Responsive Analytics of Highly-Connected Big Data

Dr. Peter Janacik, peter.janacik@tu-berlin.de
Linking concepts

Data from social networks (Twitter/Instagram), Wikipedia, Wiktionary, evocation/synonym databases, medical knowledge, etc.
Extracting sense from tokens

- Approach
  - Symbol/token: notion in general, oriented gradient cluster, recognized object, word
  - Semantic overlay absorbs flow of meaning
  - Follow-up processing
  - Polarization and gist extraction
  - Wikipedia/wiktionary as knowledge model for semantic overlay
Apache Flink

- Was initiated at TU Berlin (first under the name Stratosphere)
- MapReduce does not provide sufficient means to implement state-of-the-art analysis methods
- Flink allows to connect transformations (vertices) to a graph using data streams (directed edges)
- Distributed execution and placement within a cluster: Flink program -> subtasks -> slots
- Number of slots on one physical node is configurable but usually it is equal to number of cores
- Maximum degree of parallelism can be defined and is used by Flink during execution
Future Work

- Partitioning of data flow graph over several data centers based on available resources, data stream bandwidth, data privacy criteria
- Optimization criteria
  - Min processing time
  - Min costs
  - Best fit
  - Matching different criteria
- Dynamic migration in order to accommodate changing characteristics of physical topology (available bandwidth/resources (nodes), price, follow the sun, etc.)
Interdependence of areas

- Concept/relationship/story detection
- Visualization
- Human data/feedback generation
- Distribution at web-scale for insightful analysis

Enabled by:
- Alters models/algo behavior
- Results/recommendation to trigger
- Comprehensible presentation
- Data/results

www.cit.tu-berlin.de
Semantic graph as result of analysis
Semantic graph as result of analysis
Instagram interaction heat map
Instagram interaction heat map
But how to visualize, what these graphs are about, when there are typically millions to trillions of edges?
Approaches to visualization

- **2-dimensional**
  - Works well with most of the currently available devices, no special hardware needed
  - Supported by broad range of platforms
  - Less complex, easier to implement
  - Fewer problems with readability and overlap
Approaches to visualization

- 3-dimensional
  - Chance to make use of additional dimension to untangle the big graph
  - Different perspectives may cover different aspects/lead to different conclusions
  - Can exploit the full potential of touch interfaces
Responsive Analytics of Highly-Connected Big Data

Dr. Peter Janacik, peter.janacik@tu-berlin.de