Metric Spaces for Temporal Information Retrieval

Matteo Brucato[1], Danilo Montesi[2]

[1] University of Massachusetts Amherst, USA
[2] University of Bologna, Italy

Presented by:
Matteo Brucato
matteo@cs.umass.edu

Time and Temporal Scope

- Time is an ubiquitous **dimension** of nearly every collection of documents
  - Digital libraries, news stories, tweets, the Web, …
- Documents
  - Meta-level: creation, publication date, …
  - Content-level: **Periods of time** mentioned in the text
    ⇒ The **document temporal scope**
- Queries
  - Meta-level: issue date, …
  - Content-level: **Periods of time** mentioned in the query
    ⇒ The **query temporal scope**

Temporal Similarity: Motivation

- **Textual similarity**
  - Similarity based on **term statistics**
  - Not adequate for **temporal queries**:

    "**results elections 2008**"

    "**best movies last year**"

  - "2008" and "last year" are considered **terms** and searched **literally** in the documents

⇒ We need to model **temporal similarity**

Temporal Intervals

- Temporal **intervals** are semantically rich:
  - **Synonymy:**
    - "2013" = "last year" = "the year after 2012"
  - **Polysemy:**
    - "every friday", "yearly", "super bowl"
  - **Algebraic structure** (to correlate temporal scopes):
    - overlapping
    - containment
    - **Distance**

⇒ We can **exploit** this to improve IR models

Temporal Domain

Chronon

The *smallest* discrete unit of time (e.g., a second, a day, a year)

Temporal domain

$\Delta = [t_{min}, t_{min}], \ldots, [1990, 1991], [1990, 1992], \ldots, [t_{min}, t_{max}]$

Interpretation function

$\Psi : \text{TIMEX} \rightarrow \wp(\Delta)$

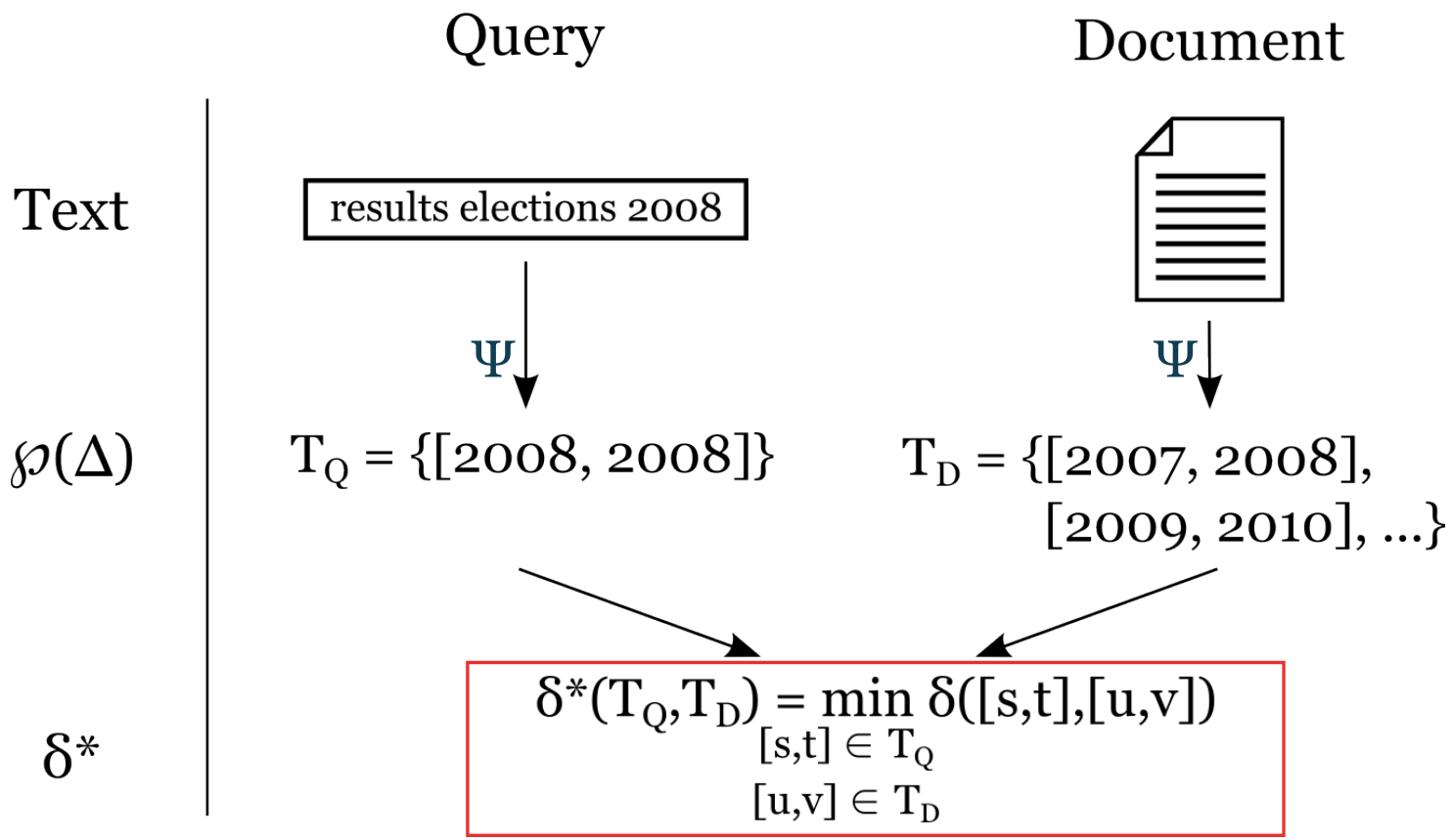
where TIMEX is the set of all possible time expressions

Temporal scope of a document D (or a query Q)

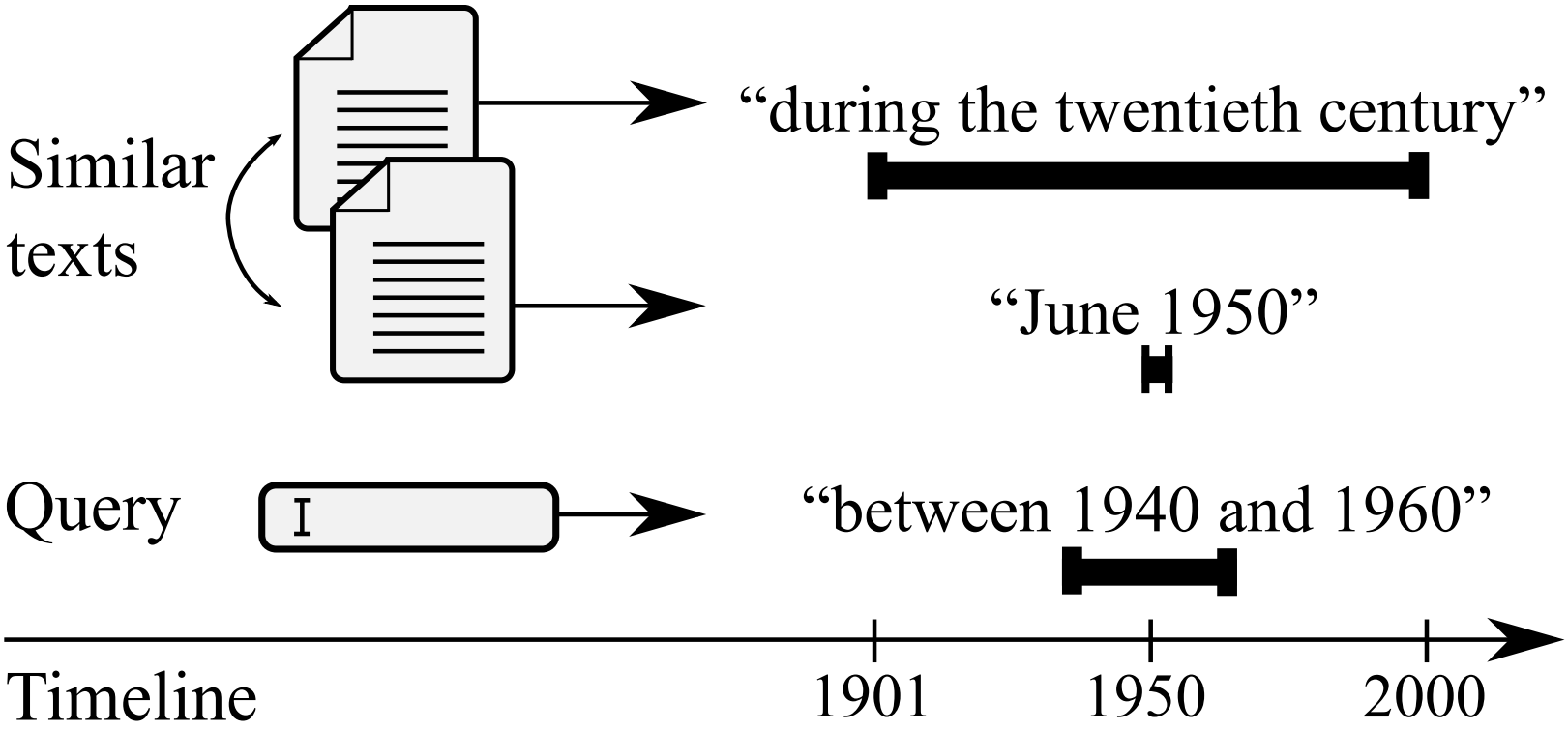
$T_D = \{ [1990, 1999], [1995, 1997], [2001, 2002] \}$

$T_Q = \{ [1991, 2001], [2002, 2003] \}$

The temporal similarity δ*

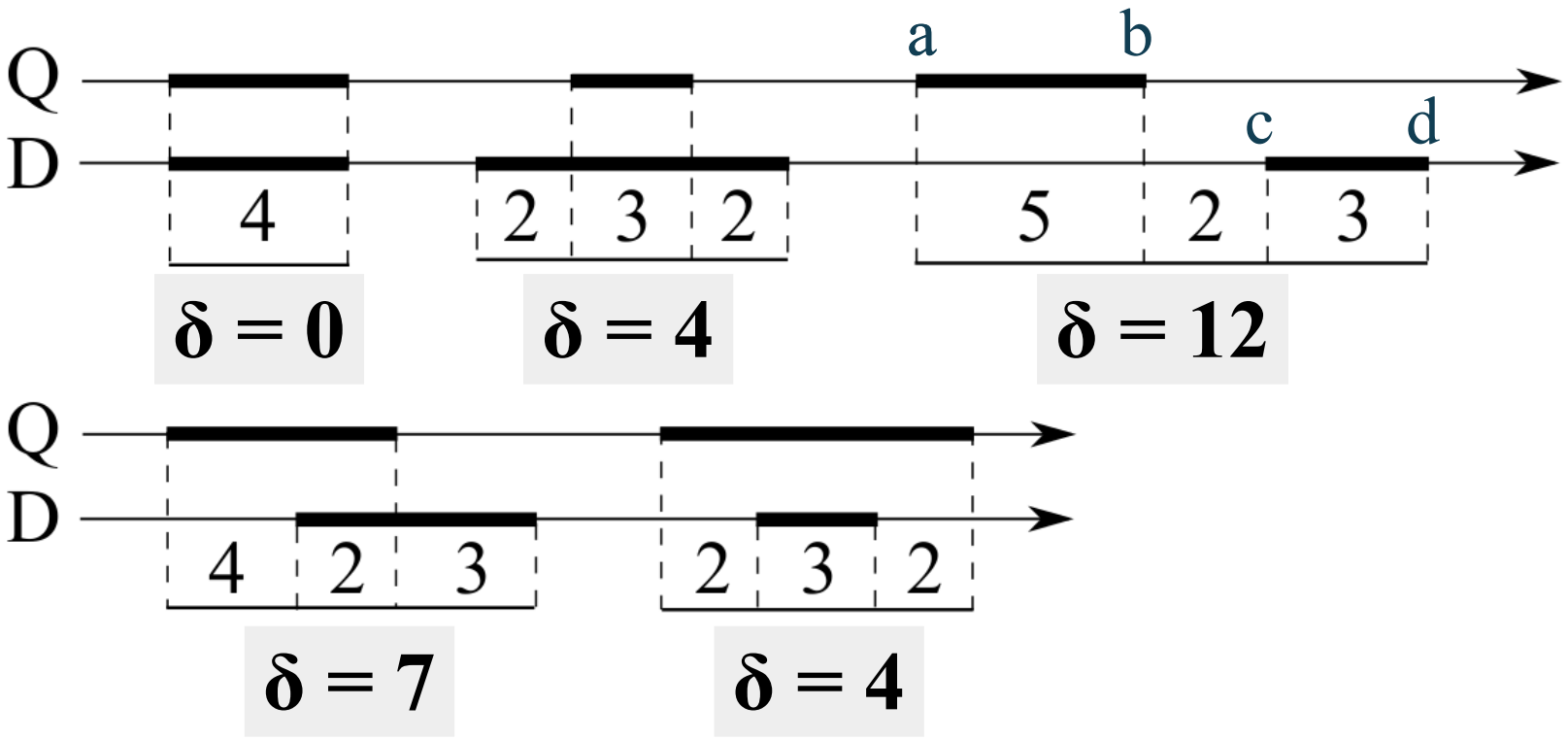


How can we effectively model δ?



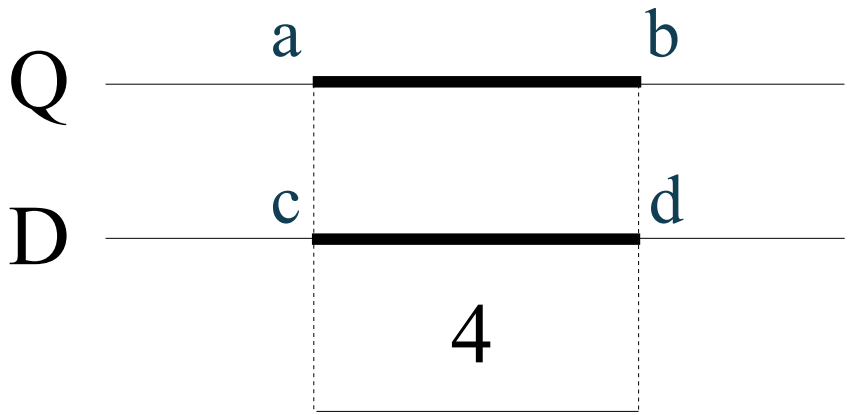
Simple solution: Manhattan Distance

$$\delta_{sym}([a,b]_Q, [c,d]_D) = |a - c| + |b - d|$$



Reasonable?

$$\delta_{sym}([a,b]_Q, [c,d]_D) = |a - c| + |b - d|$$

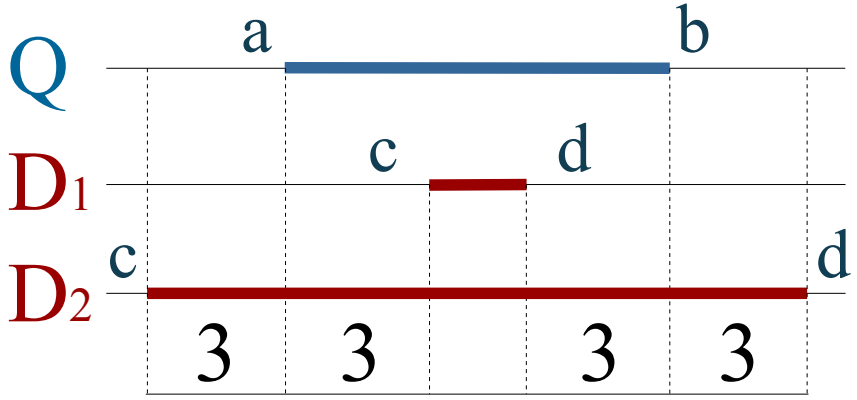


$$\delta = 0$$

This looks intuitively correct

Manhattan distance: Anomaly

$$\delta_{sym}([a,b]_Q, [c,d]_D) = |a - c| + |b - d|$$



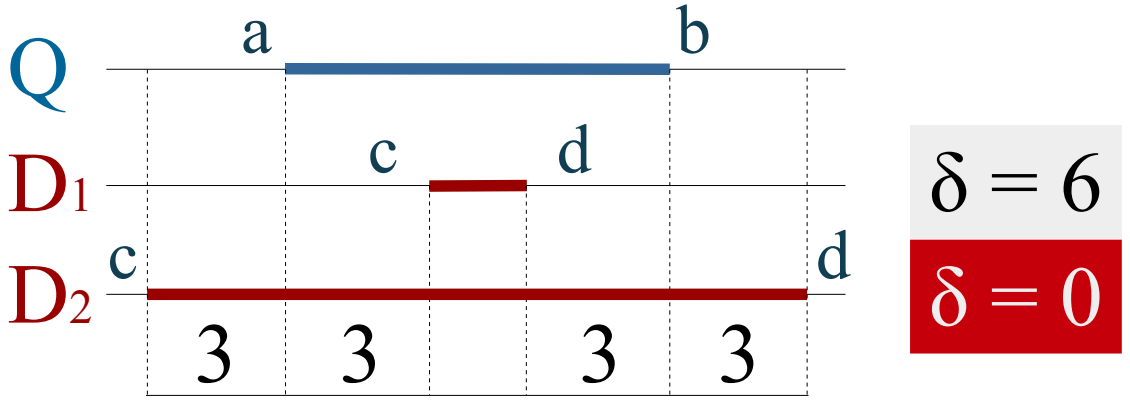
$\delta = 6$

$\delta = 6$

The two documents would have the same distance from the query...

Distance reflecting query coverage

$$\delta_{\mathbf{cov(Q)}}([a,b]_Q, [c,d]_D) = (b - a) - (\min\{b,d\} - \max\{a,c\})$$

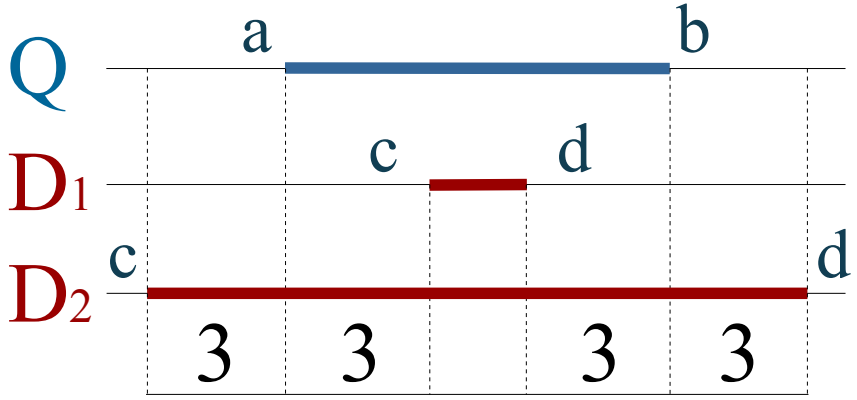


More appropriate for **"narrow" time queries:**

- Query represents the **narrowest time interval** the user is willing to accept
- Distance reflects **query coverage**

Distance reflecting document coverage

$$\delta_{\mathbf{cov(D)}}([a,b]_Q, [c,d]_D) = (d - c) - (\min\{b,d\} - \max\{a,c\})$$



More appropriate for **"broad" time queries:**

- Query represents the **broadest time interval** the user is willing to accept
- Distance reflects **document coverage**

Generalized metrics

- Metrics (e.g. Manhattan distance):
  - *Non-negativity*: $\delta(x,y) \geq 0$
  - *Coincidence*: $\delta(x,y) = 0$ *iff* $x = y$
  - *Symmetry*: $\delta(x,y) = \delta(y,x)$
  - *Triangle inequality*: $\delta(x,z) \leq \delta(x,y) + \delta(y,z)$
- The 2 new distances are **hemimetrics**:
  - *No symmetry*
  - *Partial coincidence*:
    - $\delta(x,x) = 0$
    - but we allow **y**'s, $\mathbf{y} \neq x$, such that: $\delta(x,\mathbf{y}) = 0$
- Interesting property:

$$\delta_{sym}(x,y) = \delta_{cov(D)}(x,y) + \delta_{cov(Q)}(x,y)$$

Combining text and time scores

- Temporal similarity:

  $sim_{\delta^*}(Q, D) = \exp\{-\delta^*(T_Q, T_D)\}$

- Two models of relevance
  - *Textual similarity*: sim_{kw}
  - *Temporal similarity*: sim_{δ^*}
- Combining them:

  $sim(Q, D_i) = (1 - \alpha)\, sim_{kw}(Q, D_i) + (\alpha)\, sim_{\delta^*}(Q, D_i)$

  where α is a combination parameter in [0,1]

Effectiveness Evaluation

Test Collection

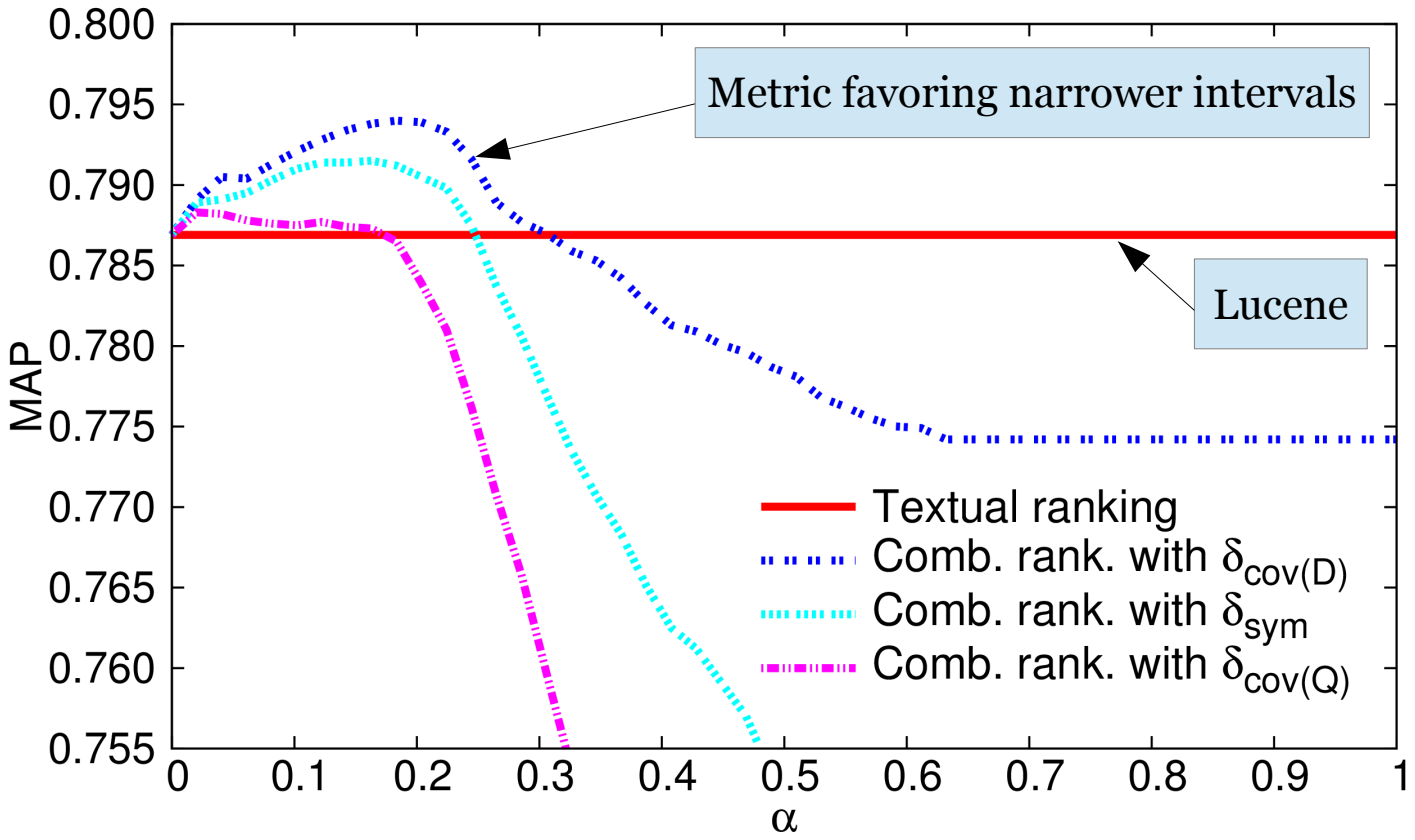
- TREC Novelty 2004:
  - 1808 articles from New York Times and other newswires
  - From January 1996 through September 2000 (almost 5 years)
  - "traditional" (and "novelty") relevance assessments
  - HeidelTime[1] and TIMEN[2] libraries to extract and normalize temporal expressions (aka “timexes”)

	Documents	Topic Descriptions	Topic Narratives
Number	1808	50	50
% containing timexes	75%	22%	10%

[1] https://code.google.com/p/heideltime/
[2] http://code.google.com/p/timen/

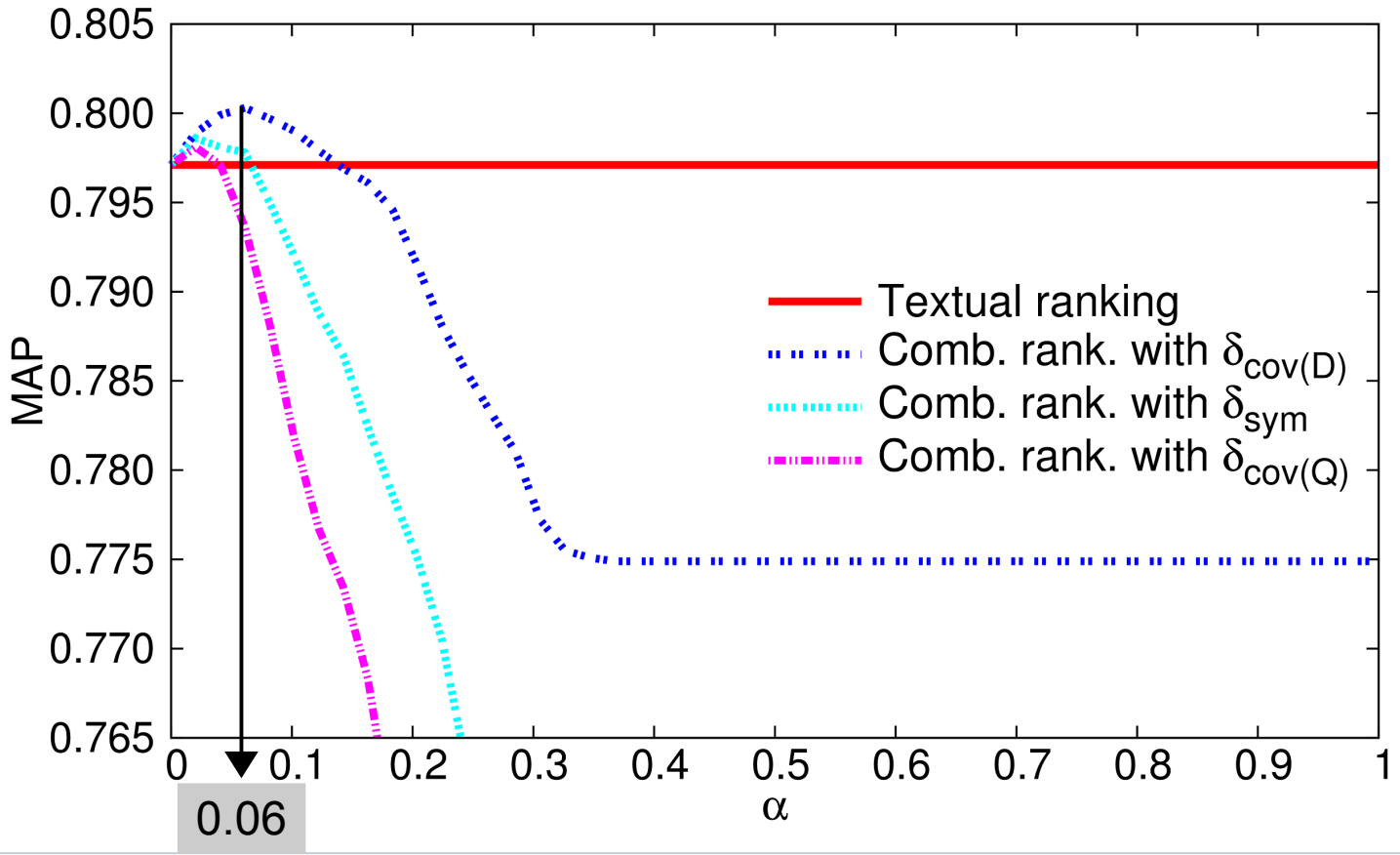
Comparing textual and combined ranking (1/2)

- *Textual queries*: Topic **titles**
- *Temporal queries*: All extracted temporal intervals



Comparing textual and combined ranking (2/2)

- *Textual queries*: Topic **descriptions**
- *Temporal queries*: All extracted temporal intervals



Impact on top-*k* for **all** queries

Considering all queries, temporal and non-temporal:

Textual Ranking (α = 0)

k	P@k	R@k	MAP@k
5	0.84	0.17	0.16
10	0.80	0.33	0.30
20	0.77	0.64	0.57

Combined Ranking (α = 0.06)

k	P@k	R@k	MAP@k
5	0.84	0.17	0.16
10	**0.81**	0.33	**0.31**
20	**0.78**	**0.65**	**0.58**

Best combination weight from previous experiment

Impact on top-*k* for **temporal** queries only

Considering only the 11 temporal queries:

Textual Ranking (α = 0)

k	P@k	R@k	MAP@k
5	**0.83**	0.18	0.17
10	0.79	0.34	0.31
20	0.76	0.66	0.57

Combined Ranking (α = 0.06)

k	P@k	R@k	MAP@k
5	0.81 ↓	0.18 ↑	0.17 ↑
10	**0.81**	**0.35** ↑	**0.32** ↑
20	**0.79** ↑	**0.69** ↑	**0.60** ↑

↓ Worst on temporal queries

↑ Better on temporal queries

Best combination weight from previous experiment

Summary of contributions

- Model for temporal scopes of documents and queries
- Three novel metrics for temporal scope similarity
- Ranking model combining textual and temporal scores
- Experimental evaluation of the effectiveness improvements over a text-only ranking
- The asymmetry and partial coincidence used for modeling the temporal similarity might have a meaning beyond just the time dimension

Closely Related Work

- Among the many interesting works on Temporal IR, these address the task from a very similar perspective:
  - Berberich, Bedathur, Alonso, Weikum in *Advances in Information Retrieval*, 2010:
    - Language modeling approach
    - Worse effectiveness with **no uncertainty** and **inclusive mode**
  - Khodaei, Shahabi, Khodaei in *International Journal of Next-Generation Computing*, 2012:
    - Emphasis on **index structures** for fast top-k retrieval
    - Ranking model considering **only overlap** (our metrics include the concept of overlap: they are more general)

Thank you!
Questions?

Matteo Brucato[1], Danilo Montesi[2]

[1] University of Massachusetts Amherst, USA
[2] University of Bologna, Italy

Presented by:
Matteo Brucato
matteo@cs.umass.edu