Search & Analytics in Archives using Semantic Annotations

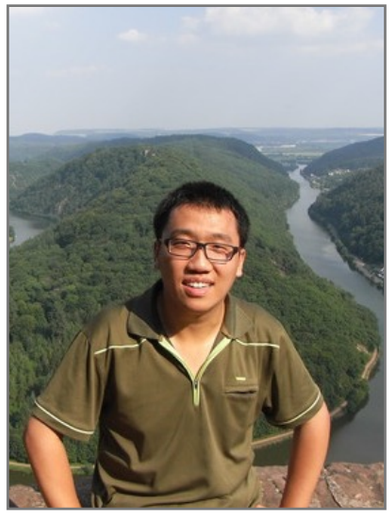
Klaus Berberich
(kberberi@mpi-inf.mpg.de)

htw saar
max planck institut informatik

Team / About

- Research area **Text + Time Search & Analytics** at MPI-INF

me



Kai Hui

Arunav Mishra

Dhruv Gupta

+ close collaboration with **Jannik Strötgen** and others

- Focus: **Efficient** and **effective** methods to **search** and **analyze text collections** with **temporal information** (e.g., temporal expressions, timestamps)

Motivation

- Idea of this talk is to give you an **overview of our research** during the past two-and-a-half years without diving too deep into **technical details**

- **Semantic annotations** as common ground of all our methods
  - What are they? How do we **obtain** and **process** them?
  - Which **models** do we use to **make sense** of them?

- Each of our **recent papers** presented on **at most two slides** (no technical details, but feel free to ask technical questions)

Outline

- Motivation
- Archives
- Semantic Annotations
- Search
- Analytics
- Outlook

Archives

- **Newspaper archives**
  - Examples:
    - **The New York Times Annotated Corpus**
    - **English Gigaword Corpus**
  - clean language, reliable meta-data (e.g., publication dates)

- **Web archives**
  - Examples:
    - **ClueWeb09/12**
    - **UKGOV**
  - messy language, unreliable meta-data (e.g., no publication dates)

Outline

- Motivation
- Archives
- Semantic Annotations
- Search
- Analytics
- Outlook

Semantic Annotations

- Our recent methods leverage **semantic annotations** to make documents **more than sequences of text tokens**
  - (disambiguated) mentions of **named entities** (e.g., persons)
  - **geographic references** (e.g., cities or countries)
  - **temporal expressions** (e.g., to past dates)

- We rely on **existing off-the-shelf tools** and **some handcrafted rules** to obtain semantic annotations for documents

(Disambiguated) Named Entities

In **Indianola, Mississippi**, … , a new museum will begin construction in June to honor a local boy who has become international star, the blues performer **B.B. King**. The **B.B. King Museum and Delta Interpretive Center** will begin with restoration of the cotton mill, … , and is expected to be open by 2007, with additions following. Artifacts from **Mr. King's** 60-year career will be housed in the museum and the interpretive center will focus on educational, cultural and character development programs for **Mississippi** youth.

- **AIDA** (https://www.ambiverse.com) as our **named entity recognition and disambiguation** (NERD) tool of choice

(Disambiguated) Named Entities

isLocatedIn

hasLatitude 33.43°

hasLongitude -90.63°

wordnet_guitarist

American_Blues_Guitarist

Riley B. King

Bluesboy

- Named entities are anchored in the **YAGO knowledge graph**, which provides us with **additional information** about them
  - **semantic types** (from WordNet and Wikipedia)
  - surface forms**, keyphrases, and links**
  - general facts

Geographic References




isLocatedIn

hasLatitude 33.43°

hasLongitude -90.63°

- **Geographic references derived** from named entity mentions by considering only entities having type `yagoGeoEntity`
- **Minimum-bounding rectangle** (MBR) indicating geographic extent of a location determined by making use of `isLocatedIn` relationship and geographic coordinates from YAGO

Geographic References

isLocatedIn
hasLatitude 33.43°
hasLongitude -90.63°

isLocatedIn
hasLatitude 34.96°
hasLongitude -89.98°

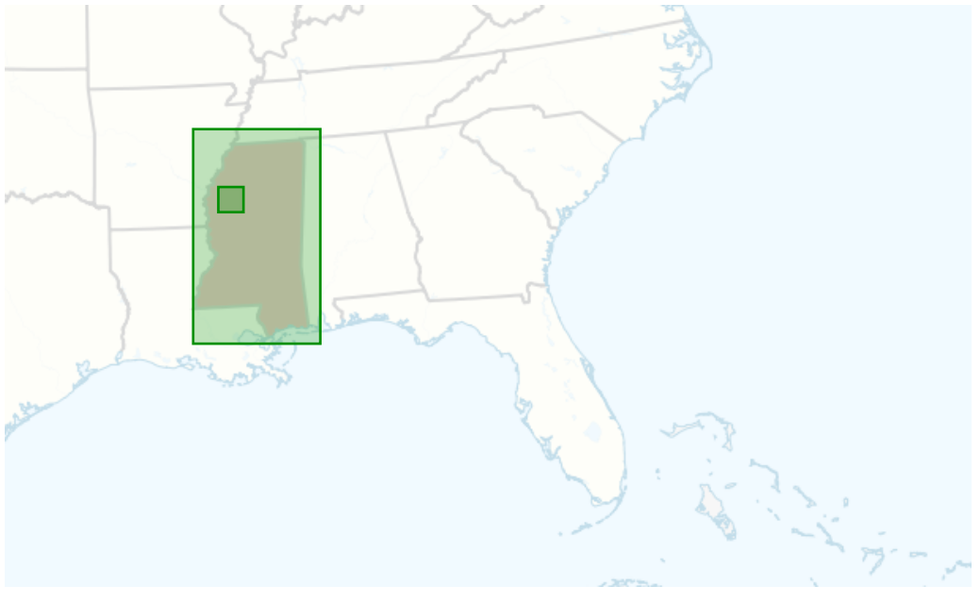
isLocatedIn
hasLatitude 31.00°
hasLongitude -90.64°

- **Outlier removal** found to be helpful (e.g., Hawaii or Alaska)
- Open question whether **richer representation** (e.g., convex hull) would be beneficial

Geographic References

- **Probabilistic generative model** for latitude-longitude pairs based on a set of MBRs (e.g., from document or query)
  - **draw a MBR** at uniform random
  - **draw a latitude-longitude pair** contained in MBR at uniform random

In **Indianola, Mississippi**, ... , a new museum will begin construction in June to honor a local boy who has become international star, the blues performer B.B. King. The B.B. King Museum and Delta Interpretive Center will begin with restoration of the cotton mill, ... , and is expected to be open by 2007, with additions following. Artifacts from Mr. King's 60-year career will be housed in the museum and the interpretive center will focus on educational, cultural and character development programs for **Mississippi** youth.



Temporal Expressions

In Indianola, Mississippi, … , a new museum will begin construction **in June** to honor a local boy who has become international star, the blues performer B.B. King. The B.B. King Museum and Delta Interpretive Center will begin with restoration of the cotton mill, … , and is expected to be open by **2007**, with additions following. Artifacts from Mr. King's 60-year career will be housed in the museum and the interpretive center will focus on educational, cultural and character development programs for Mississippi youth.

- **Temporal expressions** annotated using
  - **SUTime** (http://nlp.stanford.edu/software/sutime.shtml)
  - **HeidelTime** (https://github.com/HeidelTime/heideltime)
- **Reliable publication dates** are crucial for correct resolution of **relative temporal expressions** (e.g., “last year”, “next month”)

Temporal Expressions

In Indianola, Mississippi, … , a new museum will begin construction **in June** to honor a local boy who has become international star, the blues performer B.B. King. The B.B. King Museum and Delta Interpretive Center will begin with restoration of the cotton mill, … , and is expected to be open by **2007**, with additions following. Artifacts from Mr. King's 60-year career will be housed in the museum and the interpretive center will focus on educational, cultural and character development programs for Mississippi youth.

`[2007/01/01, 2007/12/31, 2007/01/01, 2007/12/31]`

`[2005/06/01, 2005/06/30, 2005/06/01, 2005/06/30]`

- **Meaning** of temporal expressions is often **vague** (e.g., “in June”)
  - capture **earliest/latest begin/end time point** of any precise **time interval** that the temporal expression **may refer** to

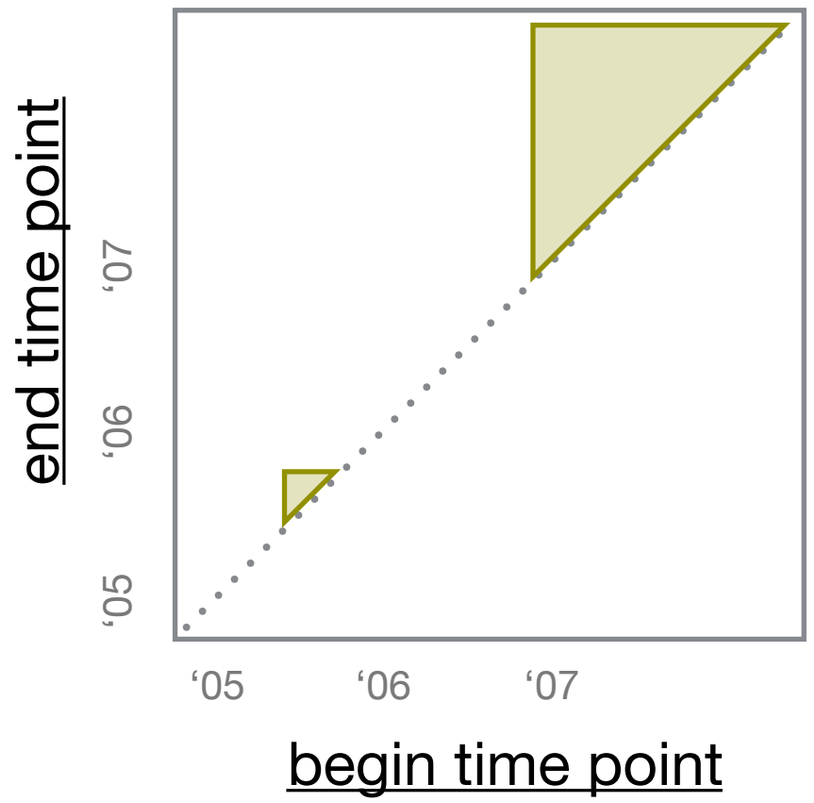
Temporal Expressions

- **Probabilistic generative model** for time intervals
  based on a set of temporal expressions
  (e.g., from document of query)
  - **draw a temporal expression** (quadruple) at uniform random
  - **draw a time interval** that the temporal expression may refer to

In Indianola, Mississippi, ... , a new museum will begin construction **in June** to honor a local boy who has become international star, the blues performer B.B. King. The B.B. King Museum and Delta Interpretive Center will begin with restoration of the cotton mill, ... , and is expected to be open by **2007**, with additions following. Artifacts from Mr. King's 60-year career will be housed in the museum and the interpretive center will focus on educational, cultural and character development programs for Mississippi youth.

`[2007/01/01, 2007/12/31, 2007/01/01, 2007/12/31]`

`[2005/06/01, 2005/06/30, 2005/06/01, 2005/06/30]`



Outline

- Motivation
- Archives
- Semantic Annotations
- Search
- Analytics
- Outlook

Identifying Interesting Time Intervals

- Idea: Given a **keyword query** (e.g., `world war`), determine **interesting time intervals** (e.g., [1914, 1918] or [1939, 1945]) that help the user **refine the query** to **explore results**

$$\mathrm{P}\left[\,[tb, te] \mid q\,\right] \quad = \sum_{d \in top(q,k)} \mathrm{P}\left[\,[tb, te] \mid d\,\right] \cdot \mathrm{P}\left[\,d \mid q\,\right]$$

D. Gupta and K. Berberich: *Identifying Time Intervals of Interest to Queries*, **CIKM 2014**

Identifying Interesting Time Intervals

- Idea: Given a **keyword query** (e.g., `world war`), determine **interesting time intervals** (e.g., [1914, 1918] or [1939, 1945]) that help the user **refine the query** to **explore results**

$$\mathrm{P}\left[\,[tb, te] \mid q\,\right] \quad = \sum_{d \in top(q,k)} \mathrm{P}\left[\,[tb, te] \mid d\,\right] \cdot \mathrm{P}\left[\,d \mid q\,\right]$$

Generative model based on temporal expressions

D. Gupta and K. Berberich: *Identifying Time Intervals of Interest to Queries*, **CIKM 2014**

Identifying Interesting Time Intervals

- Idea: Given a **keyword query** (e.g., `world war`), determine **interesting time intervals** (e.g., [1914, 1918] or [1939, 1945]) that help the user **refine the query** to **explore results**

$$\mathrm{P}\left[\,[tb, te] \mid q\,\right] = \sum_{d \in top(q,k)} \mathrm{P}\left[\,[tb, te] \mid d\,\right] \cdot \mathrm{P}\left[\,d \mid q\,\right]$$

Generative model based on temporal expressions

Document likelihood ~ query likelihood

D. Gupta and K. Berberich: *Identifying Time Intervals of Interest to Queries*, **CIKM 2014**

Temporal Diversification of Search Results

- Idea: **Re-rank documents** to **cover** all **interesting time intervals** (e.g., [1914, 1918] or [1933, 1945]) previously identified for a **keyword query** (e.g., `world war`)

$$\underset{R:|R|=k}{arg\,max} \sum_{[tb,te]} \left(\mathrm{P}\left[[tb,te] \mid q \right] \left(1 - \prod_{d \in R} \left(1 - \mathrm{P}\left[[tb,te] \mid d \right] \cdot \mathrm{P}\left[d \mid q \right] \right) \right) \right)$$

D. Gupta and K. Berberich: *Diversifying Search Results Using Time*,
ECIR 2016

Temporal Diversification of Search Results

- Idea: **Re-rank documents** to **cover** all **interesting time intervals** (e.g., [1914, 1918] or [1933, 1945]) previously identified for a **keyword query** (e.g., `world war`)

$$\underset{R:|R|=k}{arg\,max} \sum_{[tb,te]} \left(\mathrm{P}\left[[tb,te] \mid q \right] \left(1 - \prod_{d \in R} \left(1 - \mathrm{P}\left[[tb,te] \mid d \right] \cdot \mathrm{P}\left[d \mid q \right] \right) \right) \right)$$

Probability that time interval $[tb,te]$ is interesting for query q

D. Gupta and K. Berberich: *Diversifying Search Results Using Time*, **ECIR 2016**

Temporal Diversification of Search Results

- Idea: **Re-rank documents** to **cover** all **interesting time intervals** (e.g., [1914, 1918] or [1933, 1945]) previously identified for a **keyword query** (e.g., `world war`)

$$\underset{R:|R|=k}{arg\,max} \sum_{[tb,te]} \left(\mathrm{P}\left[[tb,te] \mid q \right] \left(1 - \prod_{d \in R} \left(1 - \mathrm{P}\left[[tb,te] \mid d \right] \cdot \mathrm{P}\left[d \mid q \right] \right) \right) \right)$$

Probability that time interval $[tb,te]$ is interesting for query q

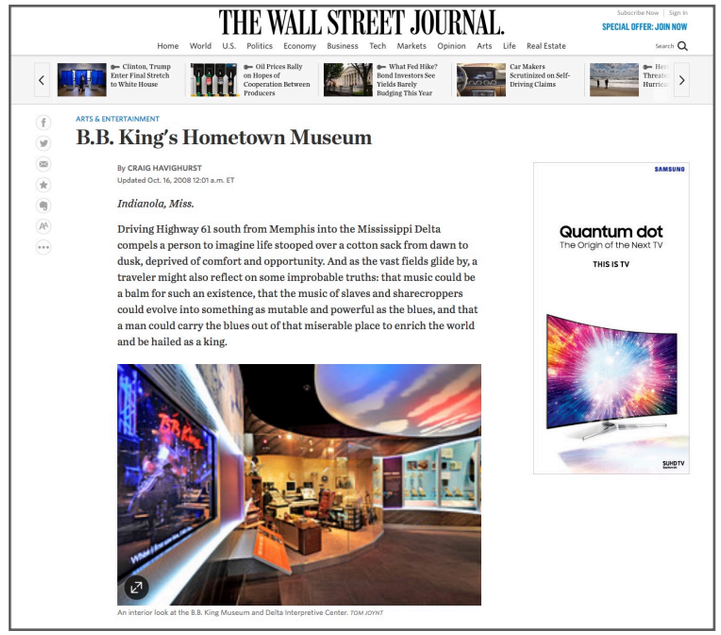
Probability that user sees at least one document in R that covers $[tb,te]$

D. Gupta and K. Berberich: *Diversifying Search Results Using Time*, **ECIR 2016**

Linking Wikipedia and News Archives

- Idea: Given an **excerpt from Wikipedia**, automatically retrieve **articles** from a **news archive** that provide **in-depth information**

> The same month, a groundbreaking was held for a new museum, dedicated to King,[44] in Indianola, Mississippi.[45] The B.B. King Museum and Delta Interpretive Center opened on September 13, 2008.[46]



THE WALL STREET JOURNAL.

SPECIAL OFFER: JOIN NOW

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Clinton, Trump Enter Final Stretch to White House

Oil Prices Rally on Hopes of Cooperation Between Producers

What Fed Hike? Bond Investors See Yields Barely Budging This Year

Car Makers Scrutinized on Self-Driving Claims

ARTS & ENTERTAINMENT

B.B. King's Hometown Museum

By CRAIG HAVIGHURST
Updated Oct. 16, 2008 12:01 a.m. ET

Indianola, Miss.

Driving Highway 61 south from Memphis into the Mississippi Delta compels a person to imagine life stooped over a cotton sack from dawn to dusk, deprived of comfort and opportunity. And as the vast fields glide by, a traveler might also reflect on some improbable truths: that music could be a balm for such an existence, that the music of slaves and sharecroppers could evolve into something as mutable and powerful as the blues, and that a man could carry the blues out of that miserable place to enrich the world and be hailed as a king.

An interior look at the B.B. King Museum and Delta Interpretive Center. TOM JOYNT

- **Divergence-based re-ranking of top-K documents**

$$KL(Q\|D) = KL(Q_{txt}\|D_{txt}) + KL(Q_{time}\|D_{time}) + KL(Q_{geo}\|D_{geo}) + KL(Q_{entity}\|D_{entity})$$

A. Mishra and K. Berberich: *Leveraging Semantic Annotations to Link Wikipedia and News Archives*, **ECIR 2016**

Outline

- Motivation
- Archives
- Semantic Annotations
- Search
- Analytics
- Outlook

Generating Event Digests

- Idea: Given an **event query** (e.g., from a Wikipedia year page), automatically generate a **summary of the event** having a **user-specified length**

1999/08/30 **East Timor** votes for independence from **Indonesia** in referendum

- **ILP** selects **sentences** from top-k pseudo-relevant documents
  - **minimize divergence(s)** between sentences and event query
  - **cover** all **temporal expressions**, **geographic references**, and **named entities** from event query
  - **adhere** to **length budget**

Event Digest (with chronological ordering on publication dates)

• **Publication Date:** *July 20, 1999* **Source Link:** *http://goo.gl/rJYDiZ*
(1) **Indonesia** is preparing to relinquish control of **East Timor** after 23 years of occupation and it believes that independence advocates are highly likely to win a referendum **next month** says an authentic internal government report that has been made available to reporters by advocates of independence. **(2)** **Late next month** estimated 400 000 **East Timorese** are to choose between broad autonomy within **Indonesia** option 1 or independence option 2.

• **Publication Date:** *August 29, 1999* **Source Link:** *http://goo.gl/Cz6Jkk*
(3) Former president **Jimmy Carter** whose human rights and diplomacy organization the **Carter Center** is monitoring the referendum here said this **this month** some top representatives of the government of **Indonesia** have failed to fulfill their main obligations with regard to public order and security.

• **Publication Date:** *November 21, 1999* **Source Link:** *http://goo.gl/hdqYm8*
(4) The last time it was **East Timor** which voted for independence from **Indonesia** in **August** only to be plunged into a spasm of violence that required an **Australian** led international military force to quell it. **(5)** **Acehs** latest push for independence began with the fall of President **Suharto** in **May 1998** and accelerated after the **East Timor** referendum.

• **Publication Date:** *September 24, 2000* **Source Link:** *http://goo.gl/AijWVY*
(6) **East Timor** has been under a transitional **United Nations** administration since the **Aug. 30** independence vote **last year**. **(7)** The groups pillaged **East Timor** after **last year's** independence vote which freed the territory from military control.

• **Publication Date:** *August 24, 2001* **Source Link:** *http://goo.gl/EAGBxC*
(8) This vote like the referendum in **1999** is being organized by the **United Nations** which has continued to administer **East Timor** a former **Portuguese colony** annexed by **Indonesia** as it struggles to its feet economically and politically.

A. Mishra and K. Berberich: *Event Digest: A Holistic View on Past Events*, **SIGIR 2016**

Mining Events

- Idea: **Identify and describe important events** for a given query

lord of the rings movie

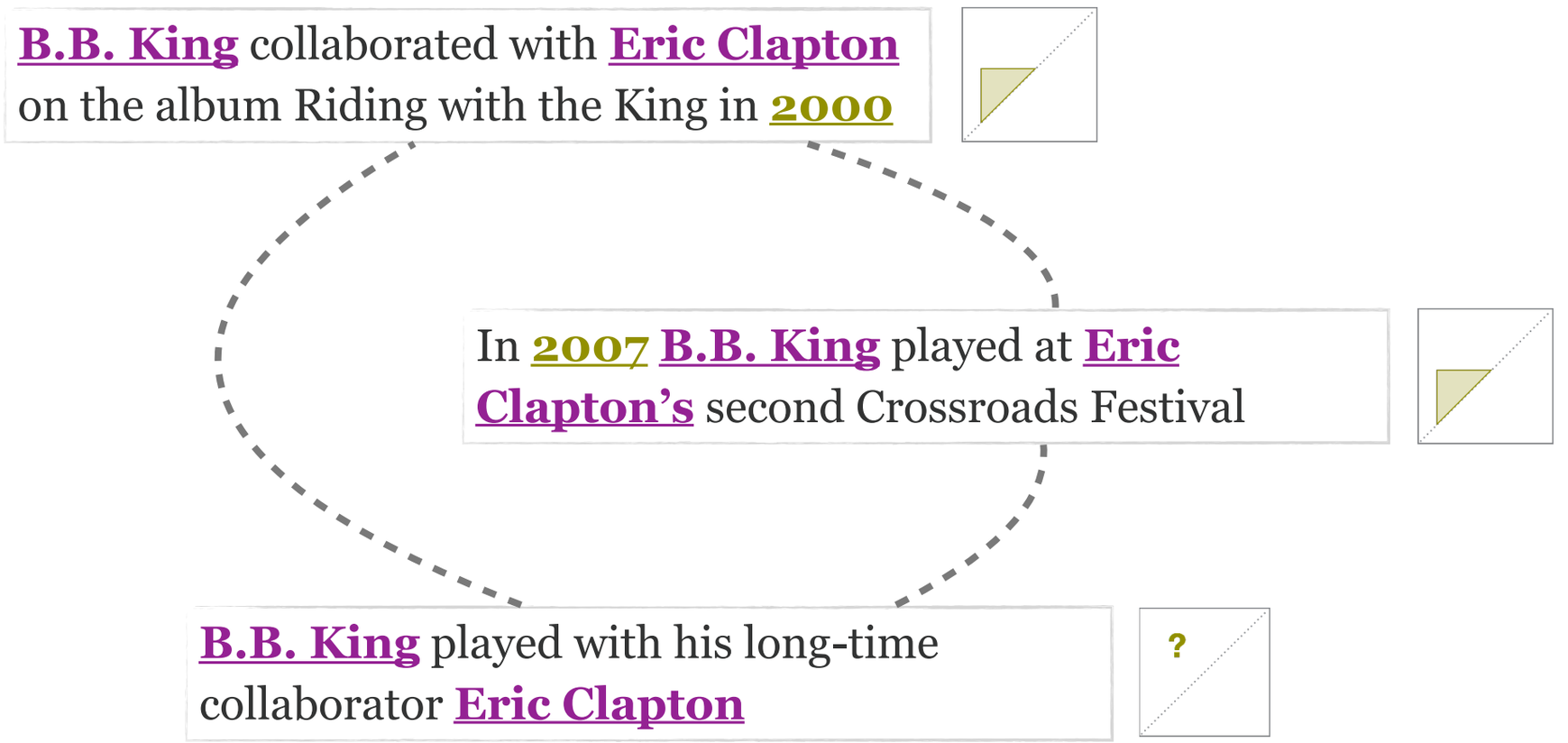
Keywords	[lord] [rings] [top] [movie] [motion] [opinion] [pictures] [article] [elvis] [jackson] [trilogy] [movies]
Time	[15-Dec-2002 , 15-Dec-2002] [01-Jan-1987 , 01-Jan-1987] [25-Jan-2004 , 25-Jan-2004] [12-Nov-2002 , 12-Nov-2002] [01-Jan-2003 , 31-Dec-2003] [01-Jan-1982 , 01-Jan-1982] [11-Jan-2004 , 11-Jan-2004] [28-Dec-2002 , 29-Dec-2002] [07-Sep-2003 , 07-Sep-2003] [01-Dec-2003 , 31-Dec-2003]
Locations	[YAGO:Weldon,_Northamptonshire] [YAGO:Wellington]
Entities	[YAGO:J._R._Tolkien] [YAGO:Weldon,_Northamptonshire] [YAGO:Wellington] [YAGO:Carol_Ann_Lee] [YAGO:Peter_Jackson]

- **Clustering of sentences** from top-k pseudo-relevant documents based on a **Dirichlet process mixture model** (DPM) with resulting **clusters as events** described by probability distributions over keywords, time intervals, geographic coordinates, and named entities

D. Gupta, J. Strötgen, and K. Berberich: *EventMiner: Mining Events from Annotated Documents*, **ICTIR 2016**

Estimating Time Models

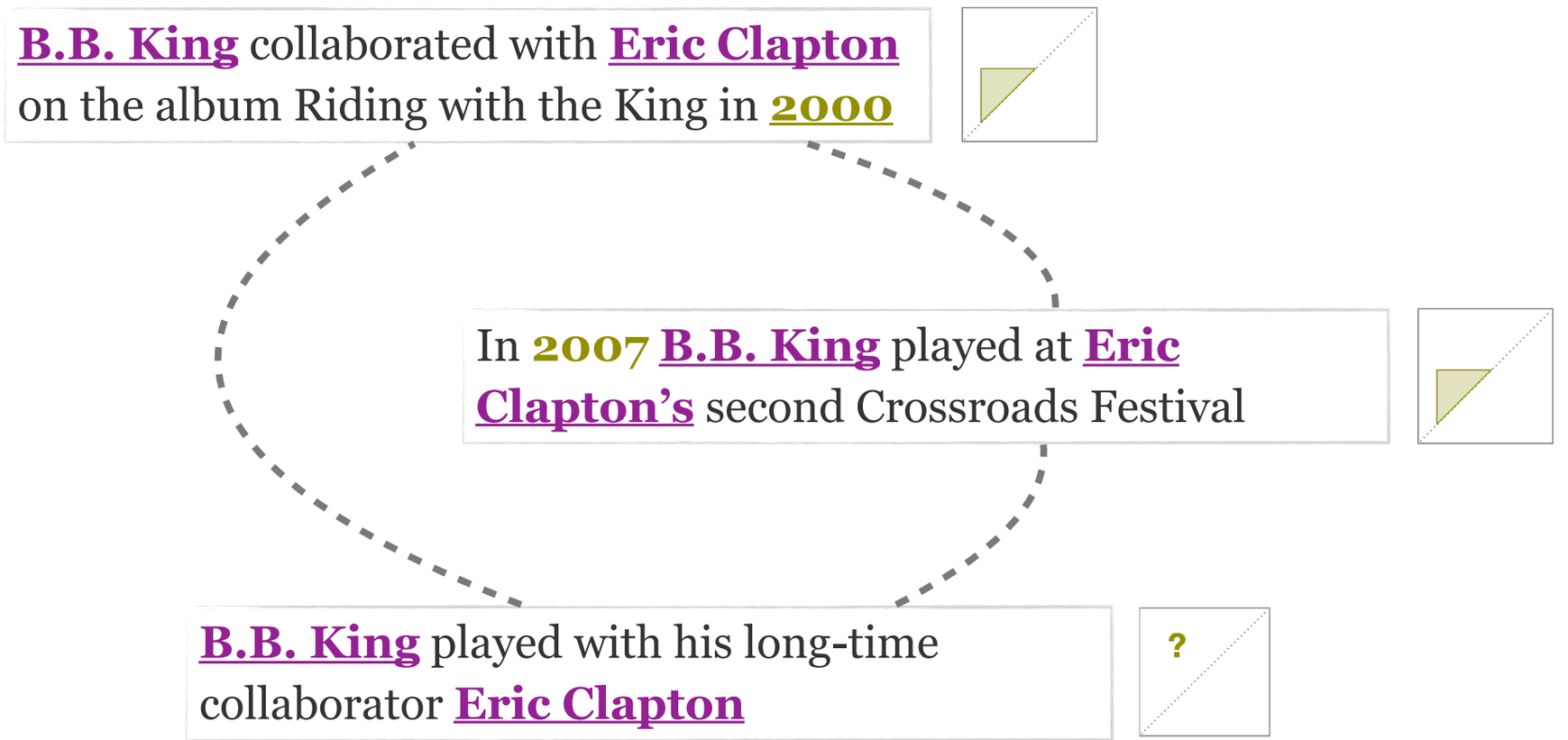
- Idea: Estimate a **time model** (i.e., probability distribution over time intervals) for **excerpts** (e.g., sentences or paragraphs) that do **not contain** temporal expressions



A. Mishra and K. Berberich: *Estimating Time Models for News Article Excerpts*, **CIKM 2016**

Estimating Time Models

- **Distribution propagation approach** with edge weights based on text similarity, conceptual similarity, and contextual similarity



A. Mishra and K. Berberich: *Estimating Time Models for News Article Excerpts*, CIKM 2016

Outline

- Motivation
- Archives
- Semantic Annotations
- Search
- Analytics
- **Outlook**

Outlook

- **Extension** of **probabilistic generative model** for time intervals to support **different temporal granularities** (day, month, year)

- **Narrative generation** for event digests to make them more natural for humans (e.g., chronological order, co-references)

- **Index structures** for documents with **semantic (and linguistic) annotations** from which all our methods can potentially profit

- **Word embeddings** for semantic annotations to have a **uniform representation** of text, temporal expressions, geographic references, and named entities and simplify our methods

Thank You!

Questions?