Advanced Topics in Information Retrieval

9. Social Media

Vinay Setty
(vsetty@mpi-inf.mpg.de)

Jannik Strötgen
(jtroetge@mpi-inf.mpg.de)

Outline

9.1. What is Social Media?

9.2. Tracking Memes

9.3. Opinion Retrieval

9.4. Feed Distillation

9.5. Top-Story Identification

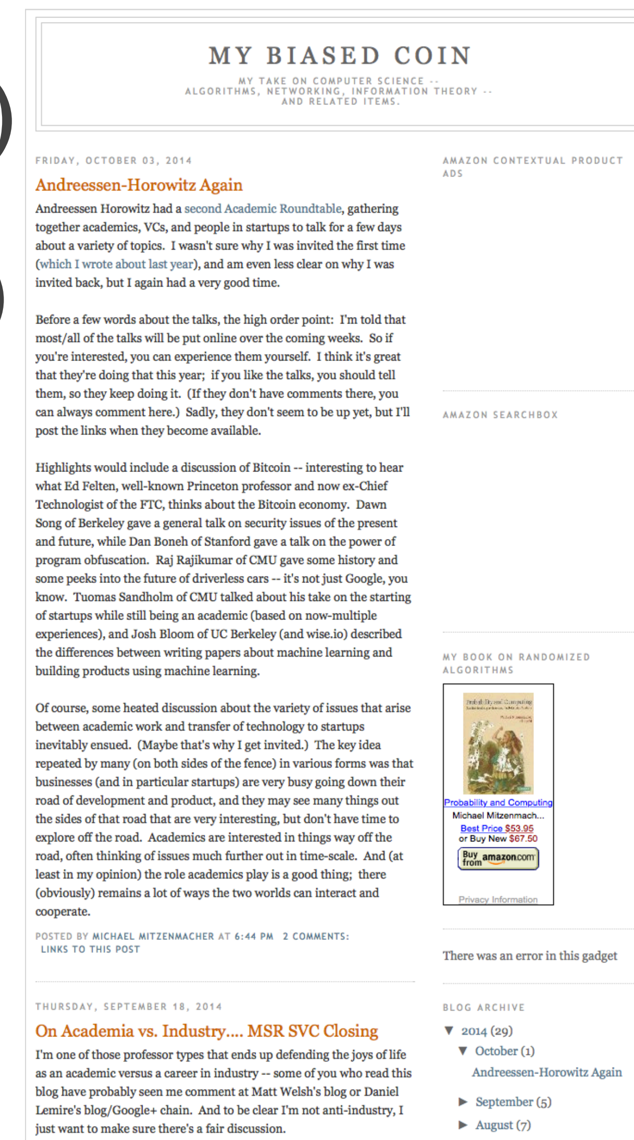
9.1. What is Social Media?

- Content creation is **supported by software** (no need to know HTML, CSS, JavaScript)
- Content is **user-generated** (as opposed to by big publishers) or **collaboratively-edited** (as opposed to by a single author)
- **Web 2.0** (if you like –outdated– buzzwords)
- Examples:
  - **Blogs** (e.g., Wordpress, Blogger, Tumblr)
  - **Social Networks** (e.g., facebook, Google+)
  - **Wikis** (e.g., Wikipedia but there are many more)
  - …

= ?!? =

Weblogs, Blogs, the Blogosphere

- **Journal-like website**, editing supported by software, self-hosted or as a service
- Initially often run by **enthusiasts**, now also common in the **business world**, and some bloggers make their living from it
- **Reverse chronological order** (newest first)
- **Blogroll** (whose blogs does the blogger read)
- **Posts** of varying length and topics
- **Comments**
- Backed by **XML feed** (e.g., RSS or Atom) for **content syndication**



max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Weblogs, Blogs, the Blogosphere

MY BIASED COIN

MY TAKE ON COMPUTER SCIENCE -- ALGORITHMS, NETWORKING, INFORMATION THEORY -- AND RELATED ITEMS.

FRIDAY, OCTOBER 03, 2014

Andreessen-Horowitz Again

Andreessen Horowitz had a second Academic Roundtable, gathering together academics, VCs, and people in startups to talk for a few days about a variety of topics. I wasn't sure why I was invited the first time (which I wrote about last year), and am even less clear on why I was invited back, but I again had a very good time.

Before a few words about the talks, the high order point: I'm told that most/all of the talks will be put online over the coming weeks. So if you're interested, you can experience them yourself. I think it's great that they're doing that this year; if you like the talks, you should tell them, so they keep doing it. (If they don't have comments there, you can always comment here.) Sadly, they don't seem to be up yet, but I'll post the links when they become available.

Highlights would include a discussion of Bitcoin -- interesting to hear what Ed Felten, well-known Princeton professor and now ex-Chief Technologist of the FTC, thinks about the Bitcoin economy. Dawn Song of Berkeley gave a general talk on security issues of the present and future, while Dan Boneh of Stanford gave a talk on the power of program obfuscation. Raj Rajikumar of CMU gave some history and some peeks into the future of driverless cars -- it's not just Google, you know. Tuomas Sandholm of CMU talked about his take on the starting of startups while still being an academic (based on now-multiple experiences), and Josh Bloom of UC Berkeley (and wise.io) described the differences between writing papers about machine learning and building products using machine learning.

Of course, some heated discussion about the variety of issues that arise between academic work and transfer of technology to startups inevitably ensued. (Maybe that's why I get invited.) The key idea repeated by many (on both sides of the fence) in various forms was that businesses (and in particular startups) are very busy going down their road of development and product, and they may see many things out the sides of that road that are very interesting, but don't have time to explore off the road. Academics are interested in things way off the road, often thinking of issues much further out in time-scale. And (at least in my opinion) the role academics play is a good thing; there (obviously) remains a lot of ways the two worlds can interact and cooperate.

POSTED BY MICHAEL MITZENMACHER AT 6:44 PM 2 COMMENTS:
LINKS TO THIS POST

THURSDAY, SEPTEMBER 18, 2014

On Academia vs. Industry.... MSR SVC Closing

I'm one of those professor types that ends up defending the joys of life as an academic versus a career in industry -- some of you who read this blog have probably seen me comment at Matt Welsh's blog or Daniel Lemire's blog/Google+ chain. And to be clear I'm not anti-industry, I just want to make sure there's a fair discussion.

AMAZON CONTEXTUAL PRODUCT ADS

AMAZON SEARCHBOX

MY BOOK ON RANDOMIZED ALGORITHMS

Probability and Computing
Michael Mitzenmach...
Best Price $53.95
or Buy New $67.50
Buy from amazon.com
Privacy Information

There was an error in this gadget

BLOG ARCHIVE

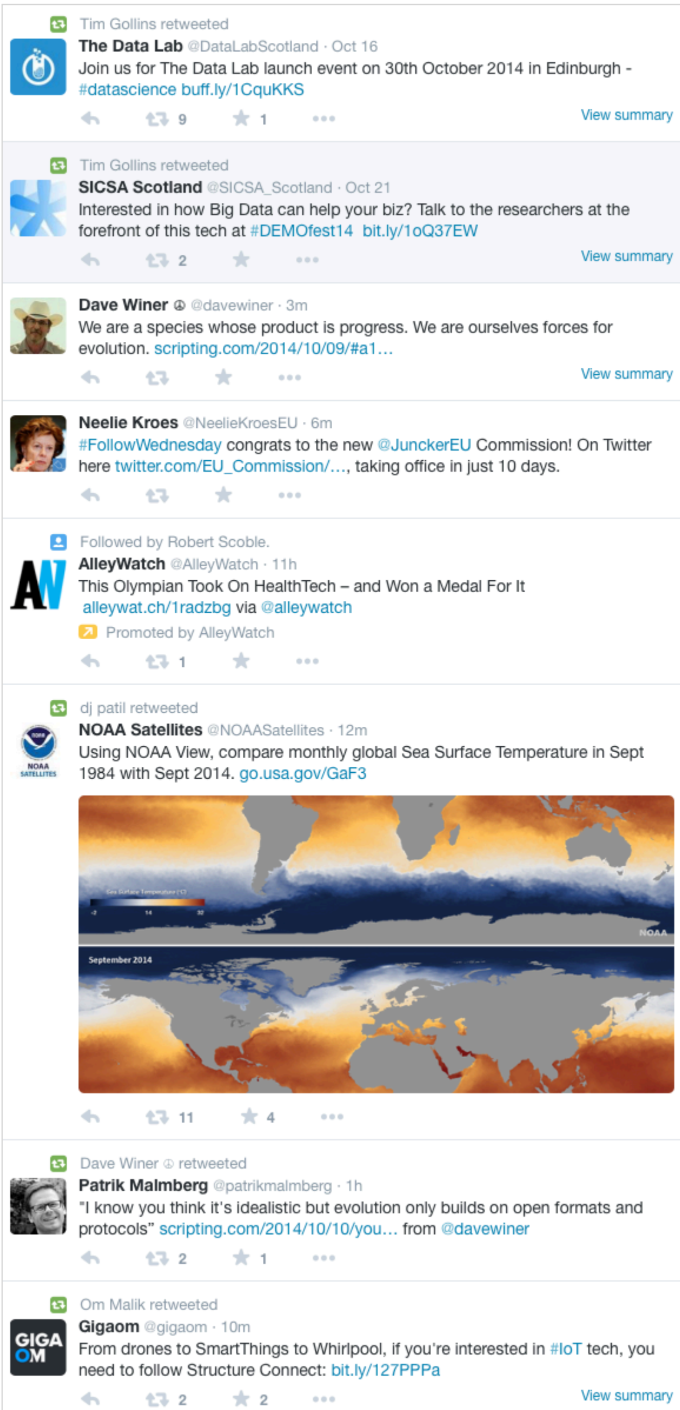
▼ 2014 (29)
▼ October (1)
Andreessen-Horowitz Again
► September (5)
► August (7)

http://mybiasedcoin.blogspot.de

- WordPress.com
  - ~ 60M blogs
  - ~ 50M posts/month
  - ~ 50M comments/month
- Tumblr.com (by Yahoo!)
  - ~ 208M blogs
  - ~ 95B posts
  - ~ 100M posts/day

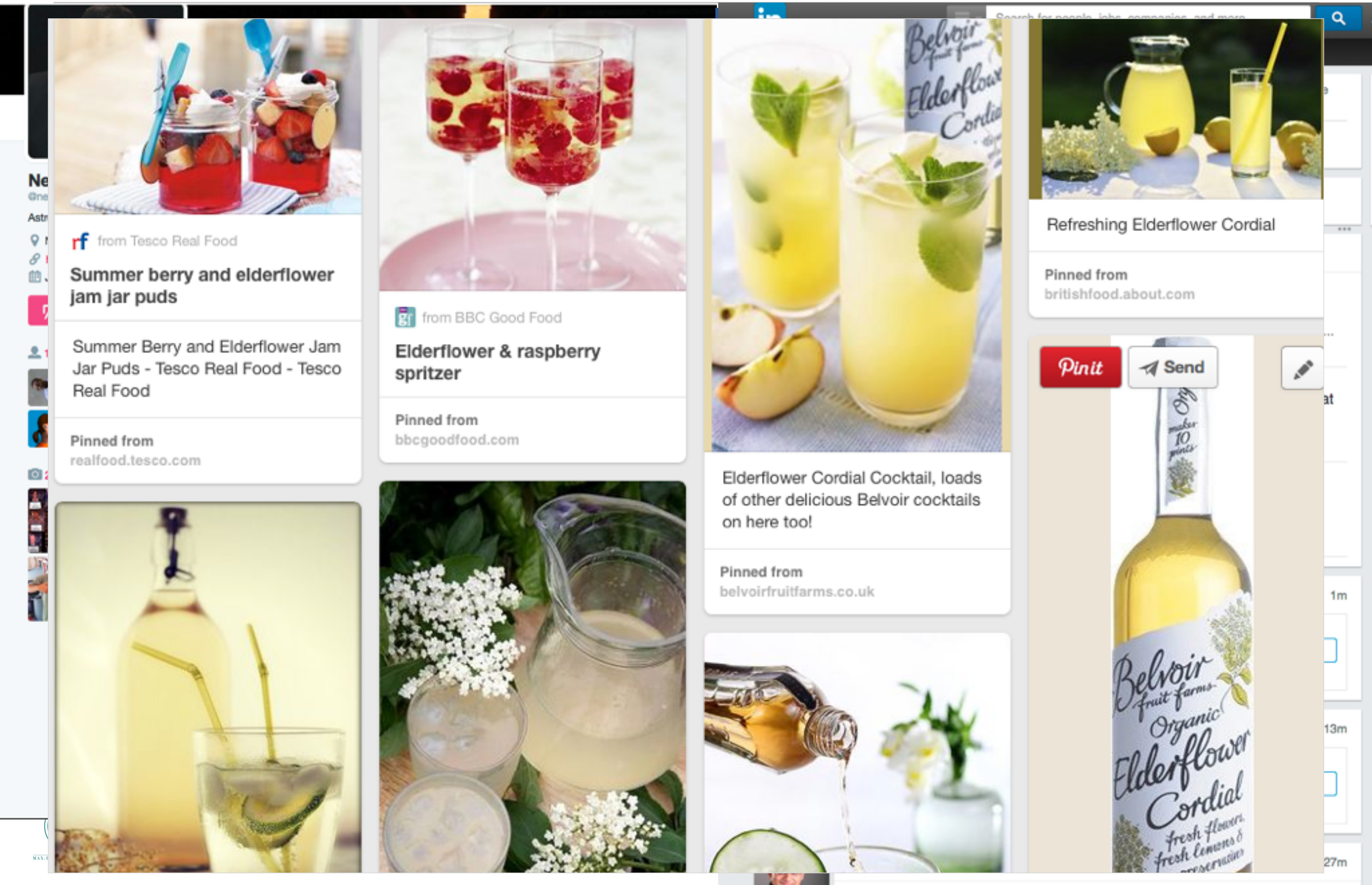
max planck institut informatik

Twitter



- Micro-blogging service created in March '06
- Posts (tweets) limited to **140 characters**
- **271M** monthly active **users**
- **500M tweets/day = ~6K tweets/second**
- **2B queries** per day
- 77% of accounts are outside of the U.S.
- Hashtags (#atir2016)
- Messages (@vinaysetty)
- Retweets

Facebook, Twitter, LinkedIn, Pinterest, …



Challenges & Opportunities

- Content
  - **plenty of context** (e.g., publication timestamp, relationships between users, user profiles, comments, external urls)
  - **short posts** (e.g., on Twitter), **colloquial/cryptic language**
  - **spam** (e.g., splogs, fake accounts)
- Dynamics
  - **up-to-date content** – real-world events covered as they happen
  - **high update rates** pose severe engineering challenges (e.g., how to maintain indexes and collection statistics)

max planck institut
informatik

How do People Search Blogs?

- Mishne and de Rijke [8] analyzed a **month-long query log** from a blog search engine (blogdigger.com) and found that
  - queries are **mostly informational** (vs. transactional or navigational)
    - **contextual**: in which context is a specific **named entity** (i.e., person, location, organization) mentioned, for instance, to find out opinions about it
    - **conceptual**: which blogs cover a specific **high-level concept or topic** (e.g., stock trading, gay rights, linguists, islam)
    - contextual more common than conceptual both for **ad-hoc and filtering queries**
  - most popular topics: **technology, entertainment, and politics**
  - many queries (15–20%) **related to current events**

How do People Search Twitter?

- Teevan et al. [10] **conducted a survey** (54 MS employees), compared **query logs** from web search and Twitter, finding that queries on Twitter
  - are often related to **celebrities, memes, or other users**
  - are **often repeated** to monitor a specific topic
  - are on average **shorter** than web queries (1.64 vs. 3.08 words)
  - tend to return **results** that are **shorter** (19.55 vs. 33.95 words), **less diverse**, and more often relate to **social gossip** and **recent events**
- People also directly **express information needs** using Twitter: **17% of tweets** in the analyzed data correspond to **questions**

max planck institut informatik

What Data?

- **Feeds** (e.g., blog, twitter user, facebook page)
- **Posts** (e.g., blog posts, tweets, facebook posts)
- We'll consider
  - **textual content** of posts
  - **publication timestamps** of posts
  - **hyperlinks** contained in posts
- We'll ignore
  - other links (e.g., friendship, follower/followee)
  - hashtags, images, comments

Tasks

- **Meme tracking** grouping of memes to track them over period of time
- **Post retrieval** identifies posts relevant to a specific information need (e.g., how is life in Iceland?)
- **Opinion retrieval** finds posts **relevant** to a specific **named entity** (e.g., a company or celebrity) which **express an opinion** about it
- **Feed distillation** identifies feeds relevant to a topic, so that the user can subscribe to their posts (e.g., who tweets about C++?)
- **Top-story identification** leverages social media to determine the most important news stories (e.g., to display on front page)

Outline

9.1. What is Social Media?

9.2. Tracking Memes

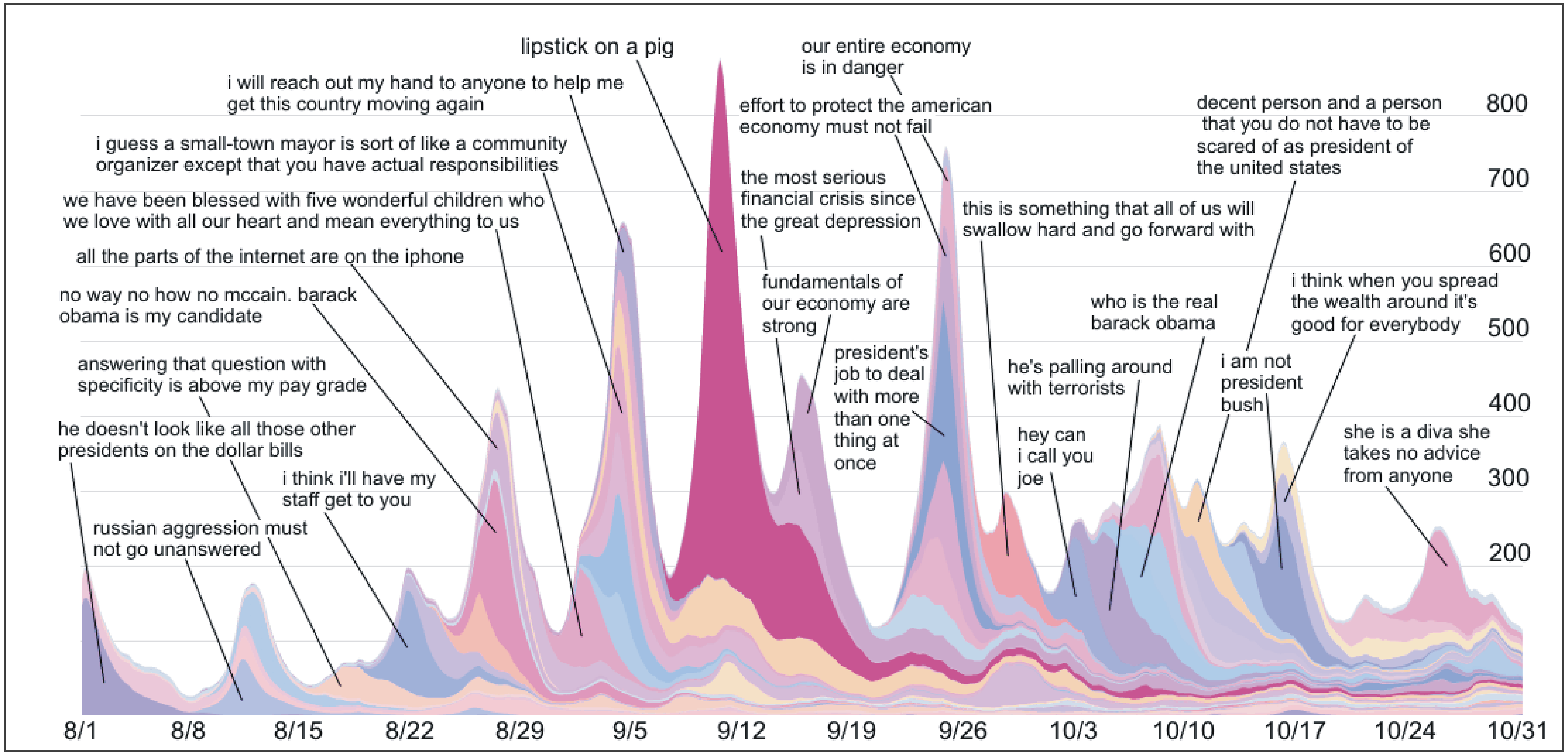
9.3. Opinion Retrieval

9.4. Feed Distillation

9.5. Top-Story Identification

9.2. Tracking Memes

- Leskovec et al. [5] track **memes** (e.g., "lipstick on a pig") and visualize their volume in traditional news and blogs



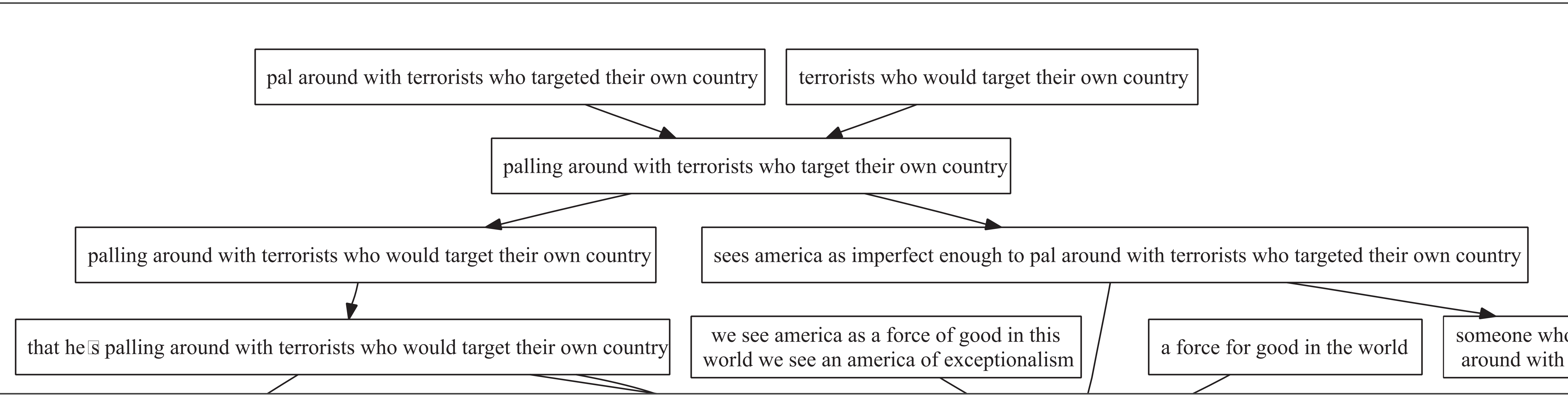
- Demo: http://www.memetracker.org

max planck institut informatik

Phrase Graph Construction

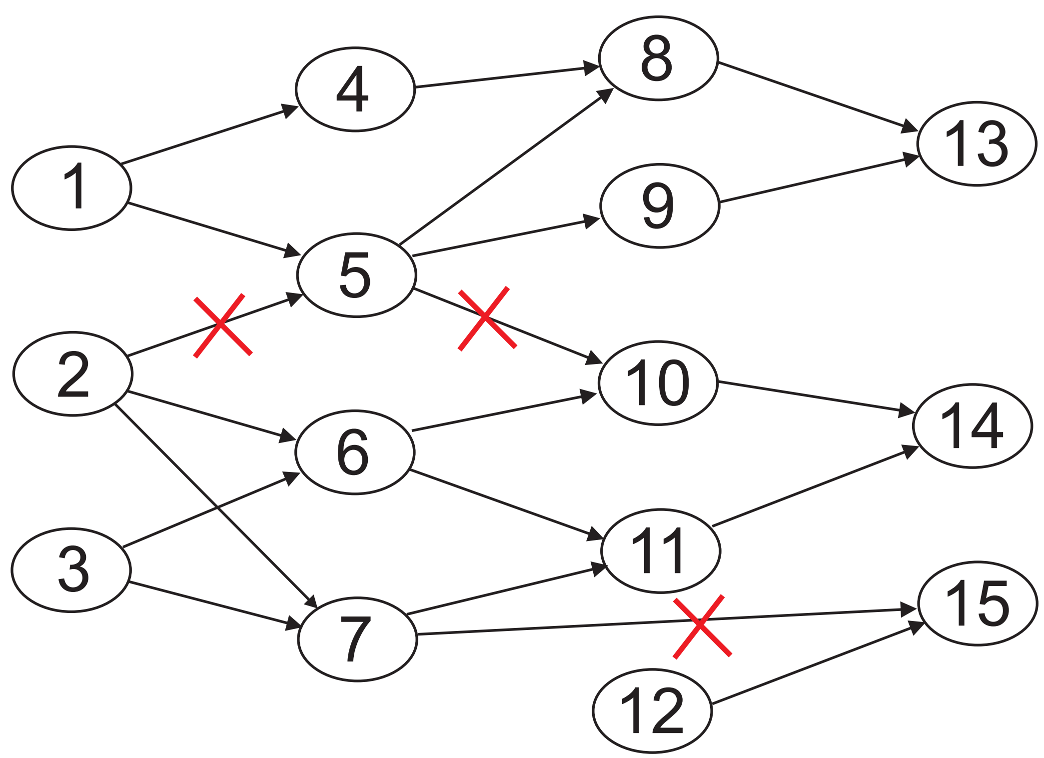
- Problem: Memes are **often modified** as they spread, so that first **all mentions of the same meme** need to be identified
- Construction of a **phrase graph** G(V, E):
  - **vertices** V correspond to **mentions of a meme** that are reasonably long and occur often enough
  - **edge** (u,v) exists if meme mentions u and v
    - u is **strictly shorter** than v
    - either: have **small directed token-level edit distance** (i.e., **u** can be transformed into **v** by adding at most **ε** tokens)
    - or: have a **common word sequence** of length at least **k**
  - **edge weights** based on **edit distance** between **u** and **v** and how often **v** occurs in the document collection

Meme Phrase Graph



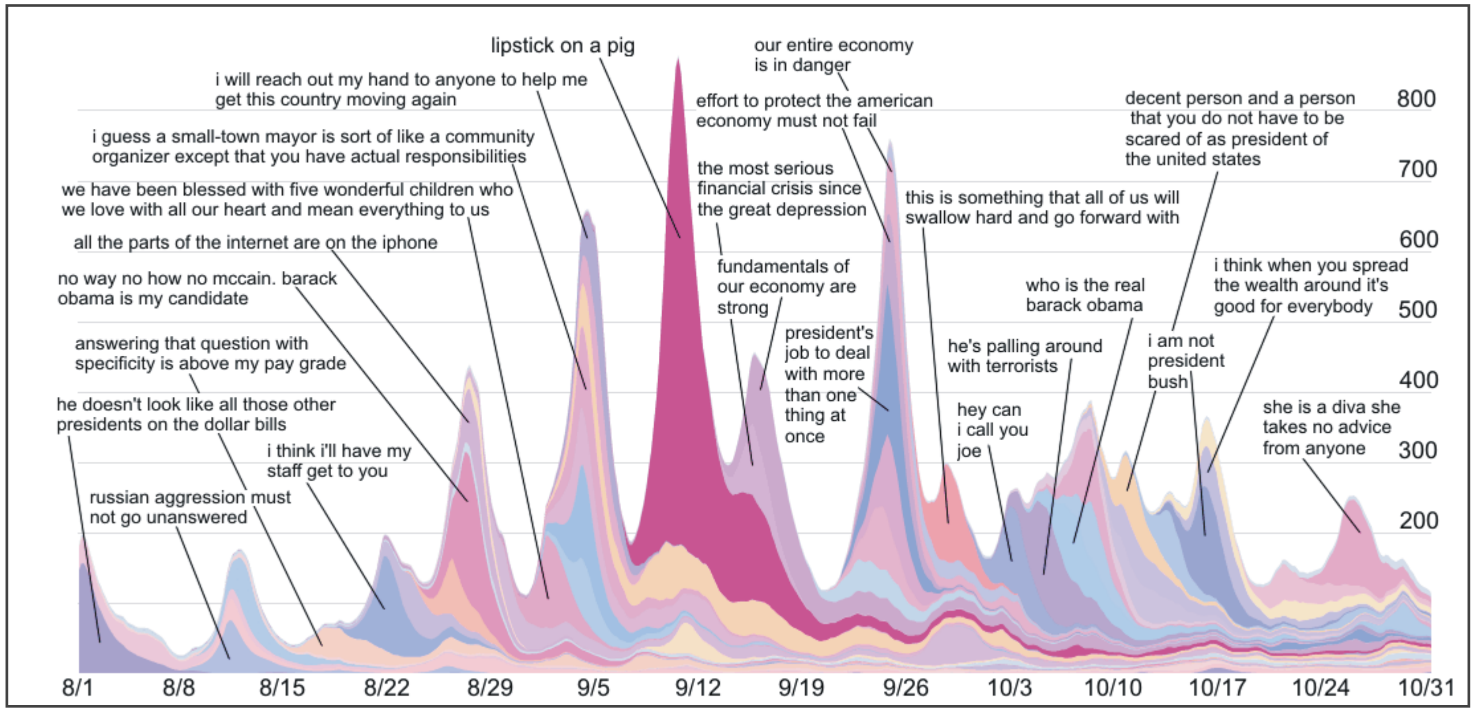
Phrase Graph Partitioning

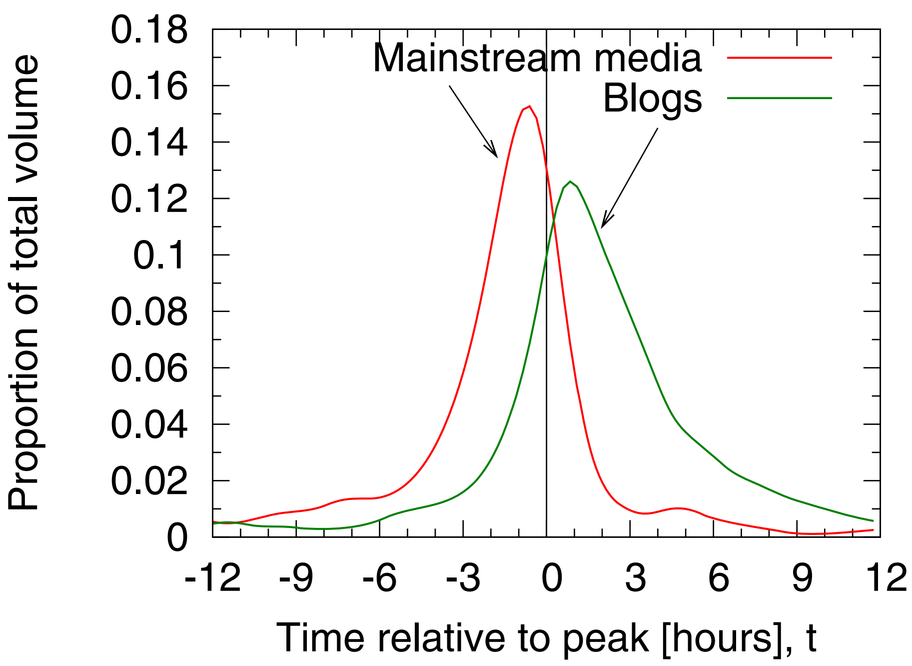
- Phrase graph is an **directed acyclic graph** (DAG) by construction
- Partition G(V, E) by **deleting a set of edges having minimum total weight**, so that each resulting **component is single-rooted**
- Phrase graph partitioning is *NP*-hard, hence addressed by **greedy heuristic algorithm**



Applications

- **Clustering of meme mentions** allows for insightful analyses, e.g.:
  - **volume of meme** per time interval
  - **peak time** of meme in traditional news and social media
  - **time lag** between peek times in traditional news and social media





Outline

9.1. What is Social Media?

9.2. Tracking Memes

9.3. Opinion Retrieval

9.4. Feed Distillation

9.5. Top-Story Identification

9.3. Opinion Retrieval

- **Opinion retrieval** finds posts **relevant** to a specific **named entity** (e.g., a company or celebrity) which **express an opinion** about it
- Examples: (from TREC Blog track 2006)
  - macbook pro
  - jon stewart
  - whole foods
  - mardi gras
  - cheney hunting

Title:
whole foods

Description:
Find opinions on the quality, expense, and value of purchases at Whole Foods stores.

Narrative:
All opinions on the quality, expense and value of Whole Foods purchases are relevant. Comments on business and labor practices or Whole Foods as a stock investment are not relevant. Statements of produce and other merchandise carried by Whole Foods without comment are not relevant.

- **Standard retrieval models** can help with finding relevant posts; but how to determine **whether a post expresses an opinion?**

Opinion Retrieval Task Example

```
<top>
<num> Number: 863

<title> netflix

<desc> Description:
Identify documents that show customer opinions
of Netflix.

<narr> Narrative:
A relevant document will indicate subscriber
satisfaction with Netflix.  Opinions about
the Netflix DVD allocation system, promptness
or delay in mailings are relevant.
Indications of having been or
intent to become a Netflix subscriber that do
not state an opinion are not relevant.
</top>
```

max planck institut
informatik

Opinion Dictionary

- What if we had a **dictionary of opinion words?** (e.g., like, good, bad, awesome, terrible, disappointing)
- Lexical resources with **word sentiment information**
  - **SentiWordNet** (http://sentiwordnet.isti.cnr.it/)

  - **General Inquirer** (http://www.wjh.harvard.edu/~inquirer/)
  - **OpinionFinder** (http://mpqa.cs.pitt.edu)

max planck institut informatik

Opinion Dictionary

- He et al. [4] construct an **opinion dictionary** from training data
  - consider only words that are neither too frequent (e.g., and, or) nor too rare (e.g., aardvark) in the post collection D
  - let D_{rel} be a set of **relevant posts** (to any query in a workload) and $D_{relopt} \subset D_{rel}$ be the subset of **relevant opinionated posts**
  - two options to **measure opinionatedness of a word v**
    - **Kullback-Leibler Divergence**

$$op_{KLD}(v) = P[\, v \mid D_{relopt} \,] \log_2 \frac{P[\, v \mid D_{relopt} \,]}{P[\, v \mid D_{rel} \,]}$$

    - **Bose Einstein Statistics**

$$op_{BO}(v) = tf(v, D_{relopt}) \log_2 \frac{1+\lambda}{\lambda} + \log_2 (1+\lambda) \quad \text{with} \quad \lambda = \frac{tf(v, D_{rel})}{|D_{rel}|}$$

Re-Ranking

- He et al. [4] **measure opinionatedness of a post** d as follows
  - consider the set Q_{opt} of k **most opinionated words** from the dictionary
  - issue Q_{opt} as a query (e.g., using Okapi BM25 as a retrieval model)
  - the retrieval status value score(d, Q_{opt}) measures how opinionated d is
- Posts are ranked in response to query Q (e.g., whole foods) according to a (linear) **combination of retrieval scores**

$$score(d) = \alpha \cdot score(d, Q) + (1 - \alpha) \cdot score(d, Q_{opt})$$

with $0 \leq \alpha \leq 1$ as a **tunable mixing parameter**

Sentiment Expansion

- Huang and Croft [5] **expand the query** with **query-independent** (Q_I) and **query-dependent** (Q_D) opinion words; posts are then ranked according to

$$score(d) = \alpha \cdot score(d, Q) + \beta \cdot score(d, Q_I) \\ + (1 - \alpha - \beta) \cdot score(d, Q_D)$$

with $0 \leq \alpha, \beta \leq 1$ as a **tunable mixing parameters** and retrieval scores based on **language model divergences**

- **Query-independent opinion words** are obtained as
  - **seed words** (e.g, good, nice, excellent, poor, negative, unfortunate, …)
  - **most frequent words** in opinionated corpora (e.g., movie reviews)

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Sentiment Expansion (Query Independent)

- Examples: (of most frequent words in different corpora)
  - **Cornell movie reviews**: like, even, good, too, plot
  - **MPQA opinion corpus**: against, minister, terrorism, even, like
  - **Blog06(op)**: like, know, even, good, too
- Observation: Query-independent opinion words are very general (e.g., like, good) or specific to the corpus (e.g., minister, terrorism)

Sentiment Expansion (Query Dependent)

- **Query-dependent opinion words** are obtained as words that frequently co-occur with query terms in **pseudo-relevant documents** (following the approach by Lavrenko and Croft [6])
- Given a query q, identify the set of R of top-k pseudo-relevant documents, and top-n words having highest probability

$$P[\,w \,|\, R\,] \propto \sum_{d \,\in\, R} P[\,w \,|\, d\,] \prod_{v \,\in\, q} P[\,v \,|\, d, w\,]$$

$$P[\,v \,|\, d, w\,] = \begin{cases} \frac{tf(v,d)}{\sum_u tf(u,d)} & : \quad w \in d \\ 0 & : \quad \text{otherwise} \end{cases}$$

with parameter set as k = 5 and n = 20 in practice

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Sentiment Expansion

- Examples: (of query-dependent opinion words)
  - mozart → (like, good, too, even, death, best, great, genius)
  - allianz → (best, premium, great, value, traditional, fidelity)
  - wikipedia → (like, open, good, know, free, great, knowledge)

max planck institut
informatik

Outline

9.1. What is Social Media?

9.2. Tracking Memes

9.3. Opinion Retrieval

9.4. Feed Distillation

9.5. Top-Story Identification

max planck institut
informatik

9.4. Feed Distillation

- **Feed distillation** identifies feeds (e.g., blogs, Twitter users) that are **relevant** to a **specific (typically rather broad) topic**

- Examples: (from TREC Blog track 2007)

  - movie review
  - firearm control
  - baseball
  - garden
  - mobile phone

> **Title:**
> baseball
>
> **Description:**
> Blogs with recurring interests in Major League Baseball, or lesser leagues, for example, giving news or analysis of games or player moves.
>
> **Narrative:**
> Relevant blogs will have news or analysis from the major league baseball and other leagues. Blogs listing only product reviews, or with other nonsensical information are not relevant.

- Challenges: How to capture whether a blog **consistently covers** the given topic? How to bridge **vocabulary gap** to posts?

Language Models

- Weerkamp et al. [11] develop two approaches to feed distillation estimating **language models** for **entire blog(ger)s** and **individual posts**, respectively

- Notation:
  - a blog b is a set of posts; |b| is the number of posts by b
  - a post p is a bag of terms
  - tf(v, p) denotes the term frequency of term v in post p
  - B denotes a virtual post concatenating all posts from all blogs

max planck institut
informatik

Blogger Model (BM)

- Estimates a language model **for each blog(ger)** b

$$P[\, q \mid \theta_b \,] = \prod_{v \in q} P[\, v \mid \theta_b \,]^{\, tf(v,q)}$$

- Smooths probability estimates using the collection of blogs B

$$P[\, v \mid \theta_b \,] = (1 - \lambda_b) \cdot P[\, v \mid b \,] + \lambda_b \cdot P[\, v \mid B \,]$$

with **blog-specific smoothing parameter**

$$\lambda_b = \frac{\beta}{(1/|b| \cdot \sum_{p \in b} \sum_v tf(v,p)) + \beta}$$

thus smoothing blogs with **shorter posts more aggressively**

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Blogger Model

- **Two-step generation** of term v from blog b

$$P[\, v \mid b \,] = \sum_{p \in b} P[\, v \mid p, b \,] \, P[\, p \mid b \,]$$

assuming **conditional independence** of terms given blog

$$P[\, v \mid b \,] = \sum_{p \in b} P[\, \underbrace{v \mid p}_{\text{2. Draw term from post}} \,] \, P[\, \underbrace{p \mid b}_{\text{1. Draw post from blog}} \,]$$

- **Uniform probability** of posts given blog (i.e., equal importance)

$$P[\, p \mid b \,] = 1/|b|$$

- **Maximum-likelihood estimate** $P[\, v \mid p \,] = \dfrac{tf(v,p)}{\sum_w tf(w,p)}$

max planck institut informatik

Posting Model (PM)

- Estimates a language model **for each individual post** p

$$P[\,v \mid \theta_p\,] = (1-\lambda_p)\cdot P[\,v \mid p\,] + \lambda_p \cdot P[\,v \mid B\,]$$

with **post-specific smoothing parameter**

$$\lambda_p = \frac{\beta}{\left(\sum_w tf(w,p)\right) + \beta}$$

thus smoothing **short posts more aggressively**

- **Maximum-likelihood estimate** $P[\,v \mid p\,] = \frac{tf(v,p)}{\sum_w tf(w,p)}$

MAX-PLANCK-GESELLSCHAFT

max planck institut
informatik

Posting Model

- Likelihood of generating query q from language model of post p

$$P[\, q \mid \theta_p \,] = \prod_{v \in q} P[\, v \mid \theta_p \,]^{\, tf(v,q)}$$

- **Two-step generation** of query q from blog b

$$P[\, q \mid b \,] = \sum_{p \in b} P[\, \underbrace{q \mid \theta_p}_{\text{2. Generate query from post}} \,]\, P[\, \underbrace{p \mid b}_{\text{1. Draw post from blog}} \,]$$

- **Uniform probability** of posts given blog (i.e., equal importance)

$$P[\, p \mid b \,] = 1 / |b|$$

Query Expansion for Vocabulary Gap

- Elsass et al. [3] proposed the highly similar **Large Document Model** (~BM) and **Small Document Model** (~PM) approaches
- Focus on bridging the **vocabulary gap** between high-level topic descriptions (e.g., garden) and posts (e.g., seed, flower, crop)
- **Query expansion** with terms from **pseudo-relevant documents** retrieved from different corpora
  - **Blogs** (MAP 0.266 compared to small document model 0.315)
  - **Posts** (MAP 0.282)
  - **Wikipedia articles** (MAP 0.314)
  - **Wikipedia passages** (MAP 0.313)

NO IMPROVEMENT!

max planck institut informatik

MAX-PLANCK-GESELLSCHAFT

Query Expansion for Vocabulary Gap

- **Query expansion** based on **anchor phrases** in Wikipedia
  - **issue original query** q against Wikipedia articles as corpus
  - **consider** top-k and top-n (k < n) **results** returned by query
  - **score every anchor phrase** a occurring in any top-n result and pointing to a document d from the top-k result as

$$score(a) = \sum_{\underbrace{(a,d)}} (k - rank(d))$$

anchor phrase **a** from top-**n** article pointing to top-**k** article **d**

http://en.wikipedia.org/wiki/United_States

united states

united states of america

america

land of the free

the states

favoring **frequent anchor phrases** pointing to **highly ranked articles**

  - **expand query** with top-m anchor phrases (MAP 0.361)

IMPROVEMENT!

max planck institut
informatik

Outline

max planck institut
informatik

Online News Media

BBC NEWS

The New York Times

CNN

THE HUFFINGTON POST

REUTERS

theguardian

Sunday Times

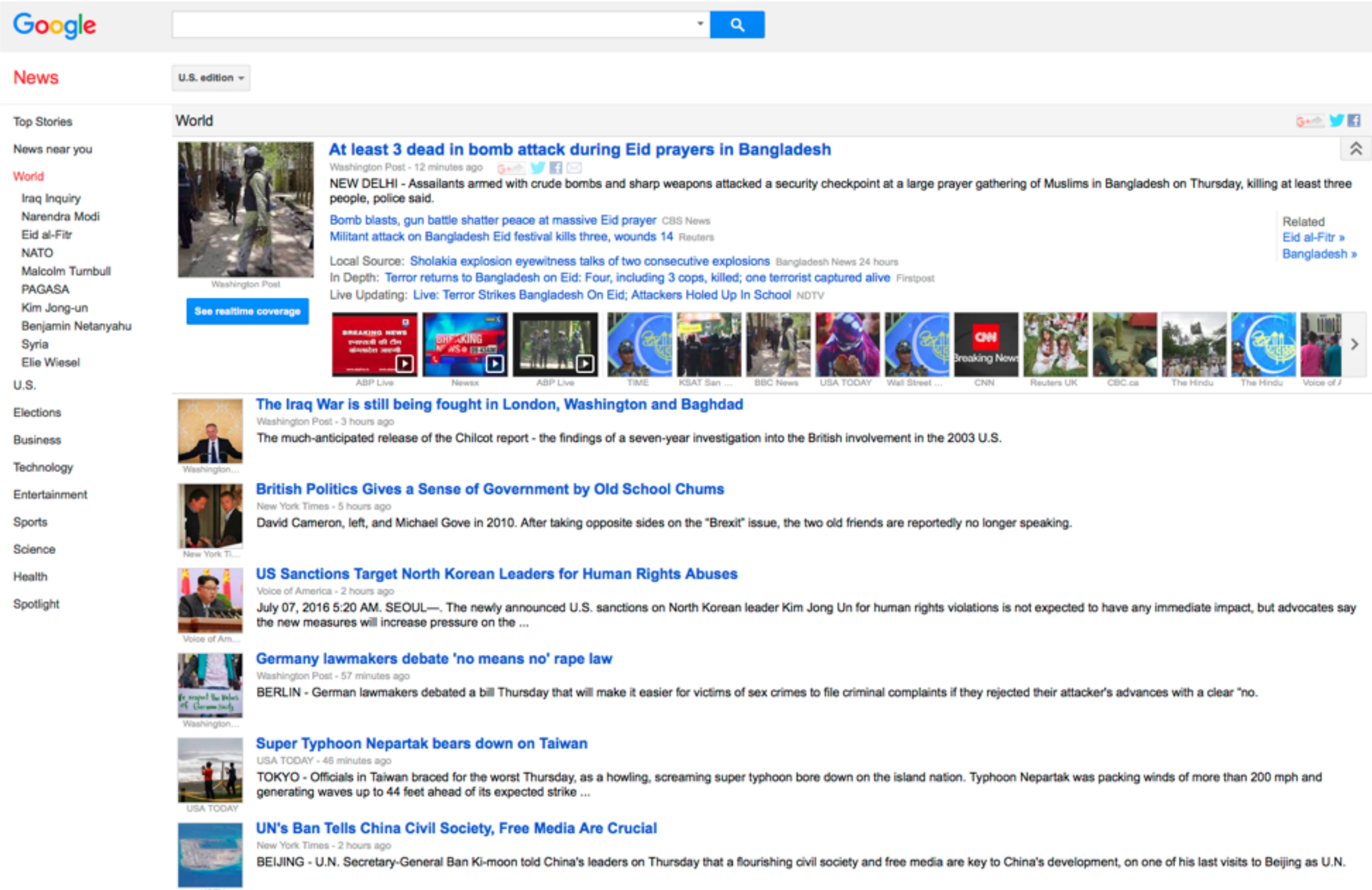
USA TODAY

A GANNETT COMPANY

THE TIMES OF INDIA

Thousands of news articles generated each day!

Google News


Google
News
U.S. edition
Top Stories
News near you
World
Iraq Inquiry
Narendra Modi
Eid al-Fitr
NATO
Malcolm Turnbull
PAGASA
Kim Jong-un
Benjamin Netanyahu
Syria
Elie Wiesel
U.S.
Elections
Business
Technology
Entertainment
Sports
Science
Health
Spotlight
World
At least 3 dead in bomb attack during Eid prayers in Bangladesh
Washington Post - 12 minutes ago
NEW DELHI - Assailants armed with crude bombs and sharp weapons attacked a security checkpoint at a large prayer gathering of Muslims in Bangladesh on Thursday, killing at least three people, police said.
Bomb blasts, gun battle shatter peace at massive Eid prayer CBS News
Militant attack on Bangladesh Eid festival kills three, wounds 14 Reuters
Local Source: Sholakia explosion eyewitness talks of two consecutive explosions Bangladesh News 24 hours
In Depth: Terror returns to Bangladesh on Eid: Four, including 3 cops, killed; one terrorist captured alive Firstpost
Live Updating: Live: Terror Strikes Bangladesh On Eid; Attackers Holed Up In School NDTV
Related
Eid al-Fitr »
Bangladesh »
Washington Post
See realtime coverage
ABP Live
Newsx
ABP Live
TIME
KSAT San ...
BBC News
USA TODAY
Wall Street ...
CNN
Reuters UK
CBC.ca
The Hindu
The Hindu
The Iraq War is still being fought in London, Washington and Baghdad
Washington Post - 3 hours ago
The much-anticipated release of the Chilcot report - the findings of a seven-year investigation into the British involvement in the 2003 U.S.
British Politics Gives a Sense of Government by Old School Chums
New York Times - 5 hours ago
David Cameron, left, and Michael Gove in 2010. After taking opposite sides on the "Brexit" issue, the two old friends are reportedly no longer speaking.
US Sanctions Target North Korean Leaders for Human Rights Abuses
Voice of America - 2 hours ago
July 07, 2016 5:20 AM. SEOUL—. The newly announced U.S. sanctions on North Korean leader Kim Jong Un for human rights violations is not expected to have any immediate impact, but advocates say the new measures will increase pressure on the ...
Germany lawmakers debate 'no means no' rape law
Washington Post - 57 minutes ago
BERLIN - German lawmakers debated a bill Thursday that will make it easier for victims of sex crimes to file criminal complaints if they rejected their attacker's advances with a clear "no.
Super Typhoon Nepartak bears down on Taiwan
USA TODAY - 46 minutes ago
TOKYO - Officials in Taiwan braced for the worst Thursday, as a howling, screaming super typhoon bore down on the island nation. Typhoon Nepartak was packing winds of more than 200 mph and generating waves up to 44 feet ahead of its expected strike ...
USA TODAY
UN's Ban Tells China Civil Society, Free Media Are Crucial
New York Times - 2 hours ago
BEIJING - U.N. Secretary-General Ban Ki-moon told China's leaders on Thursday that a flourishing civil society and free media are key to China's development, on one of his last visits to Beijing as U.N.

News Aggregators

Portal:Current events

digg

YAHOO! News

reddit

Flipboard

news360

NEWSTIFY
NEWS IS WHAT WE AIM FOR

Wikipedia Current Events Portal

July 7, 2016 (Thursday)

edit history watch

Disasters and accidents

- Super Typhoon Nepartak
  - The first major typhoon of 2016 threatens Taiwan, China and northern Luzon, Philippines. Thousands of people have been evacuated in Taiwan. (The Weather Channel), (ABC News)
  - Typhoon Nepartak is expected to make landfall on mainland China on Friday and will make flooding worse. Nearly 200 people have died in flood waters in China in the past week with 41 people missing, 1.6 million relocated and almost 50000 houses collapsed. (*The Telegraph*)

Law and crime

- A group of suspected radical Islamists hurl homemade bombs at police officers in the Kishoreganj District in central Bangladesh killing at least one officer and injuring several others. (AP via ABC News)

Politics and elections

- Australian federal election, 2016
  - Australian Prime Minister Malcolm Turnbull's Liberal/National coalition, behind Bill Shorten's Labor Party in the first 48 hours following Saturday's election, is now ahead of Labor in the Lower House, 74-71 seats, just two seats shy of the minimum needed to form a government. Minor parties and independents have won five seats; mail-in and absentee votes are still being counted. Turnbull is on the road today seeking support from a small handful of independent and small party lawmakers. (Reuters) (*The Australian*) (*Daily Mail*)

Time: 11:33 UTC | **Day:** 7 July | Purge

<<	July 2016					>>
S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

More July 2016 events...

About this page • Suggest a headline

News about Wikipedia

Ongoing events

Business

- United Kingdom withdrawal from the European Union
- 2016 Bangladesh Bank heist
- Panama Papers

Disasters

- 2016 California wildfires
- 2016 Fort McMurray wildfire

Health

- Flint water crisis
- Zika virus outbreak

Top-Story Identification

- **Top-story identification** (another task within the TREC Blog track) aims to identify the **most important news stories for a specific day d** based on their **coverage in the blogosphere**
  - **real-time** (online, limited statistics, time critical: small lag)
  - **retrospective**: (offline, full statistics)
- Notation:
  - d denotes the day of interest
  - B_d is the set of posts published at day d; p denotes a post
  - n denotes a news article (consisting of headline and content)
  - tf(v,p) is the term frequency of term v in post p

Top-Story Identification

- Lee and Lee [7] address retrospective top-story identification using **language models** estimated from news and blogs

- <u>Intuition</u>: *"News article important if discussed by many posts"*

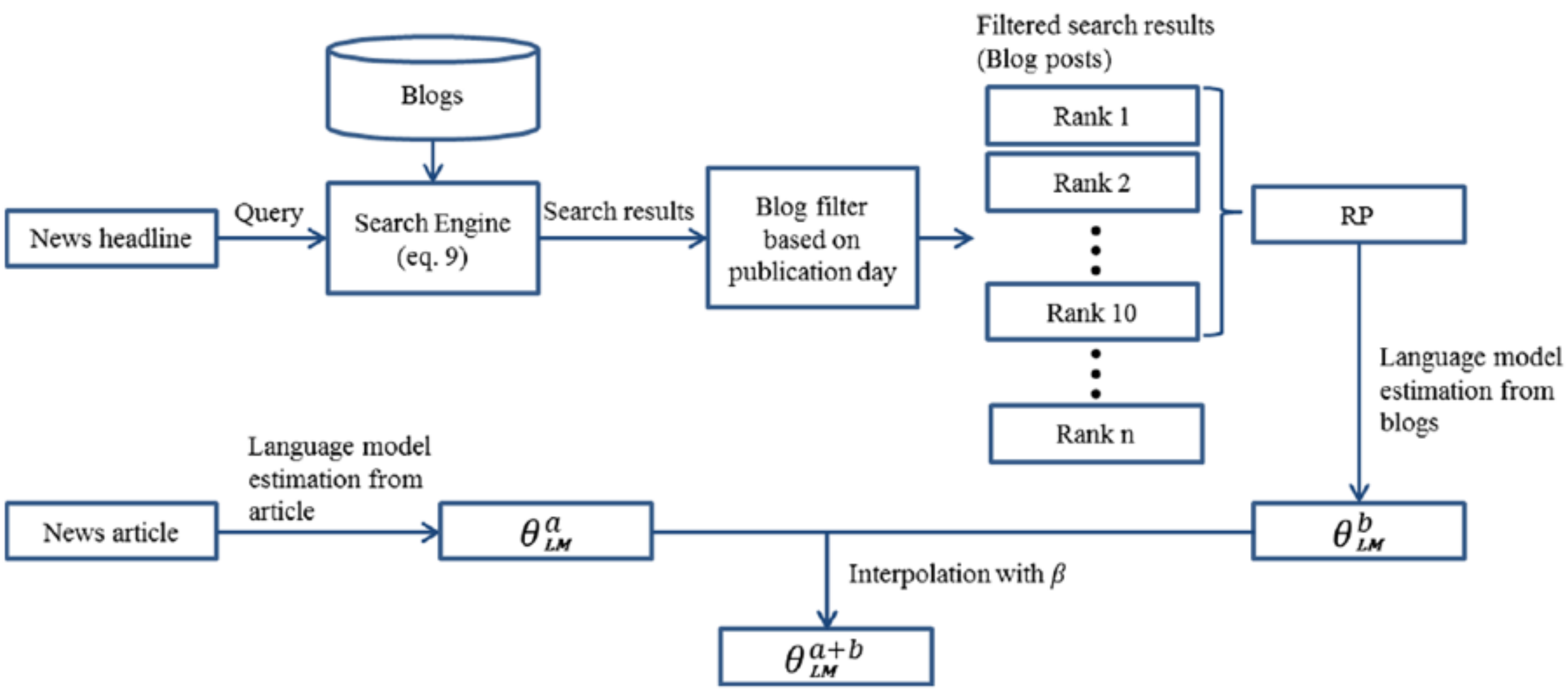
$$Importance(n,d) \propto KL(\underbrace{\theta_n}_{\text{LM representing news article } n} \parallel \underbrace{\theta_{B_d}}_{\text{LM representing posts published at day } d})$$

(Note: This is a simplified version of the approach described in [7])

- Only articles **published -1/+1 around the day of interest** d are considered as candidates and ranked by the approach

max planck institut informatik

Top-Story Identification Workflow



max planck institut
informatik

Blog Post Language Model

- Language model for **blog posts published at** d is estimated as

$$P[\, v \mid \theta_{B_d} \,] = \frac{tf(v, B_d) + \mu \cdot \frac{tf(v,B)}{\sum_w tf(w,B)}}{(\sum_w tf(w, B_d)) + \mu}$$

using Dirichlet smoothing with the collection of all posts B

max planck institut
informatik

MAX-PLANCK-GESELLSCHAFT

News-Story Language Model

- Option 1: Estimate **directly from content** of news article

$$P[\, v \mid \theta_n \,] = \frac{tf(v,n) + \mu \cdot \frac{tf(v,N)}{\sum_w tf(w,N)}}{\left(\sum_w tf(w,n)\right) + \mu}$$

VOCABULARY GAP?!?

using Dirichlet smoothing with the entire news collection N

- Option 2: Estimate from top-k **pseudo-relevant blog posts** B_n retrieved using **headline** as query and **published within -1/+1 month** of the news article; again using Dirichlet smoothing with the collection of all posts B

- Option 3: **Interpolate language models** estimated from news article content and top-k pseudo-relevant blog posts

max planck institut informatik

Summary

- **Meme tracking**
  grouping variants of memes to track them over time
- **Opinion retrieval**
  finds posts expressing an opinion about a specific named entity
- **Feed distillation**
  identifies feeds worth following for a given high-level topic
- **Top-story identification**
  spots most important news articles based on coverage in blogs
- **Vocabulary gaps**
  are a common obstacle in IR but can often be bridged
- **Language models**
  are versatile and can be used to address many (if not most) tasks

References

[1] **A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng:**
Time is of the Essence: Improving Recency Ranking Using Twitter Data,
WWW 2010

[2] **M. Efron**:
Information Search and Retrieval in Microblogs,
JASIST, 62(6):996–1008, 2011

[3] **J. Elsass, J. Arguello, J. Callan, J. G. Carbonell:**
Retrieval and Feedback Models for Blog Feed Search,
SIGIR 2008

[4] **B. He, C. Macdonald, J. He, Iadh Ounis**:
An Effective Statistical Approach for Blog Post Opinion Retrieval,
CIKM 2008

[5] **X. Huang and W. B. Croft:**
A Unified Relevance Model for Opinion Retrieval,
CIKM 2009

[6] **V. Lavrenko and W. B. Croft:**
Relevance-Based Language Models,
SIGIR 2001

References

[7] **Y. Lee and J.-H. Lee:**
Identifying top news stories based on their popularity in the blogosphere,
Information Retrieval 17:326–350, 2014

[8] **G. Mishne and M. de Rijke:**
A Study of Blog Search,
ECIR 2006

[9] **R. L. T. Santos, C. Macdonald, R. McCreadie, I. Ounis:**
Information Retrieval on the Blogosphere,
FTIR 6(1):1–125, 2012

[10] **J. Teevan, D. Ramage, M. R. Morris:**
#TwitterSearch: A Comparison of Microblog Search and Web Search,
WSDM 2011

[11] **W. Weerkamp, K. Balog, M. de Rijke:**
Blog feed search with a post index,
Information Retrieval 14:515–545, 2011

max planck institut
informatik