Exploring Significant Interactions in Live News

Erich Schubert, Andreas Spitz, Michael Gertz

March 26, 2018 — NewsIR’18 Workshop at ECIR 2018

Heidelberg University, Germany
Database Systems Research Group
What is in the news right now?
Core Idea: Cooccurrences

Focussing on the participating entities

- politicians, countries, companies, and celebrities are always in the news
- what changes is how they interact

See also: A. Spitz and M. Gertz. “Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events”. In: ACM SIGIR. 2016
Core Idea: Cooccurrences

Focussing on the participating entities

- politicians, countries, companies, and celebrities are always in the news
- what changes is how they interact

Capturing interactions

- it is not sufficient to look at one thing at a time
- instead, look at the cooccurrences of terms and entities

See also: A. Spitz and M. Gertz. “Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events”. In: ACM SIGIR. 2016
Example: Superbowl

News Cloud  ─  2018-02-05T04:20:08 UTC ─

Qualcomm
Broadcom

Edwin Jackson
Quillin

Catrall

digital currency
Bitcoin
buying

third-down
birth
announced

Khloe Kardashian
posted

Kylie Jenner
babygirl

Instagram

touchdown
passed

Travis Scott

the suffragette

Brandon Graham
Derek Barnett

T

hologram
giving
ad
Timberlake

Mobile
halftime
Super Bowl

Nicos Anastasiades
win
first Super Bowl

Nick Foles

Zach Ertz

sporting
USA TODAY
Core Ideas: Significance

Counting is not enough:

- many methods use word counts
- certain words are always frequent, others always rare
- it is interesting if a rare term or entity suddenly becomes frequent

Significance: compare frequency to expected frequency!

Details on our significance measure are in the arXiv predecessor:
Prototype Overview: Data Preparation

1. monitor live news (push notifications & RSS)
2. group articles in microbatches (25 articles)
3. crawl and extract text
4. tokenize text, detect and link entities
5. aggregate weighted cooccurrences
6. score significance based on estimated frequencies
7. update estimates for next micro-batch

Use count-min style sketches for estimation:
Prototype Overview: Visual Layout

1. select and cluster top (co-) occurrences based on significance
2. visualize as word-cloud in the browser with significance-based SNE
3. edges visualize significant cooccurrences
4. colors denote clusters
5. currently supported languages: English and German
Topic Example: Moscow Plane Crash (prior)

News Cloud ➤ 2018-02-11T12:42:02 UTC ➤

- Sheikh Maqsoud
- Afrin
- Kurdish
- crashed
- Helicopter
- the Grand Canyon
- Eurocopter EC130
- federal prison
- Mordaunt
- Penny Mordaunt
- Cathedral
- Cape Town
- ANC
- South Africa
- discussed
- Emami
- Seyed
- Ramaphosa
- Zuma
- Mabe
- Billingham
- Rajapaksa
- Kashmir
- Sirisena
- Ramaphosa
- Moula
- Ramo
- Medvedeva
- Red
- Gerard
- slopestyle
- snowboarding
- Westerville
- Dmitri Soloviev
- Bolton
- puck
Topic Example: Moscow Plane Crash (emerging)
Topic Example: Moscow Plane Crash (dominant)

News Cloud  ➤ 2018-02-11T13:30:14 UTC ➤

Purdue
oxycodone
Mordaunt
Red
Gerard

Afrin
Orsk
Kurdish

Domodedovo Airport
helicopter

region crashed

Moscow
passenger crew
plane
disappearing radar
emergency

Saratov Airlines

Billingham
Rajapaksa

Emami
Seyed

Bolton

Ramaphosa
discussed
Zuma

Ramo
Adelson

United National Party

Sirisena

An

Russian

Kashmir

Mabe

Ramaphosa
Try the live demo:

newsir-demo.ifi.uni-heidelberg.de
Try the live demo:

newsir-demo.ifi.uni-heidelberg.de

Thank you!

Questions & Discussion