

# 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*

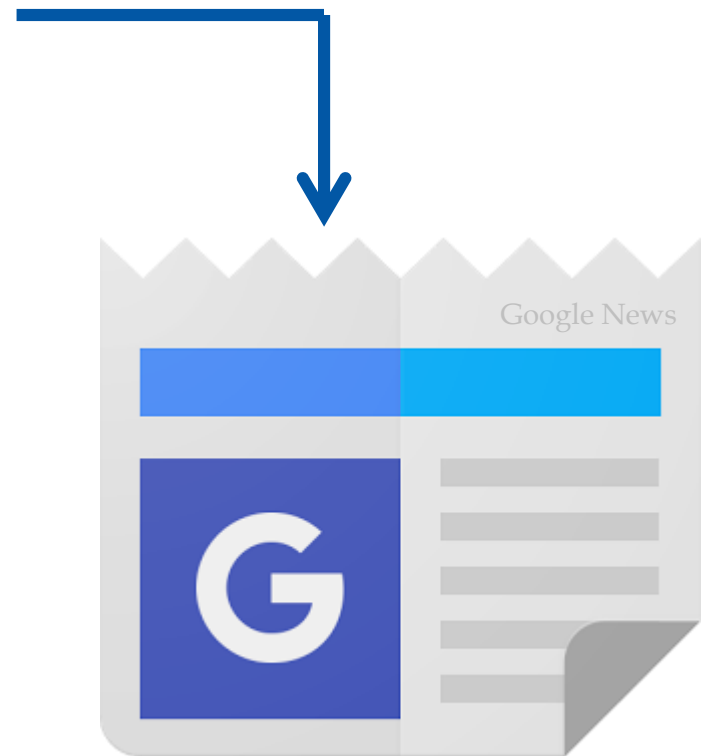
Date: 2017/03/21

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](https://en.wikipedia.org/wiki/Portal:Current_events/April_2014)

April 23, 2014 (Wednesday)

[edit](#) [history](#) [watch](#)

## Armed conflicts and attacks

### 2014 pro-Russian unrest in Ukraine:

- Ukrainian security officials discover that the Assumption Monastery in Sviatohirsk served as a military base for pro-Russian insurgents.

Ukraine's Deputy Prime Minister Vitaly Yarema announces that the eastern cities, Kramatorsk, Slaviansk, Donetsk and Luhansk have been previously seized, defying the international deal made in Geneva.

### Post-coup unrest in Egypt:

- An Egyptian general is killed and a police officer is injured during the protests. (CNN)

### Central African Republic conflict:

- Russia and China have blocked a proposal by the United States to remove the African Republic's former President Francois Bozizé and two other officials from the blacklist.

A car bomb detonates outside a police station in Nairobi, Kenya.

## Disasters and accidents

### Sinking of the MV Sewol:

- The official death toll of the ferry capsizing rises to 150 with 12 still missing.

### 2014 Katanga train derailment:

- More than 60 people are killed and 80 are seriously injured in Congo's Katanga Province. (Reuters)

## Politics and elections

- Palestinian rival factions Fatah and Hamas have announced reconciliation talks after a violent clash in 2007. (BBC)

## Long running topics

April 23, 2014 (Wednesday)

### 2014–15 Russian military intervention in Ukraine

- Ukraine forces liberate Svyatohirsk from armed groups as anti-terror operation gets underway.
- Russia's Sergey Lavrov: U.S. 'running the show' in Kiev 'without any scruples'.

### Central African Republic Civil War (2012–present)

- Exclusive: Russia, China block Central African Republic blacklist at U.N.

### Kenya car bomb kills four in Nairobi's Pangani quarter.

### Sinking of MV Sewol

- Death Toll in South Korea Ferry Disaster Reaches 150

### 2014 Katanga train derailment

- 2014 Katanga train derailment.

### Fatah–Hamas reconciliation process

- Hamas and Fatah unveil Palestinian reconciliation deal

### South Sudanese Civil War

- White House 'horrified' by massacre of hundreds in South Sudan

### Northern Iraq offensive (June 2014)

- Militants attack balloting center in Iraq, kill 10.

One-off isolated event

(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**,  $\sigma$ .
  
- 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:

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

Frequency of unique entities

# 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:

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

# Method

## Improved Popularity Estimation

- Improving Cluster Size Estimate

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

Cluster centroid

$$\rho_i = \max_{d \in c} \delta(d, \Gamma_c)$$

Radius

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

# 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:

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

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

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

# Method

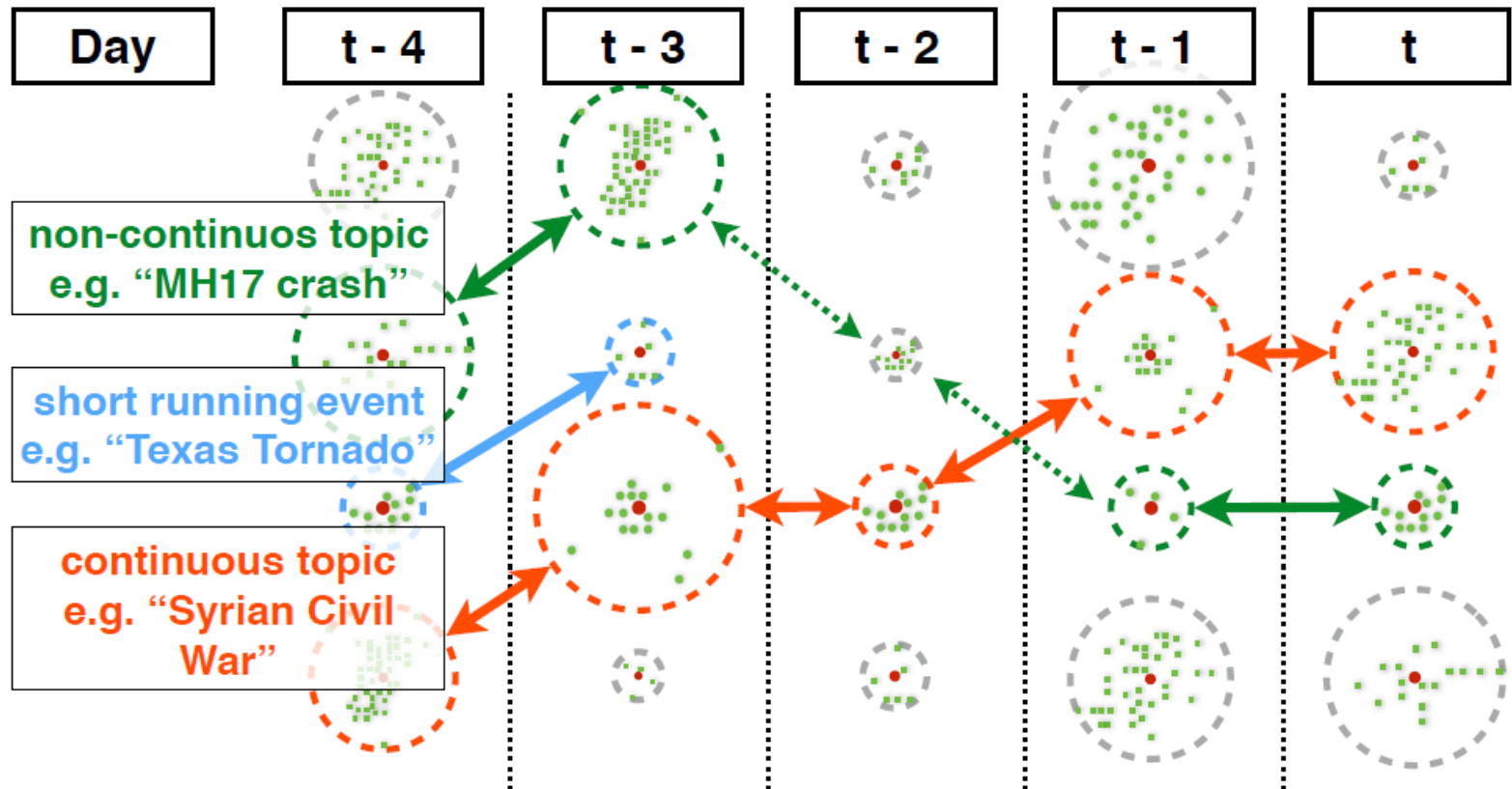


Figure 2: Example for cluster chaining

# 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:

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

Frequency of edits

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

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



# 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
<b>Stics</b>					
<i>Cluster-Size</i>	0.711	0.555	0.508	0.625	0.576
<i>Story-Rank-LM</i>	0.376	0.409	0.421	0.396	0.404
<i>Story-Rank-LM-Hist</i>	0.462	0.491	0.507	0.492	0.499
<i>Event-Rank-Pop</i>	0.650	0.615	0.582	0.627	0.608
<i>Event-Rank-Pop-Hist</i>	<b>0.929</b>	<b>0.882</b>	<b>0.882</b>	<b>0.889</b>	<b>0.885</b>
<b>Gdelt</b>					
<i>Cluster-Size</i>	0.773	0.571	0.487	0.632	0.575
<i>Story-Rank-LM</i>	0.570	0.451	0.393	0.517	0.465
<i>Story-Rank-LM-Hist</i>	0.612	0.469	0.405	0.530	0.480
<i>Event-Rank-Pop</i>	0.797	0.695	0.578	0.750	0.690
<i>Event-Rank-Pop-Hist</i> <sup>•</sup>	<b>0.923</b>	<b>0.740</b>	<b>0.675</b>	<b>0.812</b>	<b>0.750</b>

# 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.