Event Phase Extraction and Summarization

Chengyu Wang[1], Rong Zhang[1], Xiaofeng He[1], Guomin Zhou[2], Aoying Zhou[1]

[1)] Institute for Data Science and Engineering,
East China Normal University
[2)] Zhejiang Police College

Outline

- **Introduction**
- Problem Statement
- Proposed Approach
- Experiments
- Conclusion

Event Phase Extraction and Summarization (1)

- Event phase
  - Model an single event as multiple event phases
  - Each event phase relates to a single development period of a long, complicated event.
- Example: Egypt Revolution

1. Protests against Hosni Mubarak

2. Egypt under the Supreme Council

3. Egypt under President Morsi

4. Protests against President Morsi

Egypt Revolution

https://en.wikipedia.org/wiki/Egyptian_revolution_of_2011

Event Phase Extraction and Summarization (2)

- Event phase extraction and summarization
  - Input: a collection of news articles w.r.t. the same event
  - Event phase extraction: cluster news articles into different event phases
  - Event phase summarization: select top-k news headlines as the event phase summary for each event phase
- Techniques
  - Graph-based representation of news articles: Temporal Content Coherence Graph (TCCG)
  - A structural clustering algorithm to partition news articles into event phases: EPCluster
  - News headline ranking and selection: vertex-reinforced random walk process

Outline

- Introduction
- **Problem Statement**
- Proposed Approach
- Experiments
- Conclusion

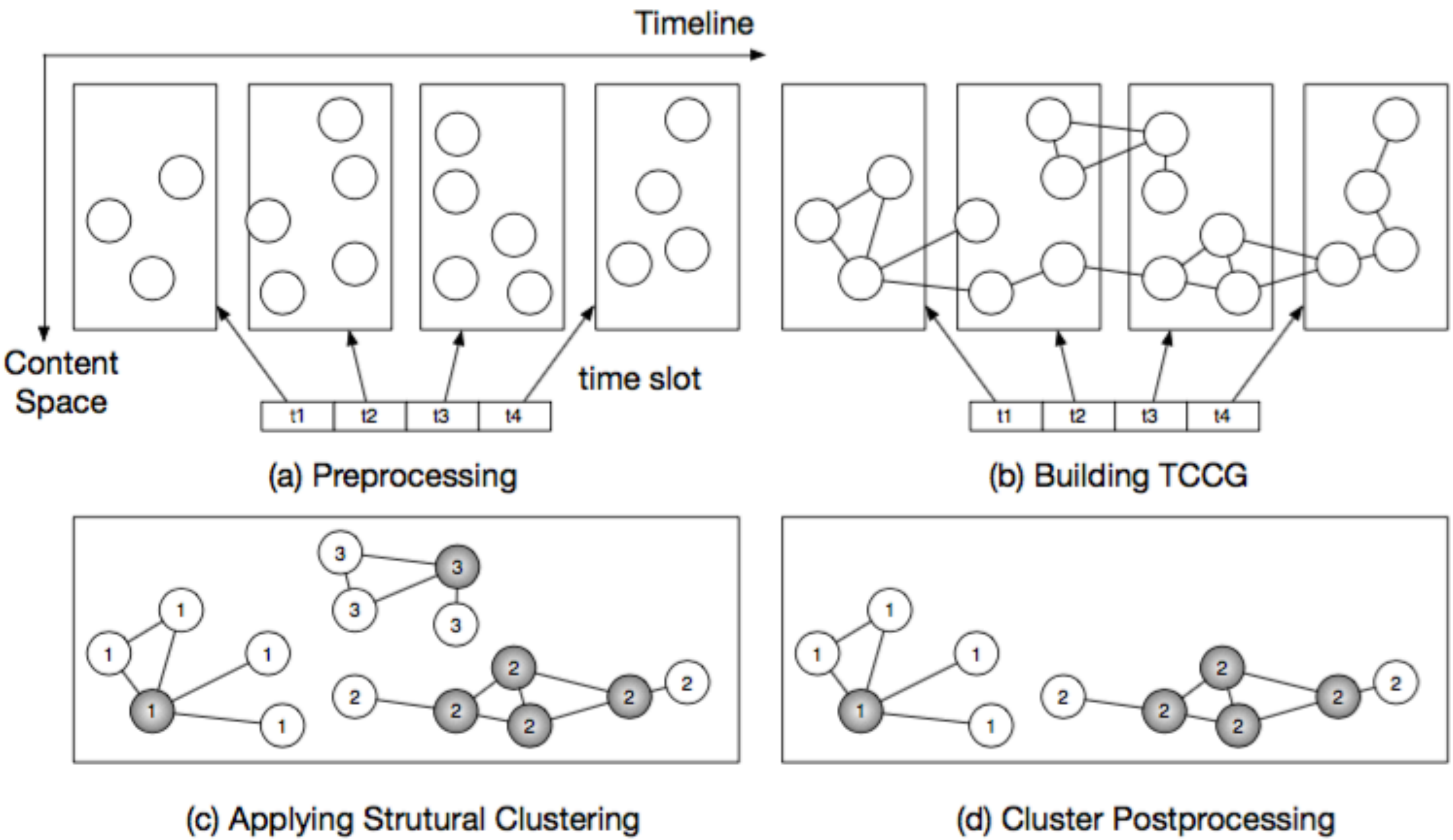
Problem Statement

- News article $d_i = (h_i, t_i, s_i)$
  - h_i: news headline
  - t_i: publication time
  - s_i: the sentence collection of news contents
- News collection $D = \{d_i\}$
- Event phase summary $P = \{(h_i, t_i)\}_{i=1}^{k}$
  - A collection of k news headline and publication time pairs
- Event phase extraction and summarization
  - Input: a news collection D
  - Output: a collection of N event phase summaries $\boldsymbol{P} = \{P_j\}_{j=1}^{N}$
  - The number N is not pre-defined.

Outline

- Introduction
- Problem Statement
- **Proposed Approach**
- Experiments
- Conclusion

Framework of Event Phase Extraction



Semantic Relatedness (1)

- Content coherence
  - Topic level similarity: Jansen-Shannon divergence between topic distributions

$$D_{JS}\left(\theta_i \,\middle\|\, \theta_j\right) = \frac{D_{KL}\left(\theta_i \,\middle\|\, \bar{\theta}\right) + D_{KL}\left(\theta_j \,\middle\|\, \bar{\theta}\right)}{2}$$

  - Entity level similarity: Tanimoto coefficient
    - C_i: count vector of key entities in d_i

$$TC(C_i, C_j) = \frac{C_i^T \cdot C_j}{\|C_i\|^2 + \left\|C_j\right\|^2 - C_i^T \cdot C_j}$$

  - Content coherence score

$$w_c(d_i, d_j) = \alpha\left(1 - D_{JS}\left(\theta_i \,\middle\|\, \theta_j\right)\right) + (1-\alpha)TC(C_i, C_j)$$

Semantic Relatedness (2)

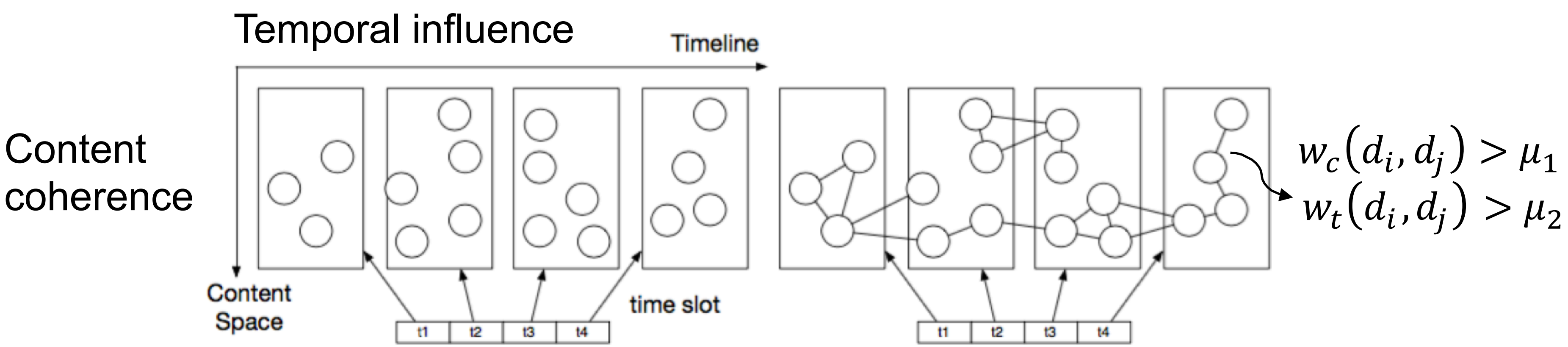
- Temporal influence
  - Use Hamming kernel to map the publication time gap to a real number in [0,1]

$$\Delta t_{i,j} = \left| t_i - t_j \right|$$

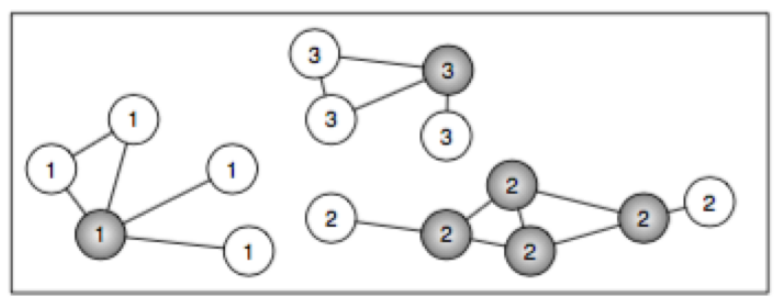
$$w_t\left(d_i, d_j\right) = \begin{cases} \frac{1}{2}\left(1 + \cos\frac{\Delta t_{i,j} \cdot \pi}{\sigma}\right), & x < 0 \\ 0, & x \geq 0 \end{cases}$$

Structural Clustering

- Temporal Content Coherence Graph (TCCG)



- EPCluster: Structural clustering algorithm
  - Parameter: $MinPts$
  - Core Object
  - Border Object
  - Noise Object



$MinPts = 3$

Cluster Postprocessing

- Goal
  - Use a classifier to filter out "small" clusters that do not correspond to an actual event phase
- Features
  - Article quantity $N(C_i) = \frac{|C_i|}{|D|} \times 100\%$
  - Time interval $T(C_i) = t_{max}^i - t_0^i$
  - Pairwise topic similarity $ATS(C_i) = 1 - \frac{2\sum_{d_m,d_n \in C_i} D_{JS}(\theta_m \| \theta_n)}{|C_i| \cdot (|C_i| - 1)}$
  - Pairwise entity similarity $AES(C_i) = \frac{2\sum_{d_m,d_n \in C_i} TC(C_m, C_n)}{|C_i| \cdot (|C_i| - 1)}$
- Prediction function $f(C_i) = \frac{1}{1+e^{-w \cdot F(C_i)}}$

News Article Ranking

- Goal
  - Assign each news article in an event phase an "informative-ness" rank value
- Vertex-reinforced random walk process
  - Graph construction: build a complete graph where the node set is news articles in an event phase
  - Prior transition probability $M^{(m,n)} = \frac{1}{Z} \cdot w_c(d_m, d_n) \cdot w_t(d_m, d_n)$
  - Rank propagation process
    - Transition matrix update:

$$T_n = [R_n R_n \cdots R_n]$$
$$M_{n+1} = \lambda T_n M_n + (1-\lambda) M_0$$

    - Rank update:

$$R_{n+1} = \lambda M_{n+1} R_n + (1-\lambda) R_0$$

华东师范大学数据科学与工程研究院
Institute for Data Science and Engineering at ECNU

Event Phase Summary Generation

- New article selection problem
  - Select k news articles from C_i (denoted as S_i) to generate the event phase summary
  - Optimization problem
    - Objective function: $\max_{S_i \subset C_i} R(S_i) = \sum_{d_j \in S_i} r(d_j)$
    - Subject to: $|S_i| = k, \forall d_m, d_n \in S_i, w_c(d_m, d_n) \le \mu_1, w_t(d_m, d_n) \le \mu_2$
  - Algorithm
    - A greedy algorithm with approximation ratio $1 - \frac{1}{e}$

Algorithm 3 News Article Selection Algorithm

Input: News cluster C_i, parameter k.
Output: Selected news collection S_i.

$S_i = \emptyset$;
while $C_i \neq \emptyset$ and $|S_i| < k$ **do**
  Select $d_n = \text{argmax}_{d_n \in C_i} R(S_i \cup \{d_n\}) - R(S_i)$ subject to $w_c(d_m, d_n) \le \mu_1, w_t(d_m, d_n) \le \mu_2, \forall d_m \in S_i$;
  $S_i = S_i \cup \{d_n\}$;
  $C_i = C_i \setminus \{d_n\}$;
end while
return S_i;

Outline

- Introduction
- Problem Statement
- Proposed Approach
- **Experiments**
- Conclusion

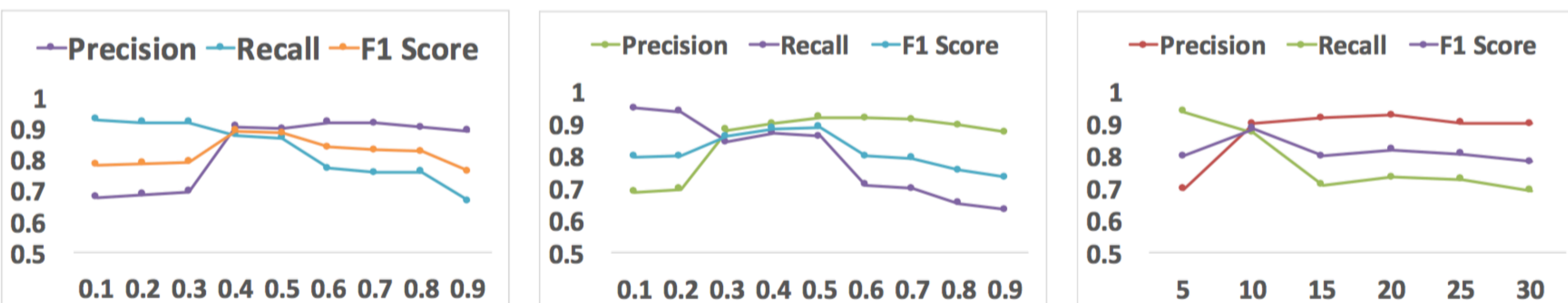
Experiments (1)

- Datasets
  - Four English news datasets regarding long-span recent armed conflicts
  - News source: 24 news agencies, e.g., Associated Press, Reuters, Guardian, etc.

Dataset	Event	#Article	Time Range
D_1	Egypt Revolution	3,869	2011.1.11 - 2013.7.24
D_2	Libya War	3,994	2011.2.16 - 2013.7.18
D_3	Syria War	4,071	2011.11.17 - 2013.7.26
D_4	Yemen Crisis	3,600	2011.1.15 - 2013.7.25

Experiments (2)

- Parameter Tuning
  - Pairwise judgment
    - Testing set: news article pairs $T_i = \{(d_m, d_n)\}$
    - Manually label whether each pair is related to the same event phase
  - Evaluation metrics: Precision, Recall and F-measure
  - Experimental results
    - $\mu_1 = 0.4, \mu_2 = 0.5, MinPts = 10$



(a) Varying μ_1 (b) Varying μ_2 (c) Varying $MinPts$

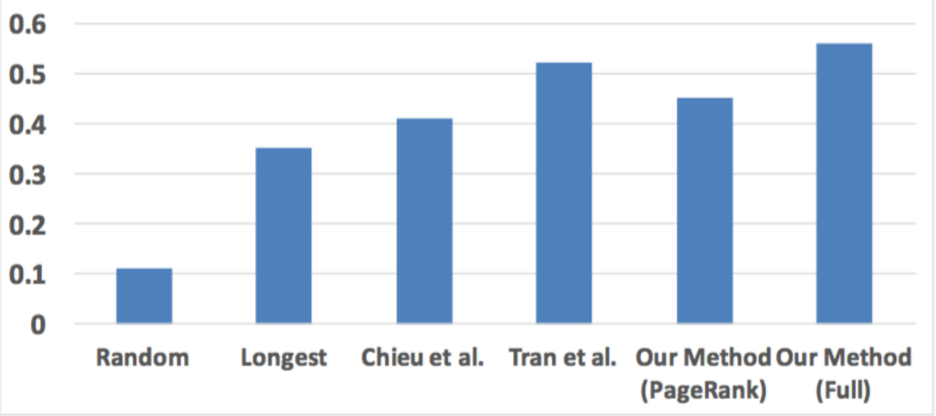
Experiments (3)

- Baselines
  - VSMCluster: KMeans using word features of TF-IDF weights
  - TopicCluster: KMeans using topic distributions based on LDA
  - SCAN: structural clustering algorithm for network partitioning
  - EPCluster-C: EPCluster without postprocessing
- Results
  - Our method EPCluster is effective for event phase extraction.

Method	**VSMCluster**	**TopicCluster**	**SCAN**	**EPCluster-C**	**Our Method**
Precision	0.35	0.52	0.78	0.81	**0.89**
Recall	0.74	0.67	0.72	**0.79**	0.78
F1 Score	0.48	0.59	0.75	0.80	**0.83**

Experiments (4)

- Baselines
  - Random: selects news articles randomly
  - Longest: selects news articles with longest headlines
  - Tran et al., Chieu et al.: timeline generation methods
  - Our Method (PageRank): the variant of our method
- Evaluation
  - Evaluate the relevance of news headlines based on gold-standard event summaries
  - Experimental results



Case Study

Event Phase #1 *Protest against Hosni Mubarak*	
2011.2.2	Egypt protests: Hosni Mubarak to stand down at next election
2011.2.11	Hosni Mubarak resigns and Egypt celebrates a new dawn
Event Phase #2 *Egypt under the Rule of Military Power*	
2011.4.9	Egyptian soldiers attack Tahrir Square protesters
2011.7.10	Protests spread in Egypt as discontent with military rule grows
Event Phase #3 *Mohammed Morsi Won Presidential Election*	
2012.5.23	First round of presidential election
2012.6.24	Election officials declare Morsi the winner
Event Phase #4 *Protest against Morsi and Muslim Brotherhood*	
2013.1.27	Egypt's Mohammed Morsi declares state of emergency, imposes curfew
2013.1.30	Egypts military chief says clashes threaten the state
Event Phase #5 *Morsi's Ousting*	
2013.7.4	After Morsi's Ousting, Egypt Swears in New Presiden
2013.7.6	Morsi's ouster in Egypt sends chill through political Islam

Outline

- Introduction
- Problem Statement
- Proposed Approach
- Experiments
- **Conclusion**

Conclusion

- Event Phase Extraction and Summarization
  - A structural clustering algorithm for event phase extraction based on TCCG
  - Summary generation via news article ranking and rank optimization
- Future work
  - Improving the performance of document summarization and timeline generation when event phases are considered

Thanks!

Questions & Answers