- Events with large numbers of visitors
- Thessaloniki International Film Festival
  - 80,000 viewers / 100,000 visitors in 10 days
  - 150 films, 350 screenings
- Fete de la Musique Berlin
  - 100,000 visitors every year
- Discovery and presentation of relevant aggregated social media (e.g. film ratings from tweets)

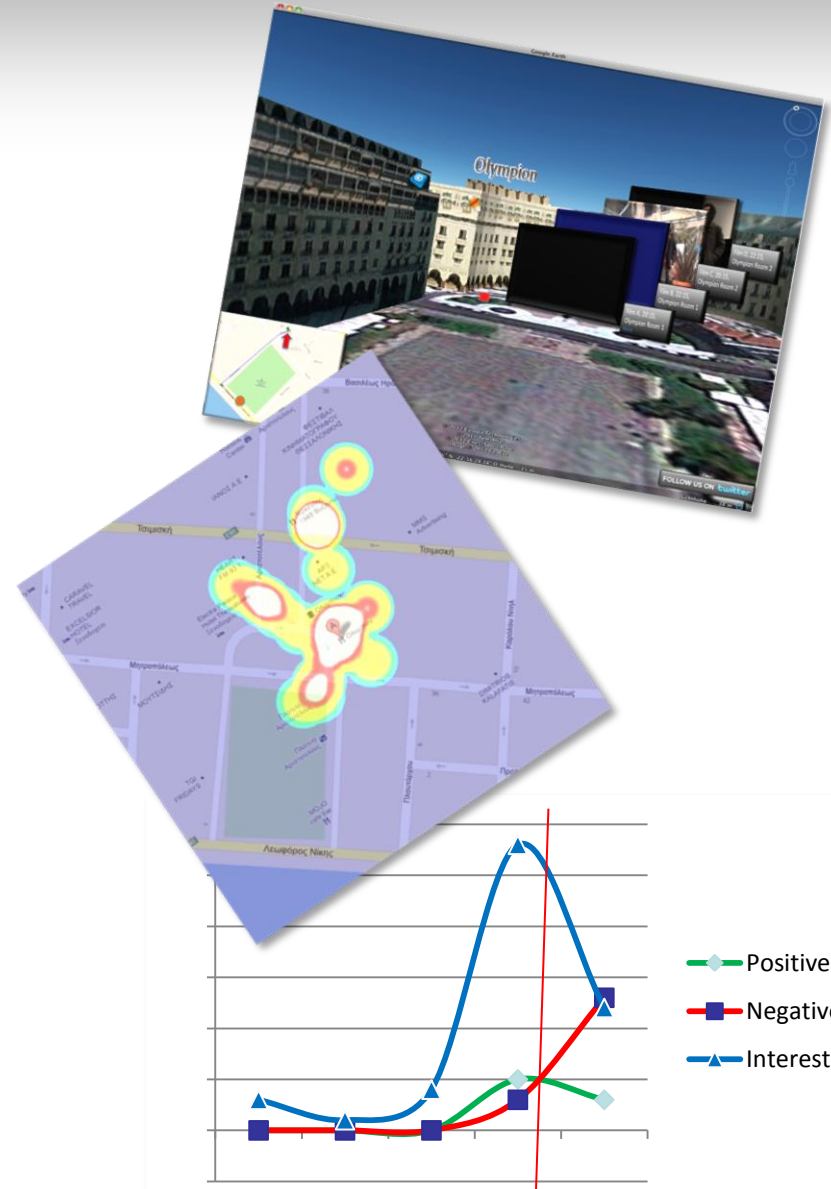


THESSALONIKI  
INTERNATIONAL  
FILM FESTIVAL  
[www.filmfestival.gr](http://www.filmfestival.gr)

Fête de la  
MUSIQUE  
21 JUIN

# User Requirements

- Social media aggregation
- Sentiment analysis for screenings and events
- Real-time check-in heatmaps
- AR 3D interfaces with RT information layers
- Social media-based film recommendations
- Smart location-based recommendations

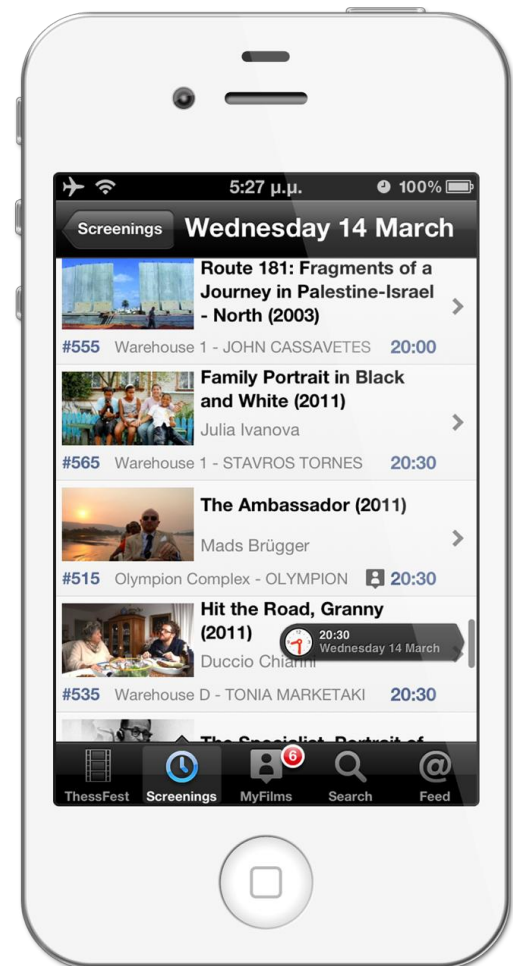
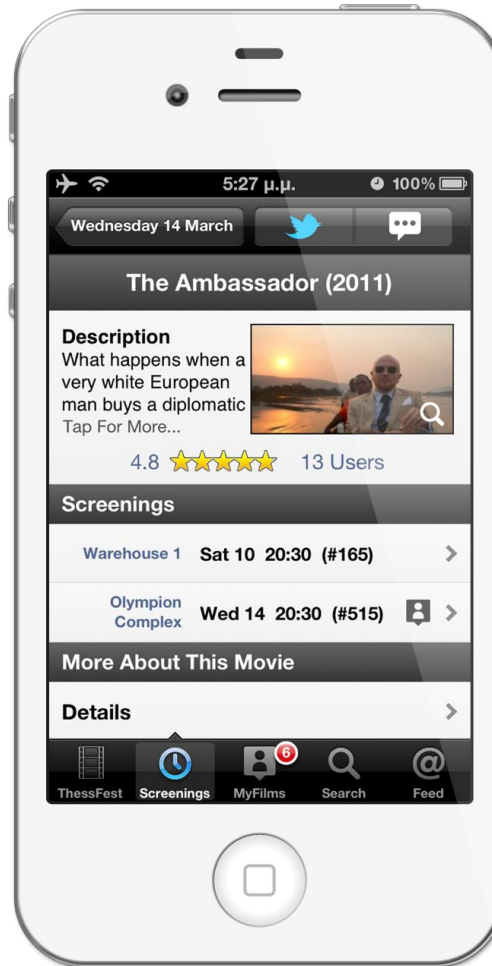




# ThessFest

- Thessaloniki International Film Festival
- Support twitter/comment usage within the app
- Ratings and comments per film
- Feedback aggregation
  - Votes
  - Tweets
- Real-time feedback to the organisation and visitors

ThessFest



# ThessFest

- Gather “realistic” user requirements
- Early showcase and evaluation of SocialSensor technologies in real-world event scale
- Engage users and create an informed user basis
- TDF14: 9-18 March 2012
  - 400+ users
  - 6500+ user sessions
  - positive response to social media
- Next version
  - Updated features
  - Android version

# Topic Discovery in Tweeter Streams

Giorgos Petkos, Symeon Papadopoulos, Yiannis Kompatsiaris,  
Carlos Martin, David Corney, Ryan Skraba, Luca Aiello, Alex Jaimes, Yosi Mass

# Problem Formulation

- Detect trending topics in a stream of Tweets
- Topics are represented as sets of keywords that capture the essence of an emerging news story
  - Ex: “ramires”, “goal”, “1-0”, “chelsea”, “score”  
in the case of a soccer match
- Due to evolving nature of social interest, **topics are extracted per timeslot** (e.g. per minute, hour, etc.)
  - in contrast to previous work on topic detection, we are not only interested in a posteriori detecting topics in a static corpus, but in **detecting them at the time they are discussed**

# Background

Two basic classes of methods to discover topics:

- **Feature-pivot (Feat-p):**
  1. Select important (trending) terms in a stream
  2. Cluster them with related terms into topics
- **Document-pivot (Doc-p):**
  1. Cluster documents (tweets) into groups based on similarity
  2. Extract most characteristic terms for each group

# Proposed Methods

- Doc-p based on Locality Sensitive Hashing (LSH)
  - Efficient stream clustering → cluster filtering → term selection
- Graph-based feat-p approach
  - Term selection → term graph creation based on co-occurrence → graph clustering
- Feat-p based on frequent pattern detection
  - Parallel FP-Growth algorithm (Hadoop implementation) → frequent pattern ranking
- “Soft” frequent itemset mining
  - Greedy search for incrementally detecting maximal term sets of co-occurrence
- DF-IDF<sub>t</sub> approach
  - Select n-grams based on burstiness (increase in frequency relative to recent past) → hierarchical clustering of n-grams

# Leveraging Influence

- Exploiting social context in the analysis procedure
- **Assumption:** Tweets coming from influential users are expected to be more relevant / less noisy
- Extract a first set of topics with previous TD methods
- Identify a set of influencers for each of those topics
  - Create content citation graph [nodes: users, edges: retweets], term profile for each edge based on RT text
  - Term- and topic-based subgraphs
  - Influence computation based on random walks (distributed computation on top of Hadoop)
- Filter original stream of Tweets taking into account only Tweets coming from influencers

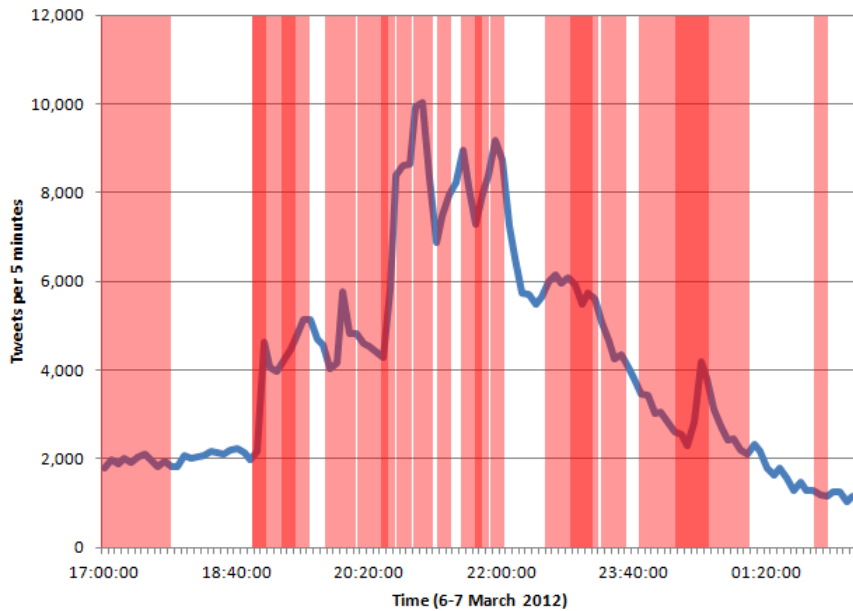


# Evaluation

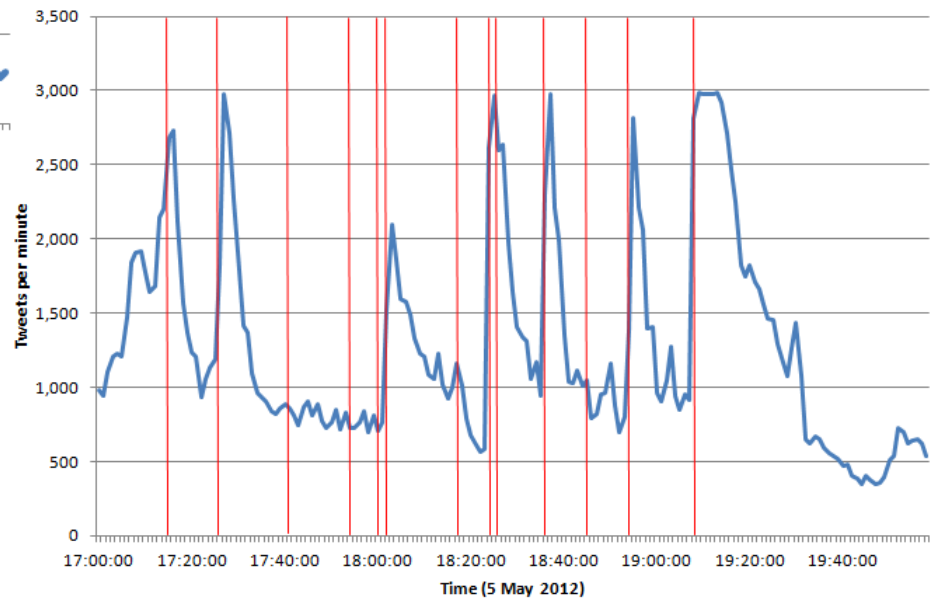
- Two datasets
  - **Super Tuesday (ST)**: event in presidential nomination race for US Republican party [**3.5M** tweets, average **131** tweets/min]
  - **FA Cup (FA)**: Chelsea vs. Liverpool [**444K** tweets, av. **148** tweets/min]
- Manual ground truth construction of set of interesting topics
  - Source: mainstream media reports
  - Representation: Set of topics (keywords + short headline) per timeslot [1-hour timeslot for ST, 1-minute for FA]
- Metrics
  - Topic recall (number of found target topics) [TRec]
  - Keyword-level recall [KRec] & precision [KPrec]
- Baseline
  - Latent Dirichlet Allocation (LDA)

# Evaluation

## Super Tuesday



## FA Cup



# Results

	FA Cup			Super Tuesday		
	TRec	KRec	KPrec	TRec	KRec	KPrec
<b>LDA</b>	<b>92.3%</b>	14.8%	<b>71.4%</b>	27.3%	22.4%	40.4%
<b>Doc-p</b>	53.8%	12.4%	46.4%	27.3%	11.6%	33.8%
<b>Graph-based</b>	84.6%	9.6%	62.5%	18.2%	7.6%	33.1%
<b>Freq pattern</b>	53.8%	30.3%	35.7%	18.2%	<b>23.3%</b>	17.6%
<b>Soft freq itemset</b>	84.6%	16.3%	58.9%	40.9%	13.0%	39.0%
<b>DF-IDF<sub>t</sub></b>	<b>92.3%</b>	<b>18.0%</b>	57.1%	<b>45.5%</b>	16.9%	<b>41.9%</b>

# Evaluation

FA cup:

- Time: 17:16, Ramires scores for Chelsea.
  - goal, 1-0, ramires, #cfc
- Time: 18:56, Andy Carroll takes a header but Cech saves Chelsea.
  - #facupfinal, saved, carroll, claiming, header, cech, @chelseafc, #cfcwembley

Super Tuesday:

- Time: 19:00: ABC/NBC/CNN project Newt Gingrich as the winner of the Georgia Primary
  - georgia, newt, wins, primary, republican, breaking, gingrich
- Time: 19:00, Newt Gingrich says “Thank you Georgia! It is gratifying to win my home state so decisively to launch our March Momentum”
  - @newtingrich, win, georgia, launch, #marchmo, gratifying, march, momentum

# Mobile browsing of large image collections on a smart phone

# Mobile photo browsing & tagging on the go



## CITY PROFILE MINING

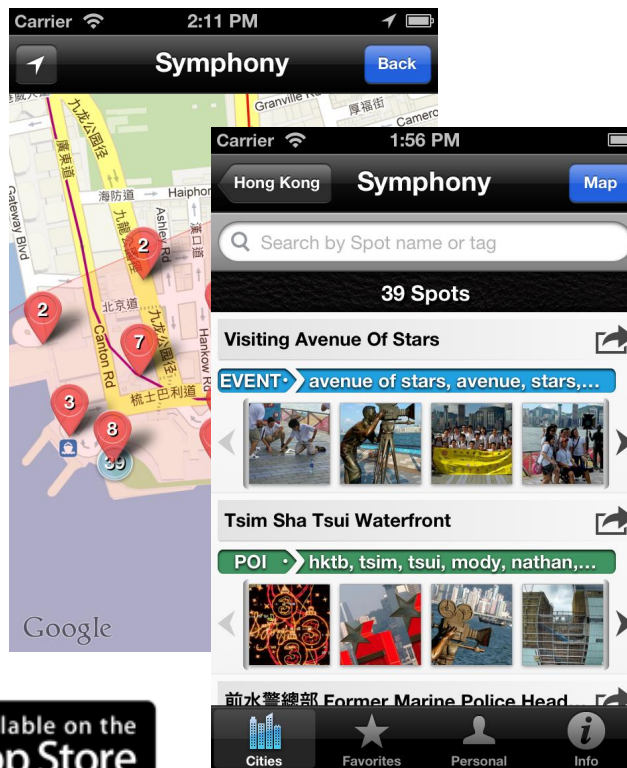
**Areas:** cluster geotagging information (BIRCH)

**Image clustering:** community detection (SCAN) on image similarity graphs

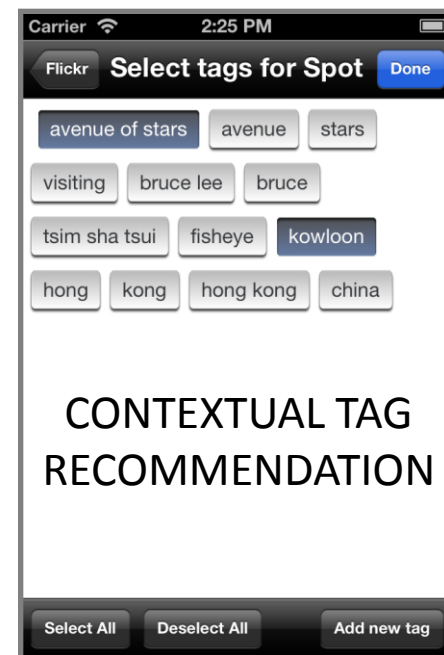
**Cluster processing:**

- classify landmarks-events
- extract titles and tags

## HIERARCHICAL PHOTO BROWSING



## OWN IMAGE TAGGING



## CONTEXTUAL TAG RECOMMENDATION

ClustTour



# City profile creation

## PHOTO COLLECTION



Area Extraction by Spatial Clustering

Area Title Extraction

Landmark & Event Detection

Temporal Analysis

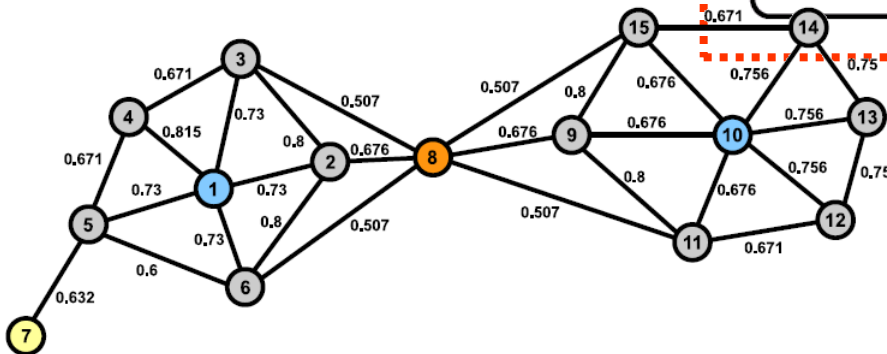
Automatic Extraction of Landmark and Event Titles

Relational Database & Web API

Mobile Client



Community detection on image similarity graphs







# Conclusions

- Great interest in both use cases
  - In news social media have transformed both news generation and consumption
- Social media data mining can provide interesting results in many applications
- Not all data always available (e.g. User queries, fb)
  - Infrastructure, Policy issues
- Technical challenges
  - Fusion (multi-modality, context), real-time, noise, big data, aggregation (web, Linked Open Data)
- Applications challenges
  - User engagement, privacy, copyright, commercialization

social sensor

Thank you!