# Scalable Detection of Emerging Topics and Geo-spatial Events in Large Textual Streams


Erich Schubert[1,2] Michael Weiler[1] Hans-Peter Kriegel[1]

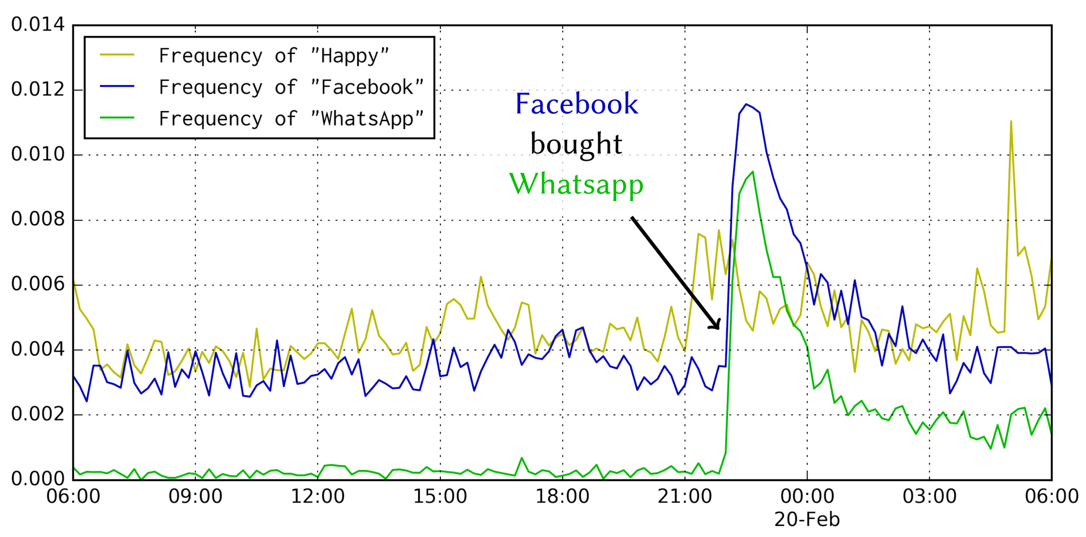
[1] Lehr- und Forschungseinheit Datenbanksysteme, Ludwig-Maximilians-Universität München

[2] Lehrstuhl für Datenbanksysteme, Ruprecht-Karls-Universität Heidelberg

{schube,weiler,kriegel}@dbs.ifi.lmu.de


## Objective and Summary

- Scalable: able to process years of news and Twitter
- Detection: topics and keywords should *not* need to be defined beforehand
- Emerging: significant increase (c.f. "Trending Topics")
- Topics: not every single message, but groups of related messages are of interest
- Geo-spatial Events: observe locality and able to detect *geographic change* and differences




## Key Ideas of our Solution

- From statistics: control charts for change detection.
- From computational linguistics: Analyze word cooccurrences for more meaningful results.
- From mathematics: Exponentially weighted moving averages for streaming operation.
- From databases: Hashing and Count-Min sketches for scalability to large data.
- From data mining: Clustering of word pairs into simple "topics" based on cooccurrences.
- From visualization: Word-cloud like visualization, but incorporating the relationships of words.
- Integrate geographic information by mapping coordinates to tokens similar to text.

The *big* challenge is scalability to millions of word pairs, at thousands of Tweets per second!

### Tracking all Cooccurences

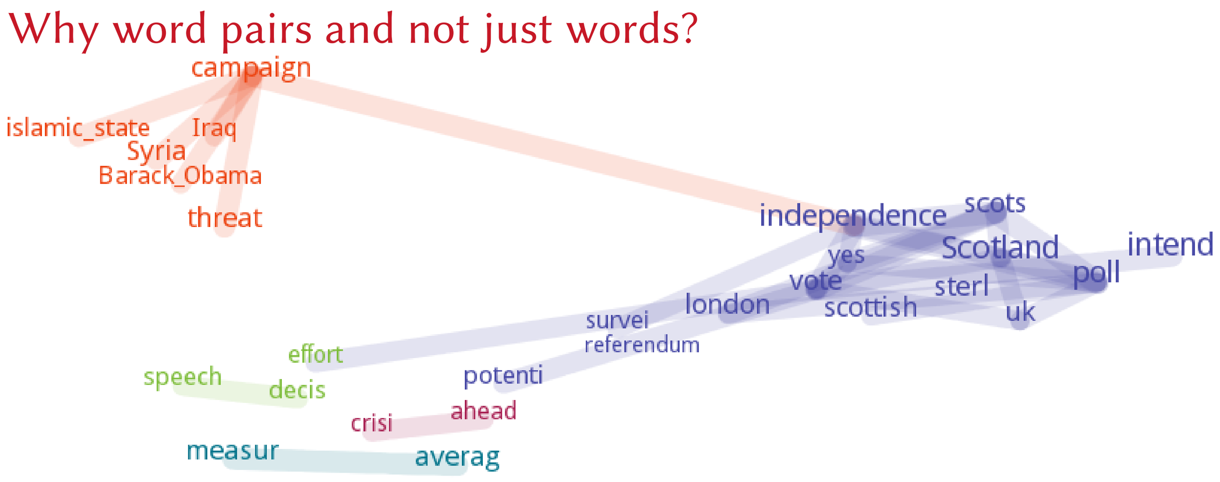
Word combinations are interesting:

- "Facebook" bought "WhatsApp"
- Edward "Snowden" traveled to "Moscow"
- "Putin", "Obama" and "Merkel" — their interactions are more interesting than their frequency

Why not the most popular terms?

- "@justinbieber" is always popular on Twitter
- Domain specific stopwords (e.g. "follow", "RT")
- Cultural-, language- and geographic differences

Why word pairs and not just words?




Pairs allow the discovery of interactions and structure.

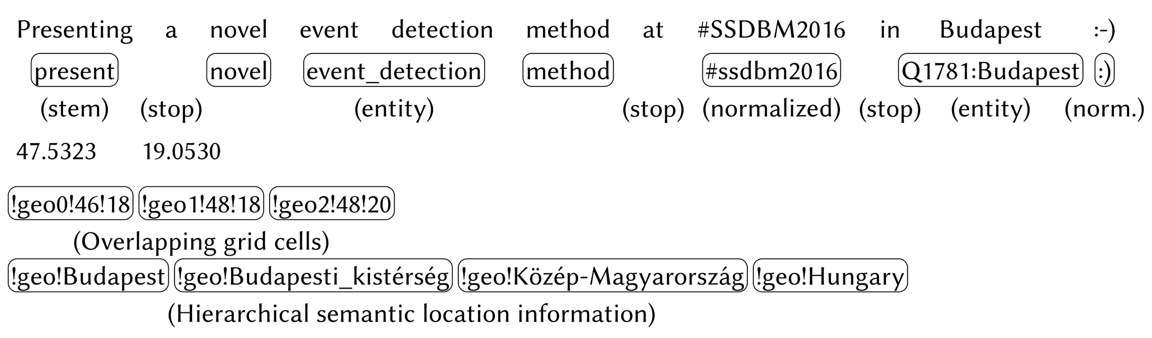
### Integrating Geographic Information

We map geographic data to tokens

$$(\text{longitude}, \text{latitude}) \rightarrow \{\text{Symbol}, \ldots\}$$

such that nearby locations produce the same symbol.

Example tokenization of a Tweet:




## Change Model and Implementation

For every word, word-word-, or word-location-pair $(w, l)$ we use a z-score-like significance:

$$z_t(w,l) := \frac{f_t(w,l) - \max\{\text{EWMA}[f(w,l)], \beta\}}{\sqrt{\text{EWMVar}[f(w,l)] + \beta}}$$

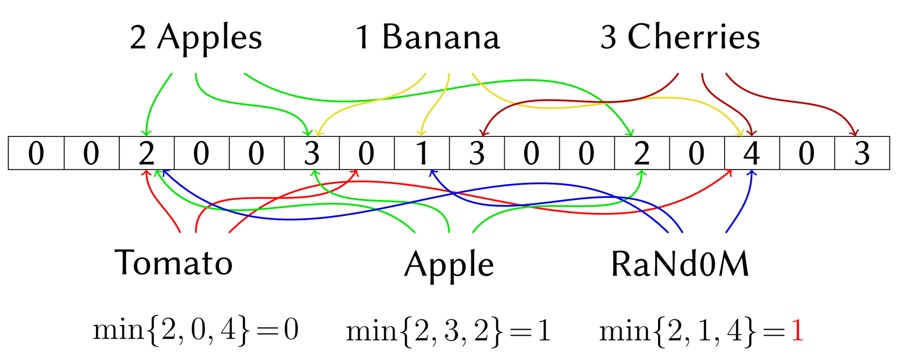
where

| | |
|---|---|
| $f_t(w,l)$ | Observed frequency of pair $(w,l)$ |
| EWMA | Exponentially-weighted moving average |
| EWMVar | Exponentially-weighted moving variance |
| $\beta$ | Laplace-style smoothing term (for rare words) |

Because we cannot afford to store and maintain all $\text{EWMA}[f(w,l)]$ values, we employ a Bloom-filter-like hashing strategy to estimate them efficiently.

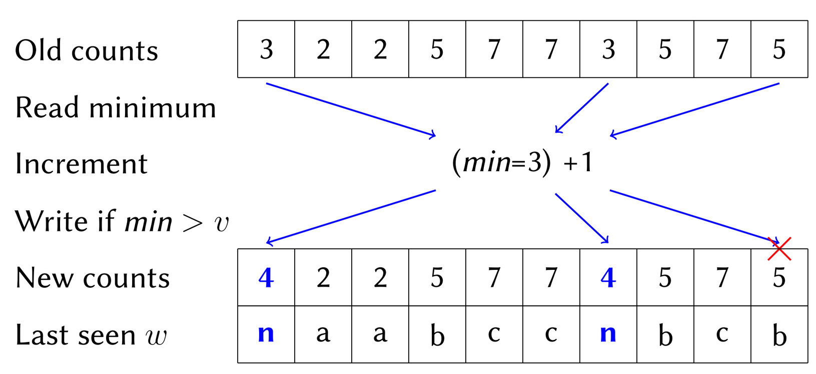
### Bloom-filter / Heavy Hitters

Counting Bloom filters increment each hash bucket. When estimating counts, the minimum found in the buckets is used as estimate. Here, $h = 3$ buckets are used:

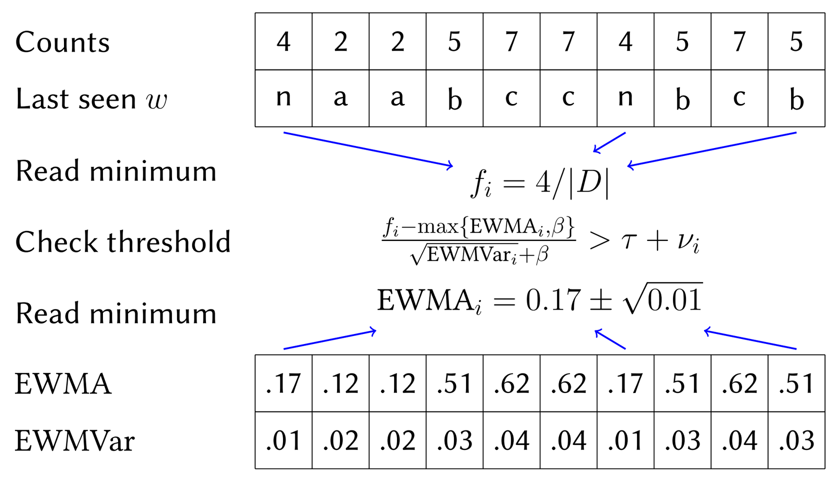



Counting Bloom filters never understimate, but if a term has $h$ hash collisions with *more frequent* terms, it may overestimate the true frequency.

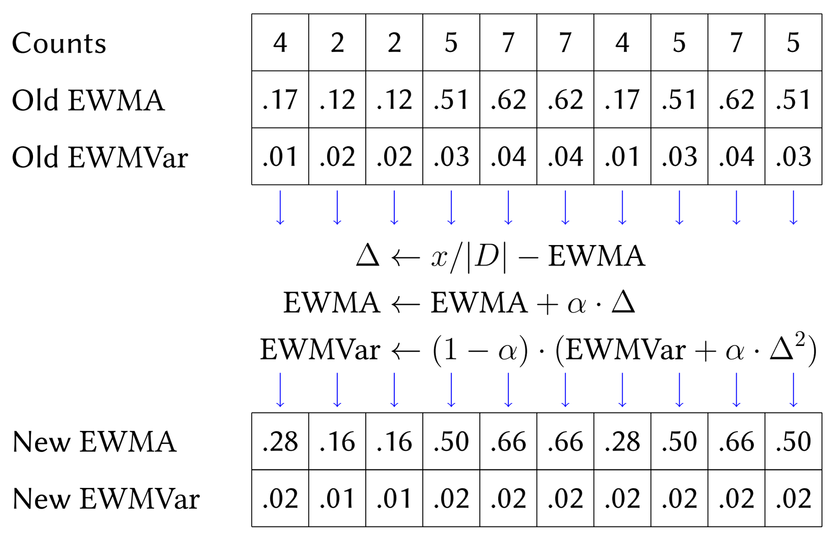
### Hash Table Maintenance




(a) Count-min sketch update (new token: $n$)




(b) Check thresholds for new events




(c) Vectorized statistics table update

## Experiments

### Data Set for SigniTrend Experiments (2014)

News: articles from 2013 of Reuters and Bloomberg news.

Twitter: 114 days of the 1% Twitter sample, originally 279 million tweets before filtering duplicates, retweets, and non-English tweets.

StackOverflow: dump of the main programming Q&A site for years 2010 to 2013.

| Data set | Documents | Paragraphs | Unique Words | Total Words | Unique Pairs | Total Pairs |
|---|---|---|---|---|---|---|
| News | 424,704 | 5,867,457 | 300,141 | 56,661,782 | 71,289,359 | 660,430,059 |
| Twitter | 94,127,149 | 94,127,149 | 25,581,022 | 245,140,695 | 179,105,233 | 473,871,456 |
| StackOverflow | 5,932,320 | 30,423,831 | 2,040,932 | 138,205,636 | 91,460,397 | 545,570,530 |

Data set statistics (after stopword removal)

### Top Events in News 2014 (Chronological)

| | |
|---|---|
| 2014-03-08 | Malaysia Airlines MH-370 missing in South China Sea |
| 2014-04-17 | Russia-Ukraine crisis escalates |
| 2014-04-28 | Soccer World Cup coverage: team lineups |
| 2014-07-17 | Malaysian Airlines MH-17 shot down over Ukraine |
| 2014-07-18 | Russian blamed for 298 dead in airline downing |
| 2014-07-20 | Israel shelling Gaza causes 40+ casualties in a day |
| 2014-08-30 | EU increases sanctions against Russia |
| 2014-10-22 | Ottawa parliament shooting |
| 2014-11-05 | U.S. mid-term elections |
| 2014-12-17 | U.S. and Cuba relations improve unexpectedly |

### Top Events in Twitter (2014)

| Score | Date | Keywords | Explanation |
|---|---|---|---|
| 174 | 03-06 | boosi releas jail | Rapper Lil Boosie released from jail early |
| 154 | 05-28 | rip author poet inspir angelou maya peac dr | Civil rights activist Dr. Maya Angelou died |
| 127 | 05-12 | elev jayz attack jay solang beyonc | Solange, Jay Z and Beyonce elevator incident |
| 98 | 03-03 | ellen degener host selfi pizza | Ellen's Oscar Selfie and Pizza |
| 76 | 05-22 | ewok | Band 5SOS changed its Twitter name to "Ewok Village" |
| 76 | 03-21 | bracket mercer duke | Mercer surprise win over Duke in March Madness |
| 73 | 05-24 | ronaldo bale gareth | Champions league final |
| 63 | 04-07 | geldof dead rip peach | Peaches Geldof died of heroin |
| 61 | 04-15 | moon eclips lunar blood | Blood moon (lunar eclipse) |
| 60 | 05-05 | shovel | Viral video: "shovel girl fight" |

### Data Set for SPOTHOT Experiments (2016)

New geography-oriented data collection:

- 5–6 million geo-tagged tweets per day (no retweets!)
- Estimated 1/3rd of all geo-tagged tweets
- September 10, 2014 to February 19, 2015
- Over 1.1 billion tweets

Selected top geographies:

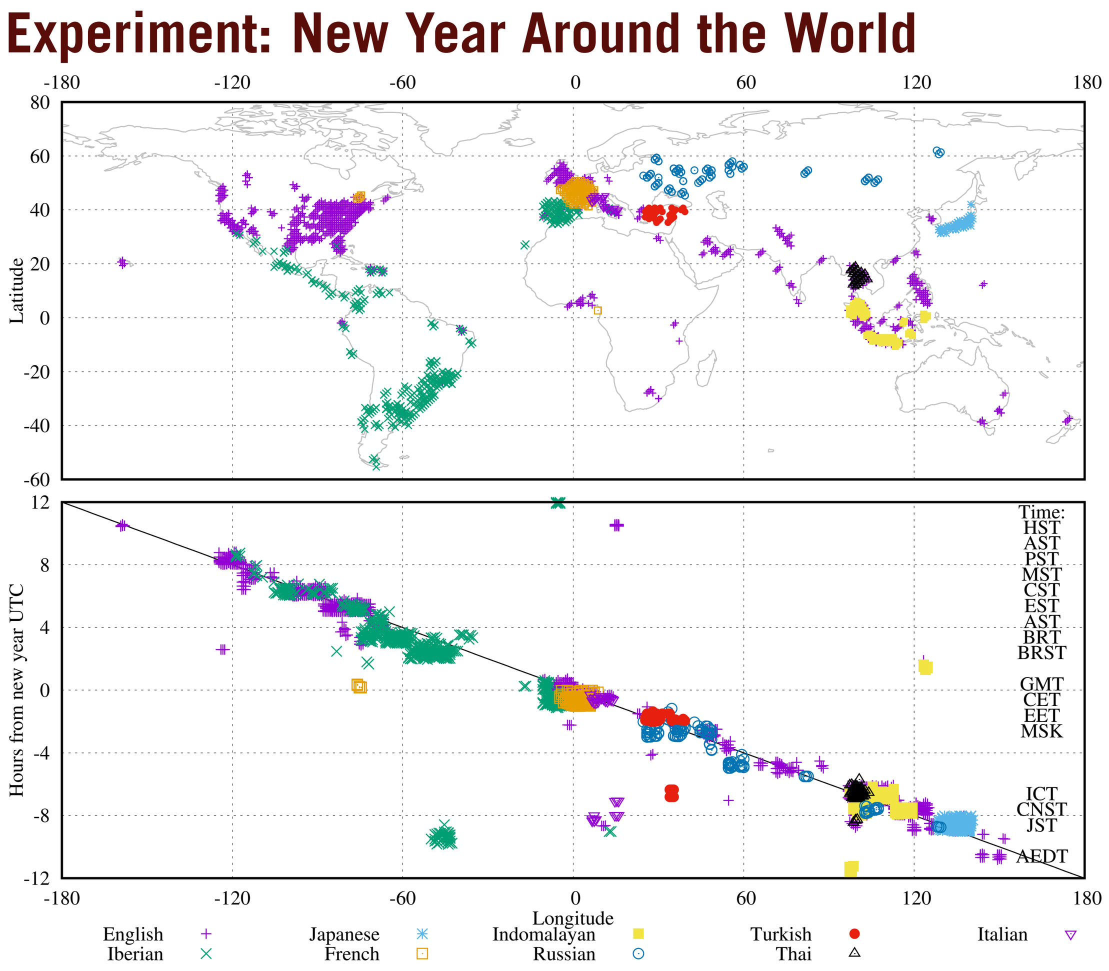
| Region | Mil. | Share | Region | Mil. | Share |
|---|---|---|---|---|---|
| United States | 287.7 | 25.4% | London | 7.6 | 0.67% |
| Brazil | 165.6 | 14.6% | New York City | 7.5 | 0.66% |
| Argentina | 73.6 | 6.5% | Tokyo | 7.4 | 0.66% |
| Indonesia | 72.0 | 6.4% | ⋮ | | |
| Turkey | 59.3 | 5.2% | Germany | 3.5 | 0.31% |
| Japan | 52.4 | 4.6% | ⋮ | | |
| United Kingdom | 49.3 | 4.4% | Berlin | 0.5 | 0.05% |
| ⋮ | | | ⋮ | | |

### Experiment: Most Significant Events

The most significant words each in its most significant location only:

| $\sigma$ | Time | Word | Location | Explanation |
|---|---|---|---|---|
| 2001.8 | 2014-10-29 00:59 | #voteluantvz | Brazil | Brazilian Music Award 2014 |
| 727.8 | 2014-09-23 02:21 | allahımsenbüyüksün | Denizli (Turkey) | Portmanteau used in spam wave |
| 550.1 | 2015-02-02 01:32 | Missy_Elliott | United States of America | Super Bowl Halftime Show |
| 413.5 | 2014-09-18 21:29 | #gala1gh15 | Spain | Spanish Big Brother Launch |
| 412.2 | 2014-11-11 19:29 | #murrayftw | Italy | Teen idol triggered follow spree |
| 293.8 | 2014-10-21 12:05 | #tarıkgüneşttyapıyor | Marmara Region | Hashtag used in spam wave |
| 271.2 | 2015-02-02 02:28 | #masterchefgranfinal | Chile | MasterChef Chile final |
| 268.1 | 2015-01-30 19:28 | #سباركيز | Saudi Arabia | Amusement park "Sparky's" |
| 257.7 | 2014-11-16 21:44 | gemma | United Kingdom | Gemma Collins at jungle camp opening |
| 249.1 | 2014-10-08 02:56 | rosmeri | Argentina | Rosmery González joined Bailando 2014 |
| 223.1 | 2015-01-21 18:51 | otortfv | Central Anatolia Region | Keyword used in spam wave |
| 212.7 | 2014-09-11 18:58 | #catalansvote9n | Catalonia | Catalan referendum requests |
| 208.4 | 2014-12-02 20:00 | #cengizhangençtürk | Northern Borders Region | Hashtag used in spam wave |
| 205.3 | 2015-01-04 15:56 | hairul | Malaysia | Hairul Azreen, Fear Factor Malaysia |
| 198.7 | 2014-12-31 15:49 | あけましておめでとうございます | Japan | New Year in Japan |
| 198.5 | 2015-01-10 20:19 | вк | Russian Federation | "Russian Facebook" VK unavailable |
| 179.7 | 2014-10-04 16:28 | #hormonestheseries2 | Thailand | Hormones: The Series Season 2 |
| 174.7 | 2014-11-28 21:29 | chespirito | Mexico | Comedian "Chespirito" died |
| 160.9 | 2014-09-21 21:27 | #ss5 | Portugal | Secret Story 5 Portugal launch |
| 157.3 | 2014-09-24 01:57 | maluma | Colombia | Maluma on The Voice Kids Colombia |

### Experiment: New Year Around the World





E. Schubert, M. Weiler, and H.-P. Kriegel. "SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds". In: *Proc. KDD.* 2014, pp. 871–880

E. Schubert, M. Weiler, and H.-P. Kriegel. "SPOTHOT: Scalable Detection of Geo-spatial Events in Large Textual Streams". In: *Proc. SSDBM.* 2016, 8:1–8:12