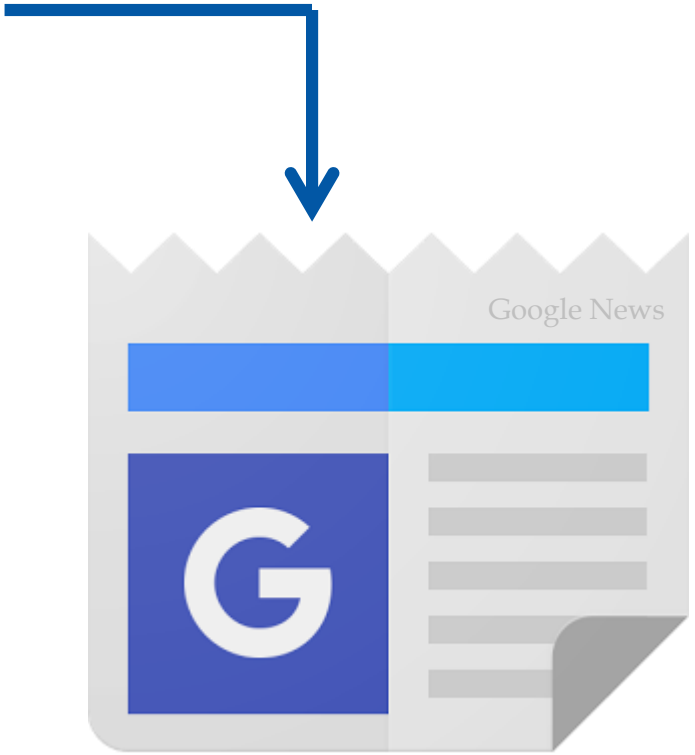
Modeling Event Importance for Ranking Daily News Events

Speaker: Shih-Han Lo
Advisor: Professor Jia-Ling Koh
Author: Vinay Setty, Abhijit Anand, Arunav Mishra, Avishek Anand
Date: 2017/03/21
Source: WSDM '17

Outline

- Introduction
- Method
- Experiment
- Conclusion

Introduction



Introduction

Motivation

- The observation that both automated aggregation and manual curation of news events need to solve two fundamental tasks:
  - Mining news events
  - Modeling news importance

Introduction

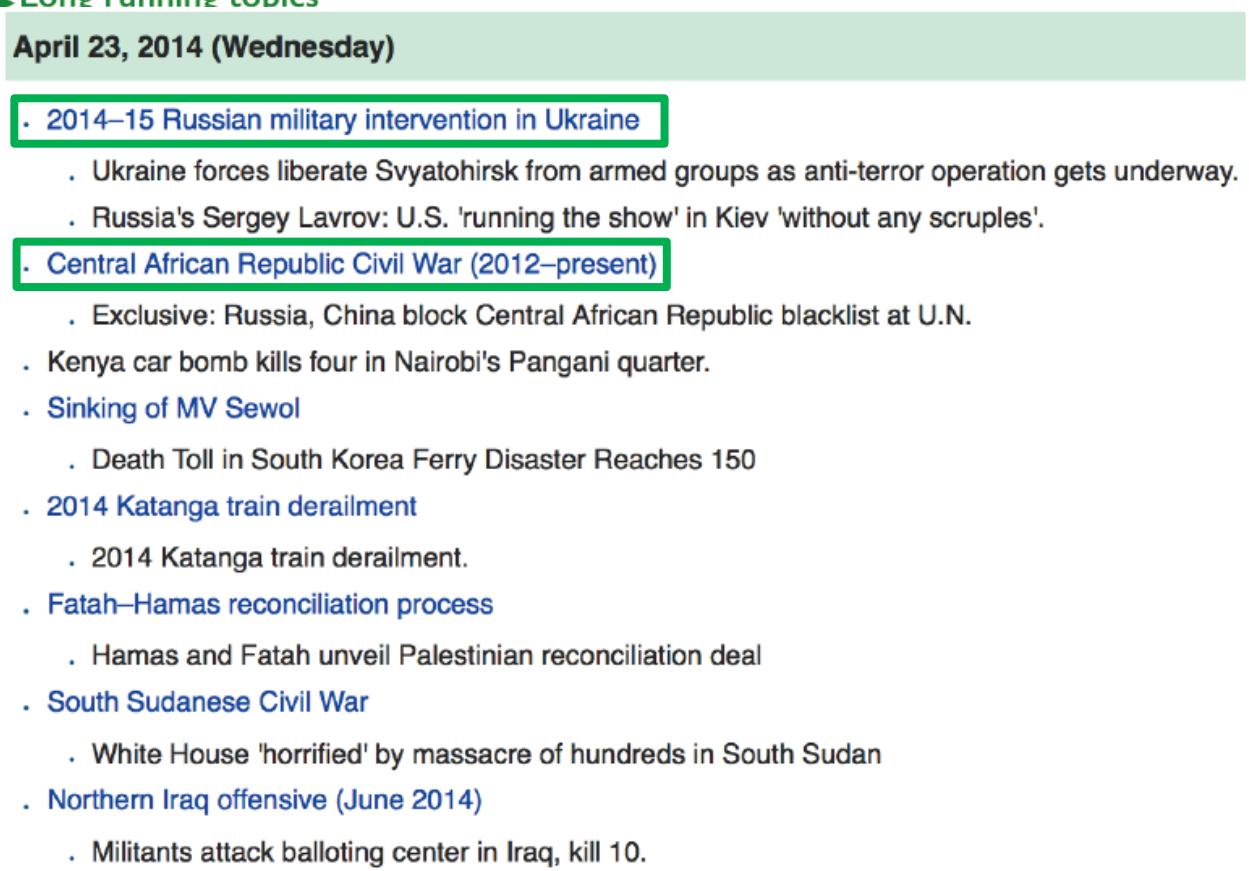
Goal

- Model the importance of wide variety of news events reported by large number of news articles.

Introduction

https://en.wikipedia.org/wiki/Portal:Current_events/April_2014

(a) Wikipedia Current Event



(b) Top news events and topics from our approach

Outline

- Introduction
- Method
- Experiment
- Conclusion

Method

Problem Definition

- News story
  - $d \in \mathcal{D}$ is a news article document.
- News event
  - c, a cluster of stories associated with a news event.
- News topic, σ.

- We approach the news ranking problem as a *Learning-to-Rank* task, specifically *SVMRank*.

Method

Mining Daily News Events

- First, we need to mine events from the news collection.
  - A bag of entities $\mathcal{E}(d)$
  - A bag of shingles $\mathcal{S}(d)$ (w-shingling, n-grams)
- We combine entities and shingles into a single bag $\mathcal{F}(d) = \mathcal{E}(d) \cup \mathcal{S}(d)$. Then:

Frequency of unique entities

$$\delta(d, d') = 1 - \frac{\sum_{e \in \mathcal{F}(d) \cap \mathcal{F}(d')} weight(e)}{\sum_{e \in \mathcal{F}(d) \cup \mathcal{F}(d')} weight(e)} \qquad (1)$$

Method

- Problem: Inability to accurately determine the true number of events
  - We resort to Locally Sensitive Hashing (LSH) with min-wise independent permutations.

- Cluster cohesiveness:

$$\phi(c) = \frac{|c|}{\sum_{d \in c} \sum_{d' \in c, d \neq d'} \delta(d, d')} \qquad (2)$$

Method

Improved Popularity Estimation

- Improving Cluster Size Estimate

$$\Gamma_c = \arg\min_{d \in c} \{\forall d' \in \{c \setminus d\} : \max \delta(d, d')\}, \quad \rho_i = \max_{d \in c} \delta(d, \Gamma_c)$$

Cluster centroid | Radius

- Maximum Sub-Cluster Density
  - k, with ρ_k as the radius containing k nearest neighbors of the centroid.
  - Find a sub-cluster which maximizes k/ρ_k (= ψ_{max}).
  - Effective size: $S_{eff} = \text{argmax}\, \psi_{max}$.

Method

- Source Diversity
  - Collection bias: Relying only on structural features may be misleading.
  - Compute a diversity score for each cluster: $\frac{cc}{log(1+S_{eff})}$
- Source Authority
  - We extract all possible news citations and construct a probability distribution based on their frequencies.

Method

Historical Importance

- Cluster Chaining
  - Previous day similarity:

$$psim(c) = \text{argmax}_{c^{t-1}}\{\delta(\Gamma_c, \Gamma_{c^{t-1}})\}.$$

  - The overall historical value for a chain initiated from *c* is:

$$\mathcal{H}_c(c) = \sum_{i \leq t} (\beta . S_{eff}(c^i) + (1 - \beta) . e^{\alpha . (i-t)} psim(c^i)) \quad (3)$$

Method

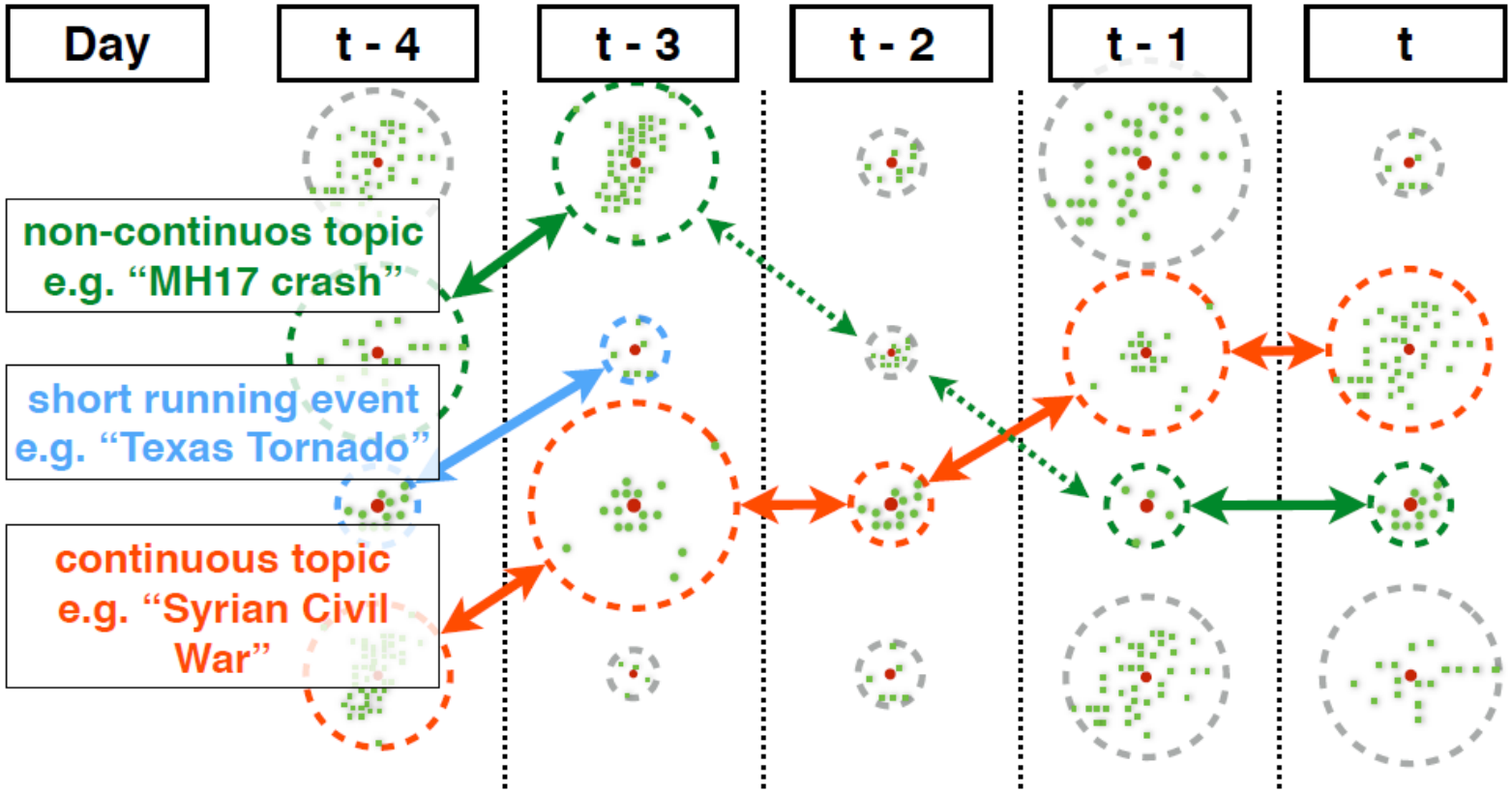


Figure 2: Example for cluster chaining

Method

- Temporal Profile from Named Events
  - *Moving Window Language Model*:

$$KL(LM_d||LM_{\sigma_s}) = \sum_w P(w|LM_d) log \frac{P(w|LM_d)}{P(w|LM_{\sigma_s})} \quad (4)$$

  - *Moving Window Entity Overlap* using the disambiguated entities:

$$J(d, \sigma_s) = \frac{|\mathcal{E}(d) \cap \mathcal{E}(\sigma_s)|}{|\mathcal{E}(d) \cup \mathcal{E}(\sigma_s)|} \quad (5)$$

Method

- Temporal Prior:

Frequency of edits

$$T(\sigma, t) = \sum_{t=t_1}^{t_2} \frac{|Edit(\sigma, t)|}{\sum_{t'=t_b(\sigma)}^{t_e(\sigma)} |Edit(\sigma, t')|} \qquad (6)$$

- Finally, we compute historical significance on a day *t*:

$$\mathcal{H}_w(c) = (\beta . S_{eff}(c)) + (1 - \beta) . \sum_{x=1}^{W} (e^{\alpha.(-x)} T(\sigma, t - x)) \quad (7)$$

Outline

- Introduction
- Method
- Experiment
- Conclusion

Experiment

Datasets

- Gdelt
  - 8 million stories.
  - Sep. 2013 – Aug. 2014 (365 days).
  - 6000 sources from 167 different countries.
- Stics
  - 1.69 million stories.
  - Jan. 2014 – Jun. 2015 (545 days).
  - 300 sources from 10 different countries.

Experiment

Benchmark

- GTS
  - We add the news stories referred in the WCEP summaries into the input collection.

- Time Lag
  - Within the 3 days window of the WCEP dates.

Experiment

Ranking Results

Dataset and Approach	Online				
	P@1	**P@5**	**P@10**	**MAP@5**	**MAP@10**
Stics					
Cluster-Size	0.711	0.555	0.508	0.625	0.576
Story-Rank-LM	0.376	0.409	0.421	0.396	0.404
Story-Rank-LM-Hist	0.462	0.491	0.507	0.492	0.499
Event-Rank-Pop	0.650	0.615	0.582	0.627	0.608
Event-Rank-Pop-Hist	**0.929**	**0.882**	**0.882**	**0.889**	**0.885**
Gdelt					
Cluster-Size	0.773	0.571	0.487	0.632	0.575
Story-Rank-LM	0.570	0.451	0.393	0.517	0.465
Story-Rank-LM-Hist	0.612	0.469	0.405	0.530	0.480
Event-Rank-Pop	0.797	0.695	0.578	0.750	0.690
Event-Rank-Pop-Hist•	**0.923**	**0.740**	**0.675**	**0.812**	**0.750**

Outline

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

- We introduced the problem of ranking a daily batch of events for large heterogeneous news corpora.

- With the use of improved popularity and historical features for events in a learning to rank framework we came up with an effective daily event ranking.