Towards an Advanced System for Real-Time Event Detection in High-Volume Data Streams

Andreas Weiler, Svetlana Mansmann, Marc H. Scholl
University of Konstanz, Germany

Motivation

High-Volume Data Streams impose new challenges on the way such data volumes are processed and analyzed in real-time or near real-time.

- How can database systems be used for real-time processing?
- How can we handle the massive amount of data in the stream?
- How can on-line event detection in data streams be efficiently implemented?
- What kinds of events can be detected? What classifies a detected item as an event?

System Description

Data Facts

- Fast, uneven, noisy, and information-rich Stream of short messages
- More than 30 million Twitter messages per day
- About 10% with attached geographic information
- Over 60 fields of meta-data per message

Event Detection

Processing Pipeline

- Tokenization: discard stopwords & short terms
- POS Tagging: limit to nouns and proper nouns
- Let \( x_n \) (n is the actual occurrence) be the number of tweets run through the system since the last occurrence of the term.
- If result \( \text{res} = \log(x_n) - \log(x_{n-1}) - \log(x_{n-2}) \) is less than the threshold \( T_1 \), increase number of hits, else decrease number of hits.

Future Work

- Improve Event Detection
- Database with Term Patterns
- Event Enrichment: Correlation with further data sources
- Event Classification and Ranking
- Evaluation of long term Event Detection

References


Case Study: Earthquake Detection

Detection of an earthquake happening on the Indonesian island of Sumatra at 08:38:38 AM UTC on Wednesday, April 11th 2012.

- Resulting event terms in the hour 08:00 to 09:00 AM UTC
- Relationships of found event terms
- Sentiment of tweets with "gempa" (engl. "quake"), "tsunami", and "aceh"
- Geospatial Density Maps showing all tweets with geospatial information