TimeMachine:
Timeline Generation for Knowledge-Base Entities

Date : 2015/10/1

Author: Tim Althoff, Xin Luna Dong, Kevin Murphy, Safa Alai, Van Dang, Wei Zhang

Source: ACM KDD’15

Advisor: Jia-ling Koh

Speaker: Han, Wang

Outline

- **Introduction**
- Approach
- Experiment
- Conclusion

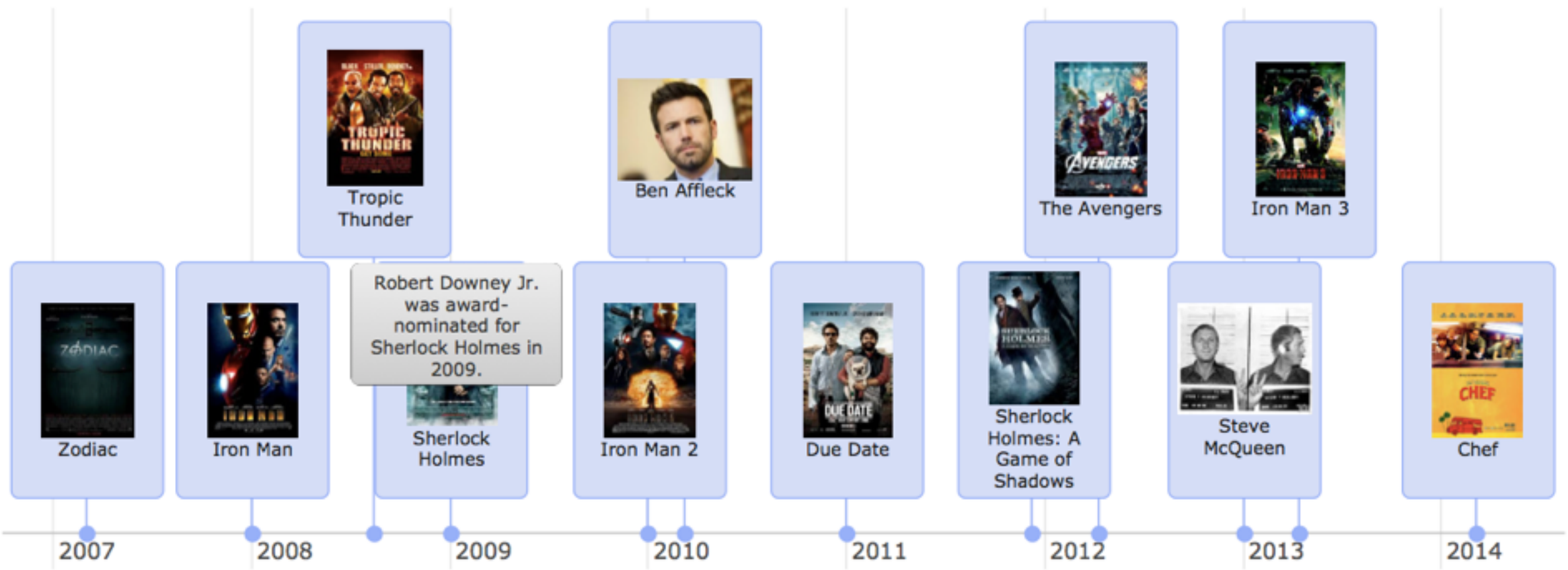
Introduction

- Motivation:
  -Overview of the most important events in an organized and readable format would serve users better.

- Pervious works:
  NO Selecting sub-events
  NO encourage diversity
  NO user interaction

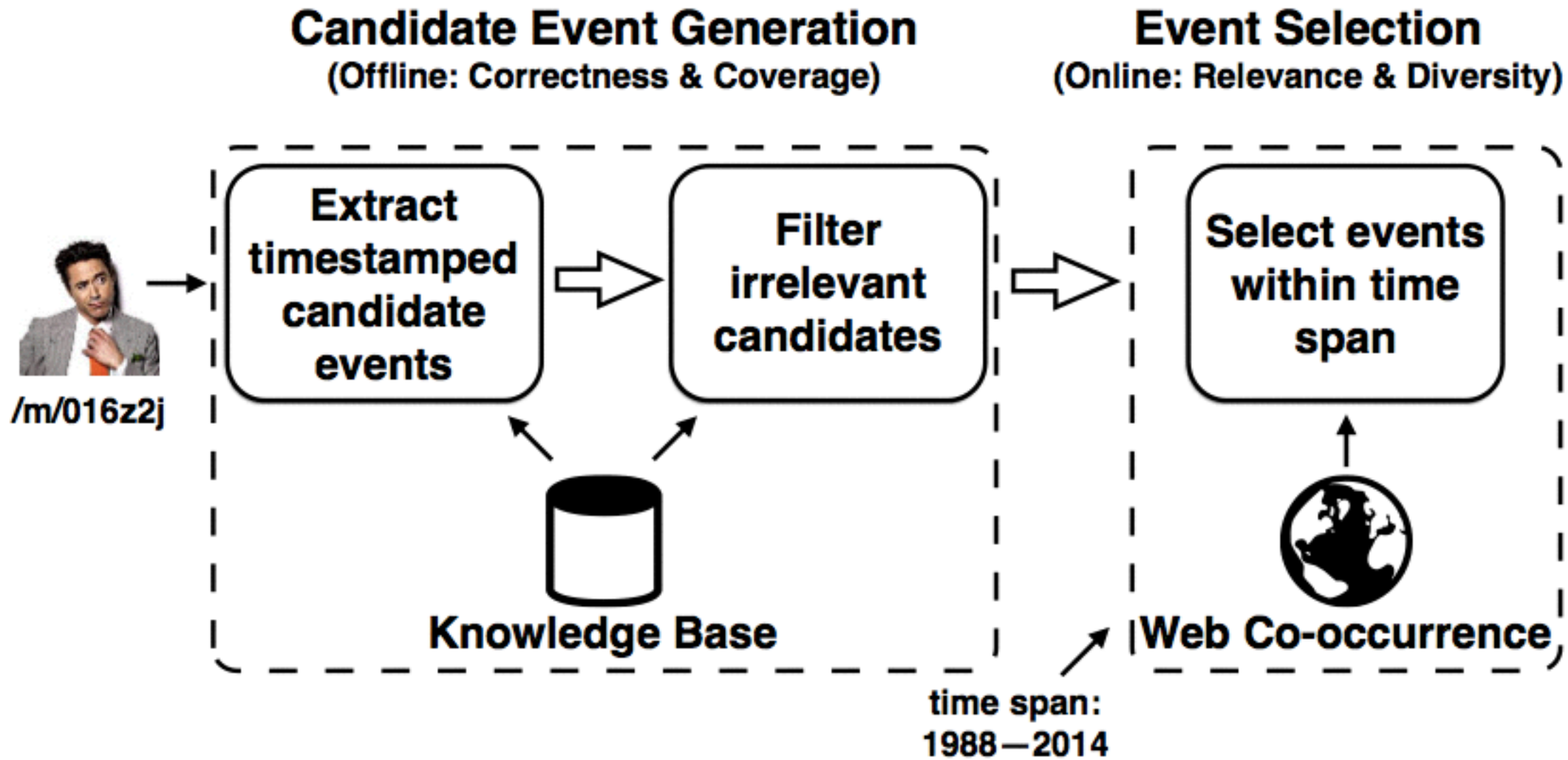
Introduction

- Goal: http://cs.stanford.edu/~althoff/timemachine/
- ☑ Relevance Display
- ☑ Temporal Diversity
- ☑ Content Diversity



Introduction

- Framework:



Outline

- Introduction
- **Approach**
- Experiment
- Conclusion

Approach

- **Event Generation:**
  Freebase knowledge base.

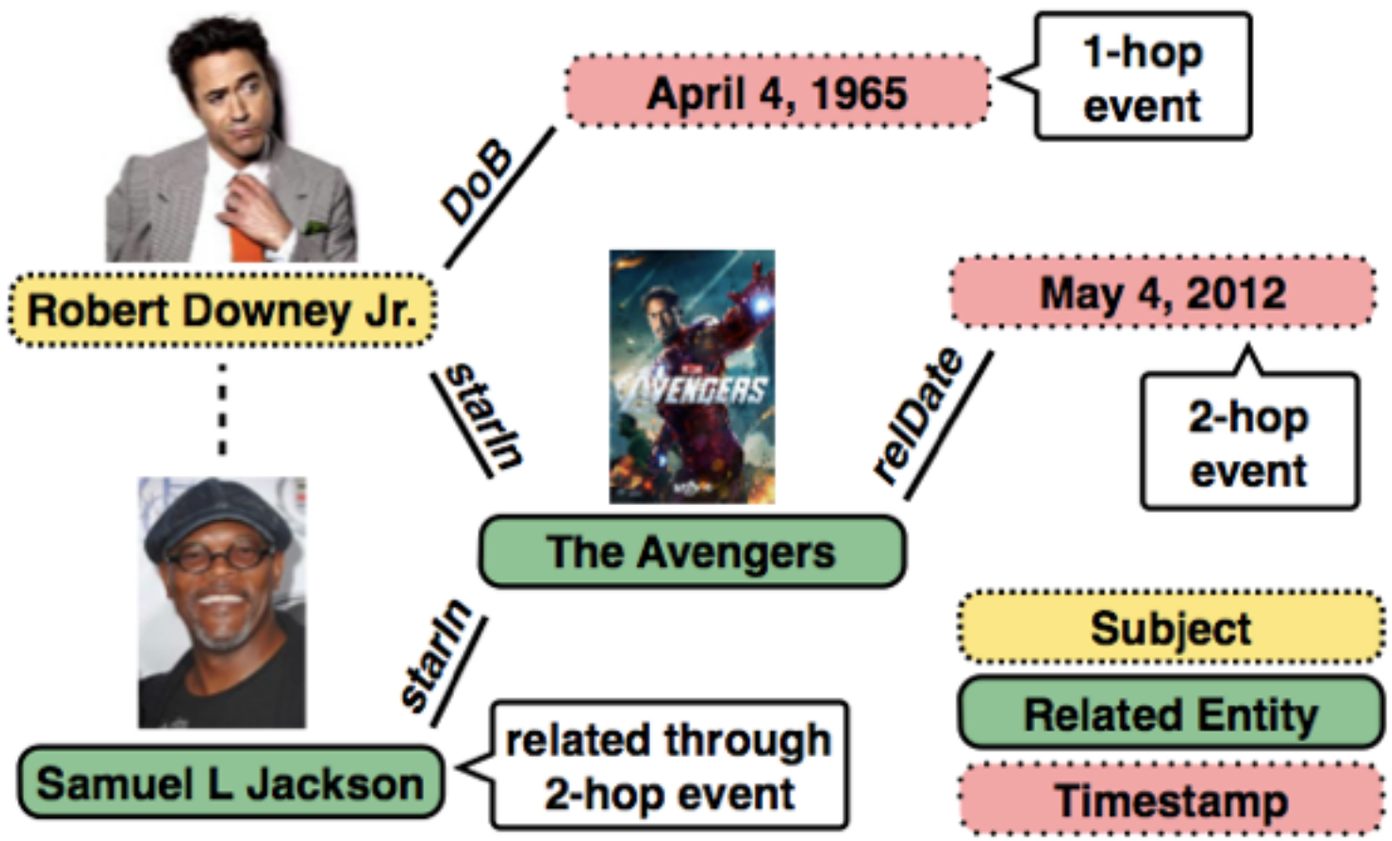
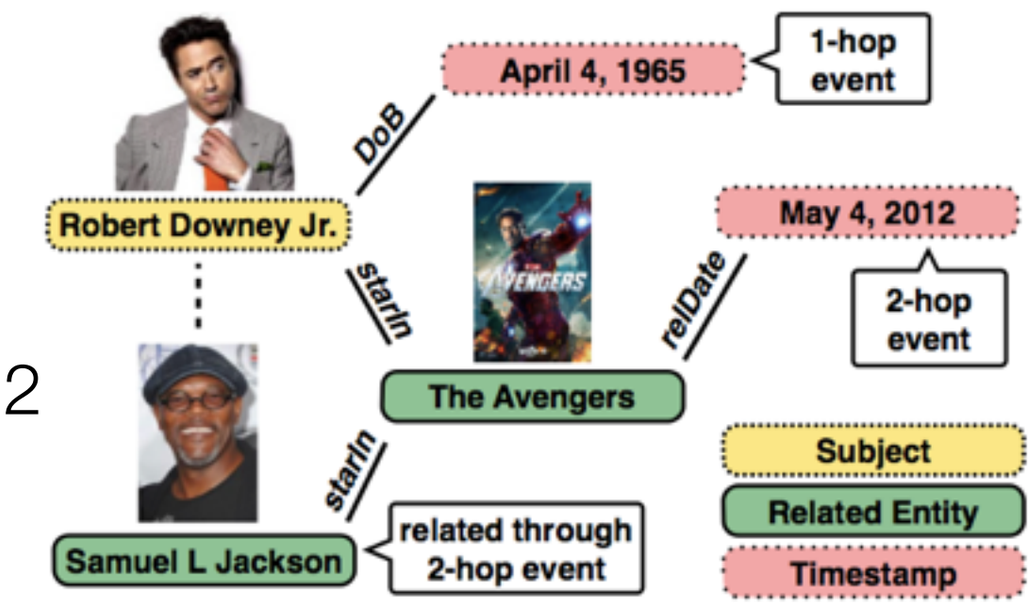


Figure 3: An illustration of the event candidate generation step. Events are short paths that are associated with a timestamp.

Approach

- **Event Generation:**



* **Simple events:**
  Length = 1 or 2, s-> re -> t
  Ex: Robert -> The Avengers -> 5/4'12

* **Compound events:**
  Two or more simple events share the same related entity and timestamp, but differ in subject.

* **Event Description:**
  Nature Language form, Define 100 most frequently occurring paths.

Approach

- **Event Filtering:**

  Remove irrelevant events.

- **Frequency Filter:**

  The idea is that an event that is commonly associated with a large number of subjects is unlikely to be particularly interesting. (IDF).

$$C(\pi_t) = |\{(re, t) : N(\pi_t, re, t) > \theta_1\}$$

$$C(\pi_t)/N(\pi_t) > \theta_2$$

- **Existence Filters:**

  Filters out events that occurred before an entity began to exist.

Approach

- **Event Filtering:**

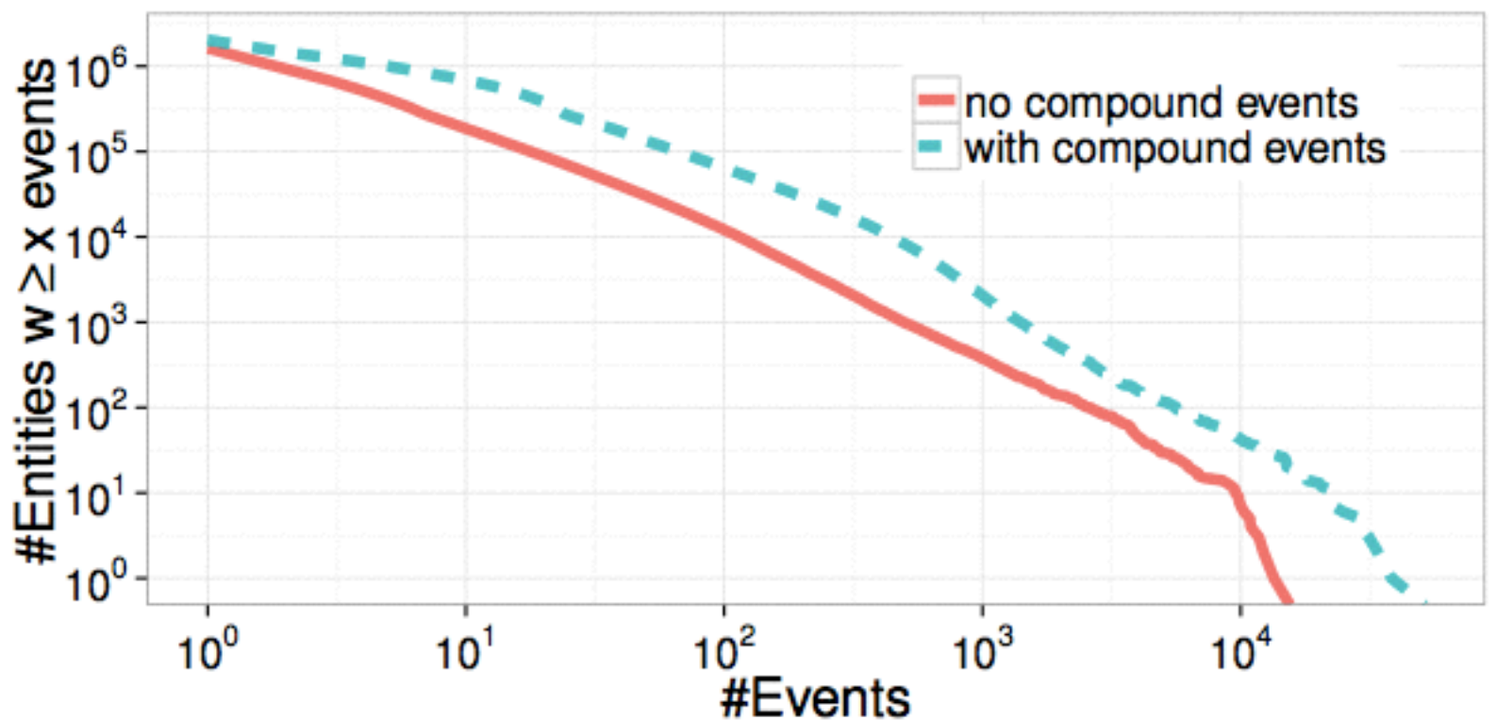


Figure 5: Log-log coverage plot showing the number of entities with X or more candidate events and illustrating the impact of adding compound events (dashed line).

- ☑ history rich -> at least 100 candidate events

Approach

- **Event Selection:**

* **Timeline Optimization Problem**
  REL evaluates how relevant each event is to the subject and how diverse they are, and temporal constraint CONSTRAINTS requires that all events can fit into the provided space without overlapping or occluding each other.

$$T^* = \arg\max_{T \subseteq E} \textsc{Rel}(s, T)$$
$$\text{s.t. } \textsc{Constraints}(T, E, \mathcal{W}, w, n)$$

* **Relevance Function:**

$$\textsc{Rel}(s, T) = \lambda\, \textsc{ERel}(s, T) + (1 - \lambda)\, \textsc{DRel}(s, T)$$

Approach

- **Event Selection:**

* Entity Relevance:

$$\text{EREL}(s, T) = w_1^e \, \text{E2E}(s, T) + w_2^e \, \text{E2EPATH}(T) + w_3^e \, \text{G2E}(T)$$

- E2E: relevance of specific events

$$\text{E2E}(s, T) = \sum_{re \in \{\text{RE}(e) \mid e \in T\}} \text{E2ECOOC}(s, re)$$

$$\text{E2ECOOC}(s, re) = \text{NPMI}(s; re) = \frac{\text{PMI}(s; re)}{-\log p(s, re)}$$

$$\text{PMI}(s; re) = \log \frac{p(s, re)}{p(s)\, p(re)}$$

Normalized pointwise mutual information

Approach

- **Event Selection:**

* Entity Relevance:

- E2EPath: relevance of paths

$$\text{E2EPath}(T) = \sum_{p \in \{\pi_{re}(e) \mid e \in T\}} \operatorname*{mean}_{e \in \mathcal{E},\, \pi_{re}(e) = p} \text{E2ECooc}(\text{Sub}(e), \text{RE}(e))$$

- G2E: how globally the relevant events are

$$\text{G2E}(T) = \sum_{re \in \{\text{RE}(e) \mid e \in T\}} \text{GlobalImportance}(re)$$

Approach

- **Event Selection:**

* Date Relevance:

$$\textsc{DRel}(s, T) = w_1^d \, \mathrm{E2D}(s, T) + w_2^d \, \mathrm{E2D}\textsc{Path}(T)$$

- E2D and E2DPath are similar to E2E and E2EPath

$$\mathrm{E2D}(s, T) = \sum_{t \in \{\tau(e) \mid e \in T\}} \mathrm{E2D}\textsc{Cooc}(s, t)$$

$$\mathrm{E2D}\textsc{Path}(T) = \sum_{p \in \{\pi_t(e) \mid e \in T\}} \operatorname*{mean}_{e \in \mathcal{E},\, \pi_t(e) = p} \mathrm{E2D}\textsc{Cooc}(\textsc{Sub}(e), \tau(e))$$

Approach

- **Event Selection:**

* **Temporal Diversity Constraint:**
  The layout constraint requires that all events can fit into a timeline of width W and height H without overlap.
  $T \sqcap R = \{e \in T \mid \tau(e) \in R\}$

$$\textsc{Constraints}(T, E, \mathcal{W}, w, n) \text{ if } \forall t \in \mathbb{R} : |T \sqcap [t, t + t_w)| \leq n$$

$$n = \left\lfloor \frac{Timeline\ Height}{Block\ Height} \right\rfloor$$

* Greedy algorithm has certain approximation guarantee.

Outline

- Introduction
- Approach
- **Experiment**
- Conclusion

Experiment

- Setup:

* Amazon Mechanical Turk 1154 raters to vote

* 250 popular entities(75 artists, 75 actors, 50 politicians, 50 athletes)

Name	w_1^e	w_2^e	w_3^e	w_1^d	w_2^d	TD	CD
FULL	1	10^{-2}	10^{-4}	1	10^{-2}	1	1
BASE	0	0	1	0	0	1	1
FULL-E2D	1	10^{-2}	10^{-4}	0	0	1	1
FULL-E2E	0	0	10^{-4}	1	10^{-2}	1	1
FULL-TD	1	10^{-2}	10^{-4}	1	10^{-2}	0	1
FULL-CD	1	10^{-2}	10^{-4}	1	10^{-2}	1	0

Table 1: Summary of experimental configurations. We fix $\lambda = 0.75$ throughout. TD = temporal diversity, CD = content diversity. FULL-TD means the full model without temporal diversity, etc. Note that all methods remove duplicate events, which is a minimal form of content diversity, but if CD=1, we ensure diversity amongst types of events (entities and paths) as well; see Section 4.5 for details.

Experiment

- Evaluation:
  Collapsed the rater votes to a 3-point scale: prefer control (full model), neutral, or prefer experiment (ablated model).

- $V(e,m,r) \in \{F,T,A\}$, $N(e,m,v) = |\{r \in R : V(e,m,r) = v\}|$
  $M(e, m) = \arg\max_v N(e, m, v)$

$$RAggr(m) = \frac{|\{V(e,m,r) = M(e,m) : e \in E, r \in R\}|}{|\{V(e,m,r) : e \in E, r \in R\}|}$$

$$RPref(m) = \frac{|\{M(e,m) = F : e \in E\}|}{|\{M(e,m) \in \{F,A\} : e \in E\}|}$$

Experiment

Ablated model	#Tasks	#Raters	RAggr	RPref
BASE	1250	344	77.0%	83.8% ***
FULL-E2D	1250	463	75.7%	59.8% **
FULL-E2E	1250	676	73.2%	64.3% ***
FULL-TD	150	53	75.3%	86.7% ***
FULL-CD	1250	665	81.0%	91.1% ***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Summary of the user studies. Each row shows an ablated version that was compared to the full model. The asterisks represent the p-value corresponding to a Binomial hypothesis test that compares the RPref value to 50%.

Experiment

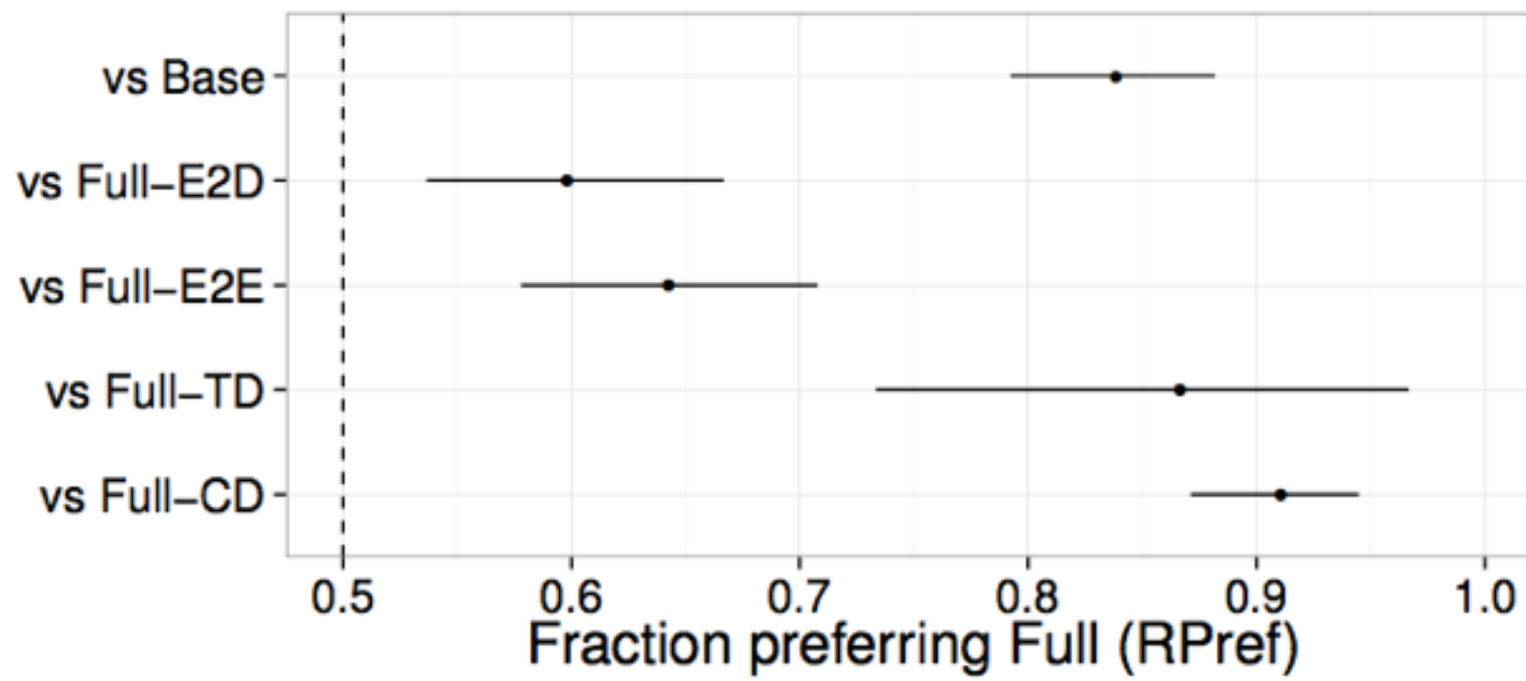


Figure 6: Fraction of entities for which raters preferred Full approach over ablated version (RPref) along with bootstrapped 95% confidence intervals. Results show significant preference for our proposed approach in all cases.

Outline

- Introduction
- Approach
- Experiment
- **Conclusion**

Conclusion

- Presented a system called TimeMachine for automatic timeline generation for entities in a knowledge base.
- Formulated in a submodular optimization framework that jointly optimizes for relevance, content diversity and temporal diversity and web-based co-occurrence signals are used.
- Proved that an efficient greedy approximation algorithm achieves near-optimal performance.