Analysis of Social Media Streams

Florian Weidner

Dresden, 21.01.2014
Outline

1. Introduction
2. Social Media Streams
   - Clustering
   - Summarization
3. Topics
   - Detection
   - Tracking
4. Conclusion
1. Introduction

- A lot of data
  ➔ hidden and obvious information
- Important for users, organization, ...
- Algorithms for static data well researched
- However:
  Processing of streams is still „in it‘s early stages“[1]

➔ State of the art overview
2. Social Media Streams

- High frequency
- Continuous
- Different kind of data
  - Text, links, pictures, meta-data...
- Human language is a problem!
2.1 Social Media Streams - Clustering

- Find groups of similar instances without prior knowledge!

- Curse of dimensionality

- Outliers
2.1.1 Social Media Streams – Clustering
Cluster Droplets, Similarity & Fading Functions

- **Cluster Droplet (CD):**
  statistical information (recency, #tweets, weights,...)

- **Similarity function:**
  cosine similarity, dice coefficient,...

- **Fading Function:**
  decay of cluster
2.1.2 Social Media Streams – Clustering Variable Feature Sets

- Feature Set
- Validity Index (VI)
- Clustering Threshold (CT)
- Reselection Threshold (RT)
2.1.2 Social Media Streams – Clustering Variable Feature Sets

1. Get Text
2. Insert into cluster
3. Calculate VI
4. Compare with CT & RT
2.2 Social Media Streams - Summarization

- Input stream is huge
  - Summarize based on intervals
- Cluster can still contain a huge amount of data
  - Summarize clusters

- Single sentence vs. Multiple sentence
- New text vs. Text from stream

- Noise
2.2.1 Social Media Streams – Summarization
Word-Variance Based Approach

Phrase Reinforcement Algorithm ➔ builds a tree

Output:
Set of sentences which summarize stream!
2.2.1 Social Media Streams – Summarization
Word-Variance Based Approach

1. A tragedy: Ted Kennedy died today of cancer
2. Ted Kennedy died today
3. Ted Kennedy was a leader
4. Ted Kennedy died at Age 77
2.2.2 Social Media Streams – Summarization
Distance Metrics

• Tweet-Cluster-Vector (timestamp, meta)
• Goal: extract k Tweets which cover as much content as possible

→ Distance of Tweet to cluster centroid
→ Size of cluster
→ Centrality Scores
3. Topics

- Abstract topic vs. real-life topic (event)

- Small-scale vs. large-scaled
  - short duration and less info vs. long lasting and a lot of data

- Semantic features important!
- For events, the location is important!
- Semantic features and weblinks
3.1 Topics - Detection

- **Topic augmentation**
  - external topic as input

- **Topic detection**
  - w/o prior knowledge

- **Clustering is important/simplifies the topic detection**
3.1.1 Topics – Detection
Word-Variance

- Topics are time-dependent!

- Simple solution: increase of certain words (i.e. „earthquake“)

→ Count words in intervals and compare!
3.1.1 Topics – Detection
Word-Variance

1. Preprocessing

2. Calculate word frequencies of incoming data for each time window

3. If there is a significant increase (threshold), keep word

4. Calculate correlations for all remaining words and cluster them
3.1.2 Topics – Detection Location

- Filter and cluster incoming data according to their location (just longitude/latitude)

- Weight Tweets and clusters with help of features (textual, other)

⇒ If weight > threshold ⇒ Topic
3.1.3 Topics – Detection
Authority Score & Tweet Influence

- Key users + selected users
- Key words + selected words
  ➔ Repository

Authority Score:
  ➔ Importance of the authors of the tweets in the cluster

Topical Tweet Influence
  ➔ How many important keywords are in the cluster?
3.1.3 Topics – Detection Authority Score & Tweet Influence

1. Cluster incoming data frequently with similarity function

2. Calculate Topical User Authority Score & Topical Tweet Influence of each cluster

3. Weight words and rank them ➔ emerging topic

4. Machine Learner (6 features) ➔ hot emerging topic
3.3 Topics and Events - Tracking

- Track topic during a period of time ➔ display (only) related content

- Track spatial development ➔ evaluate geotags and keywords
3.3.1 Topics and Events – Tracking
Tracking of an interesting topic

Background Corpus → Content Model
  - Quality Features
  - Semantic Features

Foreground Corpus → Feedback Model
  - Compare Tweet with x previous and best descriptive Tweets

Display

Query for topic 😊
4. Conclusion

Many different solutions:
- Cluster Droplets, Fading & Similarity Functions
- Variable Feature Sets
- Word-Variance
- Distance
- Scores (Authority, Tweet Influence)
- Content & Feedback Model

• No holistic solution
  - Filtered stream
  - Utilization of data sources
    ➔ just single purpose solutions

• Many restrictions!
• Few open source framework
  (lot of conceptual work)
Vielen Dank für die Aufmerksamkeit!
5. References

[7] Chen Y. - Emerging topic detection for organizations from microblogs, ACM SIGIR '13, 2013
[9] Hong L. - Discovering geographical topics in the twitter stream, WWW’12, 2012