SEMANTIC PATTERNS AND TWITTER

Egyptian Revolution, Semantic Patterns, Knowledge Discovery

Peter Teufel
EGIZ/IAIK/Graz University of Technology
peter.teufel@iaik.tugraz.at
IAIK/EGIZ/GRAZ

Institute for Applied Information Processing and Communications

- Application oriented research
- Academic teaching activity
- Independent information center
- Support secure information society

- IAIK/EGIZ
- Graz University of Technology
- Security focus
- eGov, ePart, smartphones, citizen card

Tuesday, August 30, 11
• Twitter
• Knowledge Discovery in Twitter
• Semantic Patterns
• Examples
TWITTER

• 200 million users (March 2011)

• 200 million tweets per day (August 2011)

• One million apps connecting to Twitter (August 2011)

• Tweets

• Following...
TWITTER

- Public profiles of users
- Following and being followed
- Follow friends, interesting people, news coverage, companies
SEARCH QUERIES

• Search for anything on Twitter
• Text, location, user, hashtags
• Google+?
TWEETS

• Important fact: max. 140 characters

• Limits conveyed information?

• Yes, but also forces user to be precise

• Thus, a lot of information contained in a Tweet

• Concept of retweets (RT), answers (@), hashtags (#epart11)
TWEETS

• hashtags: arbitrarily chosen terms, with a # prefix

• Self organizing process (e.g. #irene, #jan25, #epart2011)

• Used to categorize Tweets, makes search easy
DATA IN A TWEET

• Interesting for Knowledge Discovery
• User
• Timestamp
• Location (not mandatory)
• Message including text, links, hashtags, other users
• Social networks, links, spreading of tweets
KNOWLEDGE DISCOVERY

• Huge amount of data
• Huge potential for knowledge discovery
• Many examples from a wide range of areas
KNOWLEDGE DISCOVERY

• News coverage, communication (Iran election in 2009, Twitter revolutions: Tunisia, Egypt, Libya)

• Sentiment analysis: products, politicians, other topics

• Earthquake epicenters

• Decease outbreak and analysis, use of antibiotics, flu

• Stock market modeling

• Predicting elections

Tuesday, August 30, 11
E-PARTICIPATION

• Either use the available data to gain information

• or use Twitter infrastructure for an eParticipation project

• Hashtags

• Focus here: A new technique for improving knowledge discovery
KNOWLEDGE DISCOVERY

- Terms, hashtags, timestamps, locations, links, users...
- Natural Language Processing and text mining only a small part of the problem
- Work on eParticipation, network security, semantic web data, event correlation
- Machine learning important in knowledge discovery: clustering, classification, understanding of relations within the data...
SEMANTIC PATTERNS

• Preprocessing, data transformation, choice of the algorithm model interpretation, many parameters depending on the desired information, data and the selected algorithm.

• Time consuming setups

• Got fed up... Idea: Semantic Patterns

• Shift from feature vectors to a semantic representation.
IDEA: SEMANTIC PATTERNS

• Represent semantic information of arbitrary co-occurring features: terms, timestamps, users, location, numerical values...

• Transform classical feature vectors into these patterns

• Patterns form the basis for arbitrary analyses
EXAMPLES

• Two examples

• Twitter data from the Egyptian Revolution

• The CIA WorldFactbook (to show that SEMPS work on arbitrary data)
EGYPTIAN REVOLUTION

• Tweets from to 2011/01/24 to 2011/02/11
• Extracted from Google Realtime (does not exist anymore)
• But Tweets also available through Twitter API
• Jan 25th (#jan25): start of mass protests
• Jan 28th to 31st: shutting down ISPs (back on on Feb 2nd)
• Feb 11th: Mubarak steps down
KNOWLEDGE DISCOVERY

• Taking terms, timestamps and hashtags

• Each tweet represents an instance

• Semantic relations within each Tweet is determined and stored in a semantic network

• Convert tweets, terms, hashtags, timestamps into Semantic Patterns
KNOWLEDGE DISCOVERY

- Semantic relations
- Semantic search queries
- Clustering
- Time based analysis
- Relevance, anomalies, classifiers, ...
SEMANTIC RELATIONS

• Semantic Pattern is a semantic fingerprint

• Relations between terms, hashtags, timestamps directly represented in the pattern

• Pattern is only a vector: similarity, interpretation, operations

Tuesday, August 30, 11
SEMANTIC RELATIONS

- hashtag: #egypt
SEMANTIC RELATIONS

• tweet: The Battle of Tahrir Square 02/02/2011
SEMANTIC RELATIONS

- term: wikileaks
SEMANTIC RELATIONS

- timestamp: 2011-2-12-12
SEMANTIC SEARCH

• Each tweet, term or timestamp represents a semantic fingerprint

• Calculating similarity between fingerprints

• Semantic aware search algorithm
SEMANTIC SEARCH

• **Query**: After access is shut down, some ponder if Internet access is a basic human right. 2011/01/29

• **Result**: Apparently switching off Twitter is becoming the standard procedure of every country facing social unrest. 2011/01/25

• **Result**: RT @sharifkouddous I will eventually lose all communication here. But I will be out in the streets tomorrow. 2011/02/01
SEMANTIC SEARCH

• Live footage shows Egypt’s army vehicles deploy among protesters at scene of violence in Tahrir square - Al Arabiya TV. 2011/02/03
SEMANTIC SEARCH

• Timestamp at February 3rd 2011

• Search for similar events on other timestamps?

• Create Semantic Pattern for 2011-2-3-0

• Search for related patterns (limit to timestamp patterns)
SEMANTIC SEARCH

• Retrieving results for 2011-2-3-0

• Example: 2011-2-5-12

• What happened there?

• **Result:** Army removing burnt Police vehicle from Tahrir Sq - dark symbol for protestors. 2011/02/05

• **Compare to:** Live footage shows Egypt’s army vehicles deploy among protesters at scene of violence in Tahrir square - Al Arabiya TV. 2011/02/03
SEMANTIC SEARCH

• Other examples:

• Tahrir: square, #egypt, #jan25, egypt, cairo, protestor

• Mubarak: Hosni, 2011-2-10-0, resign, leave, power, 2011-2-10-12

• Internet: 2011-1-28-0, shut, #egypt, service, access
CLUSTERING

• Apply a machine learning algorithm to the Semantic Patterns of the tweet

• Get an overview of the dataset by inspecting the clusters

• Use this as basis to continue with further analyses
TIMELINE

- Extract the most important Tweets in the complete timeline
- Based on their semantic activity over time
- Allows a quick overview
OTHER TIME BASED ANALYSIS

• Semantic development over time

• term: internet
OTHER TIME BASED ANALYSIS

- RT @benjamincohen Twitter and Google team up to get around Egypt's internet blackout 2011/02/01
SOMETHING COMPLETELY DIFFERENT?

• Word Fact Book

• Information about the world’s countries

• Exports, unemployment rates, birth rates, organizations, agreement, service sectors...
SOMETHING COMPLETELY DIFFERENT?

• Service sector: 70% (others: industry, agricultural)
SOMETHING COMPLETELY DIFFERENT?

- Service sector: 70%, limited to exports
SOMETHING COMPLETELY DIFFERENT?

- Service sector: 20%, limited to exports
SOMETHING COMPLETELY DIFFERENT?

- Unemployment rate: 20%, limited to agricultural products
SOMETHING COMPLETELY DIFFERENT?

• No it isn’t, but due to Semantic Patterns setup remains the same

• Features and feature values are semantically related

• Combining text features with other information

• Twitter is a wealth of interesting information that can be combined with arbitrary features

• Could learn many things for e-Participation
THANKS

• Questions?

• Data sets you want to analyze?

• Peter Teufl (peter.teufl@iaik.tugraz.at)

• http://www.iaik.tugraz.at

• http://www.egiz.gv.at