Sociologie computationnelle pour l'environnement

GreenAI Uppa, paul.gay@univ-pau.fr

Lundi 31 janvier 2022

Definition

Social Computing" refers to systems that support the gathering, representation, processing, use, and dissemination of information that is distributed across social collectivities such as teams, communities, organizations, and markets. Moreover, the information is not "anonymous" but is significantly precise because it is linked to people, who are in turn linked to other people.

Social Computing, Douglas Schuler, 1994

Motivations

Why would we do that ?

- Assist policy makers
- Target commercial brands
- Inform people
- Improve human computer interfaces

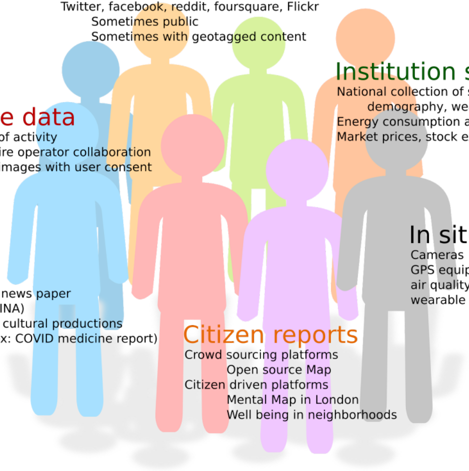
Let's start by a broad non exhaustive view

Data sources

And more precisely, let's start with the data

What can we use?

Data sources



Social Networks

Twitter, facebook, reddit, foursquare, Flickr
- Sometimes public
- Sometimes with geotagged content

Institution statistics

National collection of statisics about the population
- demography, wealth, employment

Energy consumption and bills

Market prices, stock exchange

Mobile data

GPS, sign of activity
- Require operator collaboration

Recorded images with user consent

In situ sensor

Cameras

GPS equiped cars (eg taxis in New York)

air quality sensors

wearable devices

Medias

Broadcast media, news paper

National Archive (INA)

Film, songs, other cultural productions

Publications (for ex: COVID medicine report)

Citizen reports

Crowd sourcing platforms
- Open source Map

Citizen driven platforms
- Mental Map in London
- Well being in neighborhoods

Data sources

Social Networks

Twitter, facebook, reddit, foursquare, Flickr
- Sometimes public
- Sometimes with geotagged content

Institution statistics

National collection of statisics about the population
- demography, wealth, employment
- Energy consumption and bills
- Market prices, stock exchange

Mobile data

GPS, sign of activity
- Require operator collaboration

Recorded images with user consent

In situ sensor

Cameras
taxis in New York)
air quality sensors
wearable devices

Medias

Broadcast media, news paper
National Archive (INA)
Film, songs, other cultural productions
Publications (for ex: COVID medicine report)

Citizen reports

Crowd sourcing platforms
- Open source Map

Citizen driven platforms
- Mental Map in London
- Well being in neighborhoods

Social Networks

Twitter, facebook, reddit, foursquare, Flickr, youtube
- Sometimes public
- Sometimes with geotagged content

Access to large scale covering in time and space

Data sources

Social Networks
Twitter, facebook, reddit, foursquare, Flickr
- Sometimes public
- Sometimes with geotagged content

Institution statistics
National collection of statisics about the population
- demography, wealth, employment

Energy consumption and bills
Market prices, stock exchange

Mobile data
GPS, sign of activity
- Require operator collaboration

Recorded images with user consent

In situ sensor
Cameras
... cars (eg taxis in New York)
air quality sensors
...ices

Medias
Broadcast media, news paper
National Archive (INA)
Film, songs, other cultural productions
Publications (for ex: COVID medicine report)

Citizen reports
Crowd sourcing platforms
- Open source Map

Citizen driven platforms
- Mental Map in London
- Well being in neighborhoods

Medias
Broadcast media, news paper

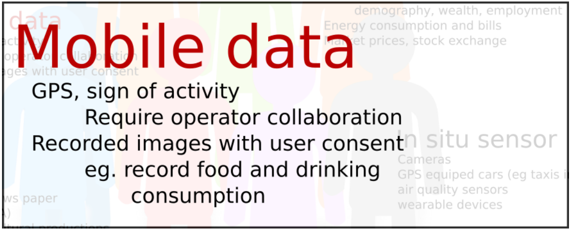
National Archive (INA)

Film, songs, other cultural productions

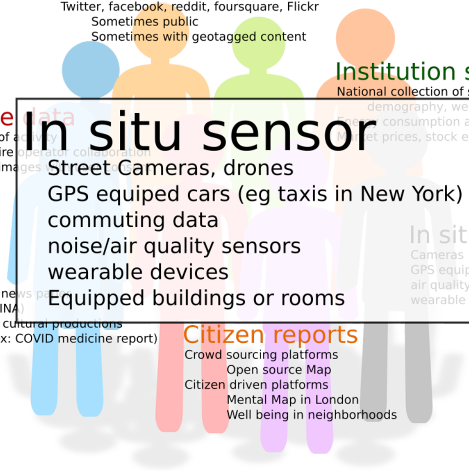
Publications (for ex: COVID medicine report)

Different facets of the same object

Data sources



Data sources



Social Networks
Twitter, facebook, reddit, foursquare, Flickr
- Sometimes public
- Sometimes with geotagged content

Institution statistics
National collection of statisics about the population
- demography, wealth, employment
- Energy consumption and bills
- Market prices, stock exchange

Mobile data
GPS, sign of activity
- Require operator collaboration

Recorded images

In situ sensor
- Street Cameras, drones
- GPS equiped cars (eg taxis in New York)
- commuting data
- noise/air quality sensors
- wearable devices
- Equipped buildings or rooms

In situ sensor
- Cameras
- GPS equiped cars (eg taxis in New York)
- air quality sensors
- wearable devices

Medias
Broadcast media, news paper
National Archive (INA)
Film, songs, other cultural productions
Publications (for ex: COVID medicine report)

Citizen reports
Crowd sourcing platforms
- Open source Map

Citizen driven platforms
- Mental Map in London
- Well being in neighborhoods

Data sources

Social Networks
Twitter, facebook, reddit, foursquare, Flickr
- Sometimes public
- Sometimes with geotagged content

Institution statistics
National collection of statisics about the population
- demography, wealth, employment

Energy consumption and bills
Market prices, stock exchange

Mobile data
GPS, sign of activity
- Require operator collaboration

Recorded ima

In situ sensor
Cameras
GPS equiped cars (eg taxis in New York)
air quality sensors
wearable devices

Medias
Broadcast media, news paper
National Archive (INA)
Film, songs, other cultural productions
Publications (for ex: COVID medicine report)

Citizen reports
Crowd sourcing platforms
- Open source Map

Citizen driven platforms
- Mental Map in London
- Well being in neighborhoods

Institution statistics
National collection of statisics about the population
- demography, wealth, employment

Energy consumption and bills
Market prices, stock exchange

Data sources

Social Networks

Twitter, facebook, reddit, foursquare, Flickr
- Sometimes public
- Sometimes with geotagged content

Institution statistics

National collection of statisics about the population
- demography, wealth, employment

Energy consumption and bills

Market prices, stock exchange

Mobile data

GPS, sign of activity
- Require operator collaboration

Recorded ima

Medias

Broadcast media, newspaper

National Archive (INA)

Film, songs, other cultural productions

Publications (for ex: COVID medicine report)

In situ sensor

Cameras

GPS equiped cars (eg taxis in New York)

air quality sensors

ble devices

Citizen reports

Crowd sourcing platforms
- Open source Map

Citizen driven platforms
- Mental Map in London
- Well being in neighborhoods

Take the data and give something back

Citizen reports

Crowd sourcing platforms
Open source Map
Citizen driven platforms
Mental Map in London
Well being in neighborhoods

Data sources

Social Networks

Twitter, facebook, reddit, foursquare, Flickr
- Sometimes public
- Sometimes with geotagged content

Institution statistics

National collection of statisics about the population
- demography, wealth, employment
- Energy consumption and bills
- Market prices, stock exchange

Mobile data

GPS, sign of activity
- Require operator collaboration

Recorded ima

Citizen reports

SenseCityVity

Set up collective action (Salvador et al 2017)

sensor

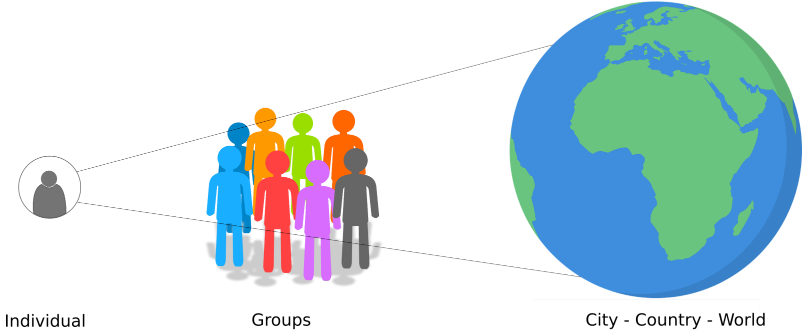
ars (eg taxis in New York)
sors
ces

Medias

Broadcast media, news paper
National Archive (INA)
Film, songs, other cultural prod
Publications (for ex: COVID me

which applications?

Social scales



Research questions at different scales

Social scales

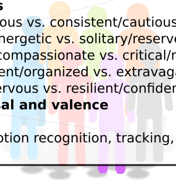
Individual

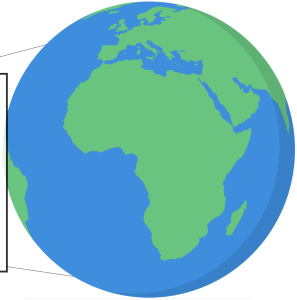
Big five personality traits

- Openness : inventive/curious vs. consistent/cautious
- Extraversion : outgoing/energetic vs. solitary/reserved
- Agreeableness : friendly/compassionate vs. critical/rational
- Conscientiousness : efficient/organized vs. extravagant/careless
- Neuroticsim : sensitive/nervous vs. resilient/confident

Characterization with **arousal and valence**

Affective Computing, Emotion recognition, tracking, detection of visual cues, gaze tracking

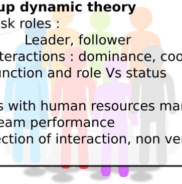




Individual | Groups | City - Country - World

Research questions at different scales

Social scales



Groups

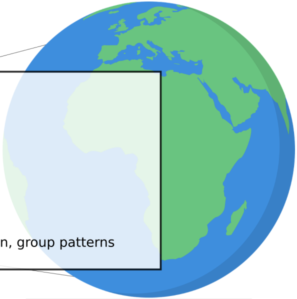
Group dynamic theory

- Task roles :
  - Leader, follower
- Interactions : dominance, cooperation
- Function and role Vs status

Links with human resources management

- Team performance

Detection of interaction, non verbal communication, group patterns



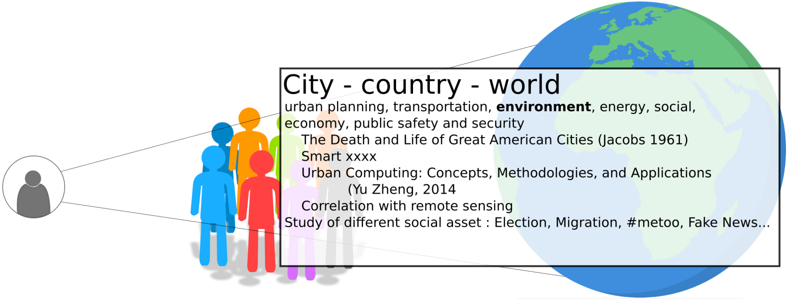
Individual

Groups

City - Country - World

Research questions at different scales

Social scales



City - country - world

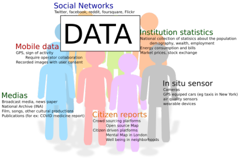
urban planning, transportation, **environment**, energy, social, economy, public safety and security

- The Death and Life of Great American Cities (Jacobs 1961)
- Smart xxxx
- Urban Computing: Concepts, Methodologies, and Applications (Yu Zheng, 2014
- Correlation with remote sensing

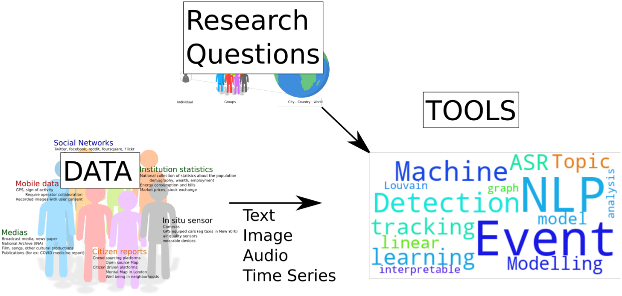
Study of different social asset : Election, Migration, #metoo, Fake News...

Research questions at different scales

Powered by machine learning



Powered by machine learning



Requirements:

- Linear Interpretable models to understand social contexts
- Data privacy, ethical bias
- Unsupervised data, collaboration with domain experts

Industrial applications

- Graph analysis companies
  - Graphika : among other things : Features reports on social network activiy
  - Linkfluence : Market research
  - LinkCurious : Connected data for crime detection
- Affective computing
  - ubiquity, Idaaas, Microsoft Azure, Affectiva
  - Eyeris : Smart in cabin sensing in vehicles.

And the environment?

Environmental have social causes and social impacts

- How environment aleas affect the different social groups ?
- Perception, community cohesion, Consequences
  - Environmental sociology (John Hannigan, 2014, Riley Dunlap 1979)
  - H. T. Williams et al. "Network analysis reveals open forums and echo chambers in social media discussions of climate change". In: **Global environmental change** 32 (2015), pp. 126–138
  - A. Ghermandi and M. Sinclair. "Passive crowdsourcing of social media in environmental research: A systematic map". In: **Global environmental change** 55 (2019), pp. 36–47
  - A. A. Anderson. "Effects of social media use on climate change opinion, knowledge, and behavior". In: **Oxford research encyclopedia of climate science**. 2017
  - J. Kaiser and C. Puschmann. "Alliance of antagonism: Counterpublics and polarization in online climate change communication". In: **Communication and the Public** 2.4 (2017), pp. 371–387
  - J. Shang et al. "Inferring gas consumption and pollution emission of vehicles throughout a city". In: **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining**. 2014, pp. 1027–1036

Application in crisis management

Let's focus on one case study

Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning.
Marc-André Kaufholda,b,∗, Markus Bayera, Christian Reutera
Information Processing & Management. 2020

Crisis informatics

Social media use during disasters and emergencies

- Valuable information (eyewitness reports, pictures, videos)
- comprehensive situational overview
- Organize help, crowsourcing, Communication and coordinations among citizens and volunteers
- rise situation awareness
- Shown to be an important vector in recent hazards (Brussels Bombing, 2012 Sandy Hurrycane, 2013 European Floods)

Issue: Information overload

- Lack of resources and skills

Crisis informatics

Social media use during disasters and emergencies

- Valuable information (eyewitness reports, pictures, videos)
- comprehensive situational overview
- Organize help, crowsourcing, Communication and coordinations among citizens and volunteers
- rise situation awareness
- Shown to be an important vector in recent hazards (Brussels Bombing, 2012 Sandy Hurrycane, 2013 European Floods)

Issue: Information overload

- Lack of resources and skills

Machine learning for information management

Supervised machine learning effective to sort this information, however:

- Requires a lot of training data
- Not adapted in the context of crisis
  - timing issue
  - Lack of clear criterion : dynamic topics and unexpected event.

Solution :

- Light Random Forest classifier with feature selection
- Active learning procedure to train the model
  - Performances are better in batch mode.
  - One batch learner model to classify, one incremental learner model to select the data to be labeled.

Experiments on two twitter datasets : European Flood and /BASF SE incident.

Machine learning for information management

Supervised machine learning effective to sort this information, however:

- Requires a lot of training data
- Not adapted in the context of crisis
  - timing issue
  - Lack of clear criterion : dynamic topics and unexpected event.

Solution :

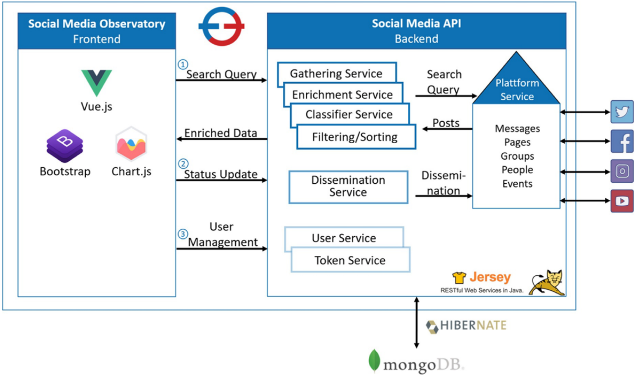
- Light Random Forest classifier with feature selection
- Active learning procedure to train the model
  - Performances are better in batch mode.
  - One batch learner model to classify, one incremental learner model to select the data to be labeled.

Experiments on two twitter datasets : European Flood and /BASF SE incident.

Reseach questions

- What are suitable criteria for relevance classification and labeling in disasters and emergencies (RQ1)?
- How can existing supervised machine learning techniques for relevance classification be improved for use in real disaster and emergency environments (RQ2)?
- How can the amount of labeled data required for relevance classification be reduced by active incremental learning and trans- parent visualization of the classifier's quality (RQ3)?
- How can the dynamic retraining of relevance classifiers be supported by user feedback performance-wise using batch learning with feature subset selection (RQ4) ?

Social Media Observatory



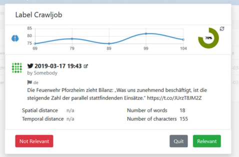
Complete system to build social media classifier

Preprocessing

- removal of characters (newline, tabulations, emojis)
- stem and lemmatization
- TF-IDF with Bag of words features
- Metadata
  - tweet geolocalisation (only 10%), author geolocalisation
  - time of emission
  - Number of retweet

How to label the tweet?

The notion of relevance is subjective, a choice is made there



- Relevant tweets
  - request for help
  - Fact AND fake news
  - Relevant in case of doubt to maximise the recall
- Non relevant Tweets
  - Condelances, Call for donation

Dataset description

- European flood datasets
  - 3923 posts over a period from 30 May to 28 June 2013
- BASF Fire
  - 3790 posts on the October 17th, 2016.

Experiments

- Building of the main Random Forest classifier
- Test of the active learning approach
- Training time is an important consideration
  - Taken for a full cross-validation and hyper parameters tuning process

Results of the Random Forest classifier

Classification Features Used	Accuracy	Precision	Recall	Time (s)
Words	90.8	91.3	81.1	850.271
Words + Number of Retweets	90.82	91.4	81.1	851.14
Words + Length	90.89	91.4	81.3	862.69
Words + Number or Retweets + Length	90.85	91.4	81.1	841.78
Words + Temporal Distance	90.93	91.4	81.3	901.22
Words + Geographical Distance (Author Distance and Tweet Distance)	91.03	91.6	81.3	1021.663
Words + Geographical Distance (Author Distance and Tweet Distance) + Temporal Distance	91.21	**91.8**	81.4	1078.092
Words + Distance (Author Distance and Tweet Distance) + Temporal Distance + Length	**91.23**	**91.8**	**81.5**	1110.276
Words + URLs	90.9	91.4	81.4	850.22
Words + Media	91	91.5	**81.5**	860.12
All Classification Features	91	91.6	81.1	1071.79
No Words + All Other Classification Features	84.35	84.4	75.1	**281.14**

Improvement with the use of metadata

Results of the Random Forest classifier

Classification Features Used	Accuracy	Precision	Recall	Time (s)
Words	90.8	91.3	81.1	850.271
Words + Number of Retweets	90.82	91.4	81.1	851.14
Words + Length	90.89	91.4	81.3	862.69
Words + Number or Retweets + Length	90.85	91.4	81.1	841.78
Words + Temporal Distance	90.93	91.4	81.3	901.22
Words + Geographical Distance (Author Distance and Tweet Distance)	91.03	91.6	81.3	1021.663
Words + Geographical Distance (Author Distance and Tweet Distance) + Temporal Distance	91.21	**91.8**	81.4	1078.092
Words + Distance (Author Distance and Tweet Distance) + Temporal Distance + Length	**91.23**	**91.8**	**81.5**	1110.276
Words + URLs	90.9	91.4	81.4	850.22
Words + Media	91	91.5	**81.5**	860.12
All Classification Features	91	91.6	81.1	1071.79
No Words + All Other Classification Features	84.35	84.4	75.1	**281.14**

Improvement with the use of metadata

	Accuracy [%]	Precision [%]	Recall [%]	Time [s]
10,153 Features	**91.64**	94	**85.2**	1120.22
148 Features	91.28	**98.2**	80.4	**204.326**

Reduction of the processing time with random feature selection

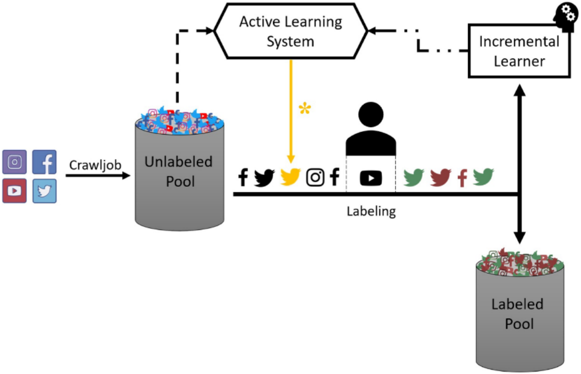
Active learning system

Use of an additional classifier to suggest new posts to the user for labeling

- Maximum uncertainty sample
  - ie Select the sample which has 50% of confidence
  - Note : beware of outliars!!
- K nearest neighbours classifier with kdtree (with K=50)
  - 3 seconds to train
  - Worst performances than Random Forest

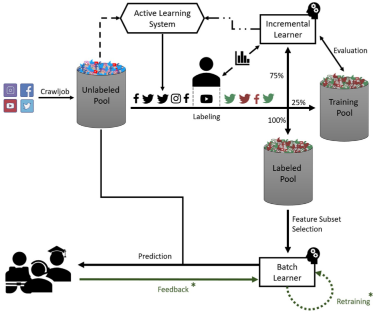
Active learning system

- Principle for the active learning system
- Active learning request at each third to fifth labeling



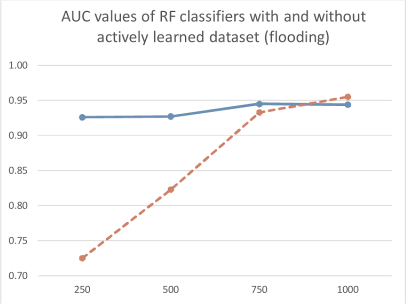
Complete Active learning system

- Adding online evaluation of the Knn classifier with 25% hold out data
- Adding correction from end user



Results for active learning

- Gains reported thanks to Active learning on their data
- Specially at the beginning



Concluding remarks on the paper

- Complete system for information overload in crisis management
- Easy inclusion of new topics

Some comments

- Probably particular to a specific dataset
- Not clear comparison with domain adaptation
- Scalability to millions of items ?