

Efficient Time-Travel on Versioned Text Collections

Klaus Berberich, Srikanta Bedathur, Gerhard Weikum

Max-Planck-Institut für Informatik, Saarbrücken



Motivation – Today...

- Versioned text collections available today: Web archives, Wikis, Information feeds,...
- Search is limited!
 - Only most recent versions are searched
 - Versions are treated as independent documents
- Time-travel search functionality is missing!

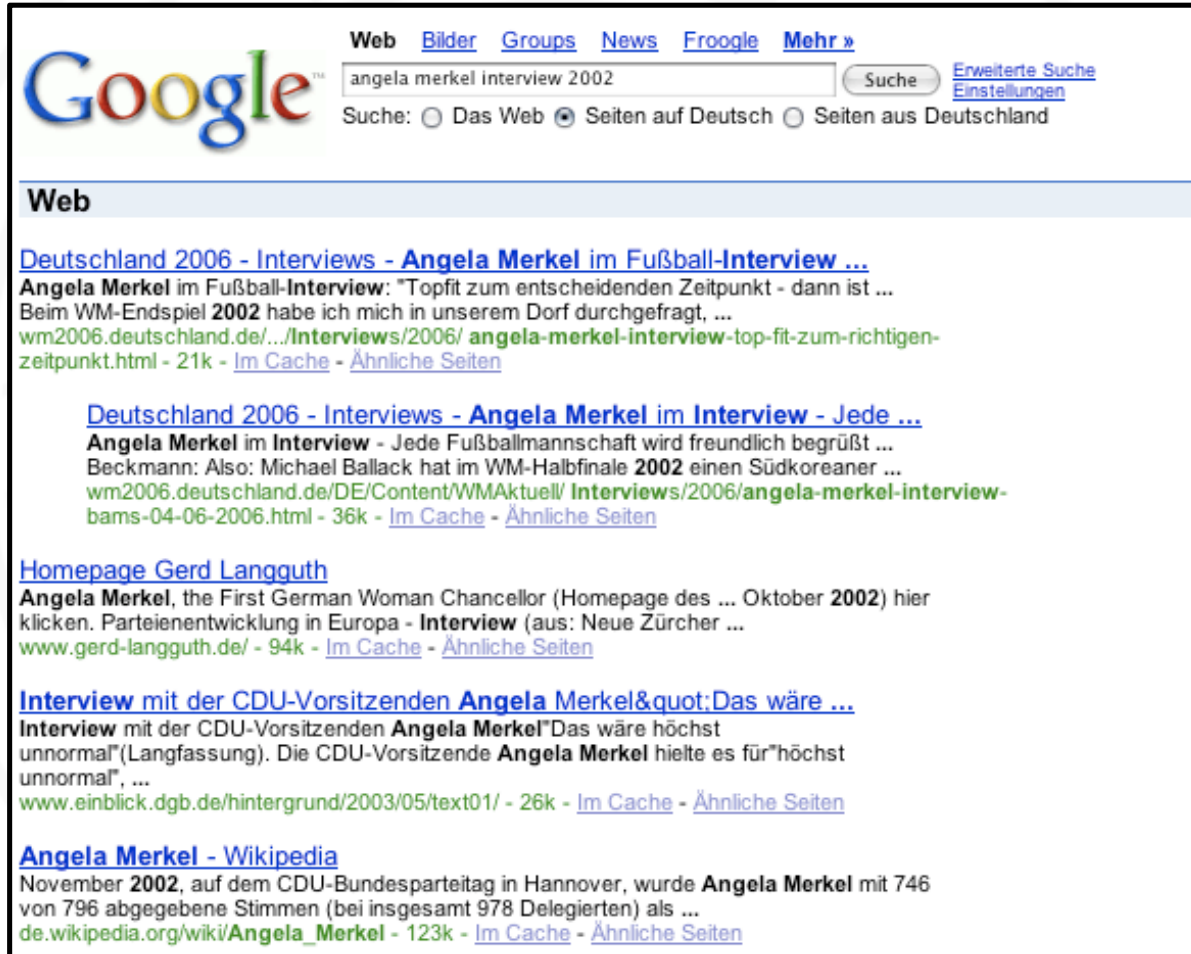


Motivation – Why time-travel search?

- Historical information needs, e.g.,
 - Web page mentioning *Web 2.0* in early 2004
 - Interview with *Angela Merkel* from 2002
 - Blogs praising Jürgen Klinsmann as a coach before FIFA World Cup 2006
- Temporal text mining applications can leverage time-travel search functionality



Motivation – Why time-travel search?



The screenshot shows a Google search interface. At the top, the Google logo is on the left, and navigation links for 'Web', 'Bilder', 'Groups', 'News', 'Froogle', and 'Mehr »' are on the right. A search bar contains the text 'angela merkel interview 2002'. To the right of the search bar is a 'Suche' button and links for 'Erweiterte Suche' and 'Einstellungen'. Below the search bar, there are radio buttons for 'Suche: Das Web', 'Seiten auf Deutsch' (which is selected), and 'Seiten aus Deutschland'. The search results are listed under the heading 'Web'. The first result is titled 'Deutschland 2006 - Interviews - Angela Merkel im Fußball-Interview ...' and includes a snippet: 'Angela Merkel im Fußball-Interview: "Topfit zum entscheidenden Zeitpunkt - dann ist ... Beim WM-Endspiel 2002 habe ich mich in unserem Dorf durchgefragt, ...'. The second result is titled 'Deutschland 2006 - Interviews - Angela Merkel im Interview - Jede ...' and includes a snippet: 'Angela Merkel im Interview - Jede Fußballmannschaft wird freundlich begrüßt ... Beckmann: Also: Michael Ballack hat im WM-Halbfinale 2002 einen Südkoreaner ...'. The third result is titled 'Homepage Gerd Langguth' and includes a snippet: 'Angela Merkel, the First German Woman Chancellor (Homepage des ... Oktober 2002) hier klicken. Parteienentwicklung in Europa - Interview (aus: Neue Zürcher ...'. The fourth result is titled 'Interview mit der CDU-Vorsitzenden Angela Merkel' and includes a snippet: 'Interview mit der CDU-Vorsitzenden Angela Merkel "Das wäre höchst unnormal" (Langfassung). Die CDU-Vorsitzende Angela Merkel hielt es für "höchst unnormal", ...'. The fifth result is titled 'Angela Merkel - Wikipedia' and includes a snippet: 'November 2002, auf dem CDU-Bundesparteitag in Hannover, wurde Angela Merkel mit 746 von 796 abgegebenen Stimmen (bei insgesamt 978 Delegierten) als ...'. Each result includes a link to 'Im Cache' and 'Ähnliche Seiten'.

ly 2004

002

coach

can
nality



Motivation – Why time-travel search?

- Historical information needs, e.g.,
 - Web page mentioning *Web 2.0* in early 2004
 - Interview with *Angela Merkel* from 2002
 - Blogs praising Jürgen Klinsmann as a coach before FIFA World Cup 2006
- Temporal text mining applications can leverage time-travel search functionality



Motivation – Challenges

- Large data volumes:
 - Internet Archive (~2 PBytes)
 - Revision history of English Wikipedia (~1 TByte)
- Existing relevance models do not handle time-varying collections statistics, e.g.,
 - growing collection size
 - *idf*-scores of increasingly popular terms
- Query response time: users are “spoiled” by Google etc., expect quick responses



Contributions

- Data, query, and relevance model for time-travel search
- Indexing infrastructure that includes
 - an adaptation of the inverted file index
 - approximate temporal coalescing as a highly effective technique to reduce index size
- Experimental evaluation of our approach on a real-world large-scale dataset



Outline

- Motivation
- Data & Query Model
- Relevance Model
- Temporal Text Indexing
- Query Processing
- Experimental Results
- Conclusions



Data & Query Model (I)

- Versioned text collection with each document potentially having many versions
- Document d is regarded as a sequence of its timestamped versions

$$d = \langle d^{t_i}, d^{t_{i+1}}, \dots \rangle$$

- Version d^{t_i} has validity time-interval $[t_i, t_{i+1})$



Data & Query Model (II)

- Time-travel query as a keyword query q that is enriched by a temporal context $[t_b, t_e]$

$$q^{[t_b, t_e]}$$

- Evaluated across all versions that existed at any point during the temporal context



Outline

- Motivation
- Data & Query Model
- Relevance Model
- Temporal Text Indexing
- Query Processing
- Experimental Results
- Conclusions



Relevance Model (I)

- Existing relevance models are time-agnostic and must therefore first be made time-aware
- OKAPI BM25 with time-dependent statistics

$$w(q^{[t_b, t_e]}, d^{t_i}) = \sum_{v \in q} w_{tf}(v, d^{t_i}) \cdot \overline{w_{idf}}(v, [t_b, t_e])$$

$$w_{tf}(v, d^{t_i}) = \frac{(k_1 + 1) \cdot tf(v, d^{t_i})}{k_1 \cdot \left((1 - b) + b \cdot \frac{dl(d^{t_i})}{avdl(t_i)} \right) + tf(v, d^{t_i})}$$

$$w_{idf}(v, t) = \log \frac{N(t) - df(v, t) + 0.5}{df(v, t) + 0.5}$$



Relevance Model (I)

- Existing relevance models are time-agnostic and must therefore first be made time-aware
- OKAPI BM25 with time-dependent statistics

$$w(q^{[t_b, t_e]}, d^{t_i}) = \sum_{v \in q} w_{tf}(v, d^{t_i}) \cdot \overline{w_{idf}}(v, [t_b, t_e])$$

$$w_{tf}(v, d^{t_i}) = \frac{(k_1 + 1) \cdot tf(v, d^{t_i})}{k_1 \cdot ((1 - b) + b \cdot \frac{dl(d^{t_i})}{avdl(t_i)}) + tf(v, d^{t_i})}$$

$$w_{idf}(v, t) = \log \frac{N(t) - df(v, t) + 0.5}{df(v, t) + 0.5}$$



Relevance Model (I)

- Existing relevance models are time-agnostic and must therefore first be made time-aware
- OKAPI BM25 with time-dependent statistics

$$w(q^{[t_b, t_e]}, d^{t_i}) = \sum_{v \in q} w_{tf}(v, d^{t_i}) \cdot \overline{w_{idf}}(v, [t_b, t_e])$$

$$w_{tf}(v, d^{t_i}) = \frac{(k_1 + 1) \cdot tf(v, d^{t_i})}{k_1 \cdot \left((1 - b) + b \cdot \frac{dl(d^{t_i})}{avdl(t_i)} \right) + tf(v, d^{t_i})}$$

$$w_{idf}(v, t) = \log \frac{N(t) - df(v, t) + 0.5}{df(v, t) + 0.5}$$



Relevance Model (I)

- Existing relevance models are time-agnostic and must therefore first be made time-aware
- OKAPI BM25 with time-dependent statistics

$$w(q^{[t_b, t_e]}, d^{t_i}) = \sum_{v \in q} w_{tf}(v, d^{t_i}) \cdot \overline{w_{idf}}(v, [t_b, t_e])$$

$$w_{tf}(v, d^{t_i}) = \frac{(k_1 + 1) \cdot tf(v, d^{t_i})}{k_1 \cdot ((1 - b) + b \cdot \frac{dl(d^{t_i})}{avdl(t_i)}) + tf(v, d^{t_i})}$$

$$w_{idf}(v, t) = \log \frac{N(t) - df(v, t) + 0.5}{df(v, t) + 0.5}$$



Relevance Model (I)

- Existing relevance models are time-agnostic and must therefore first be made time-aware
- OKAPI BM25 with time-dependent statistics

$$w(q^{[t_b, t_e]}, d^{t_i}) = \sum_{v \in q} w_{tf}(v, d^{t_i}) \cdot \overline{w_{idf}}(v, [t_b, t_e])$$

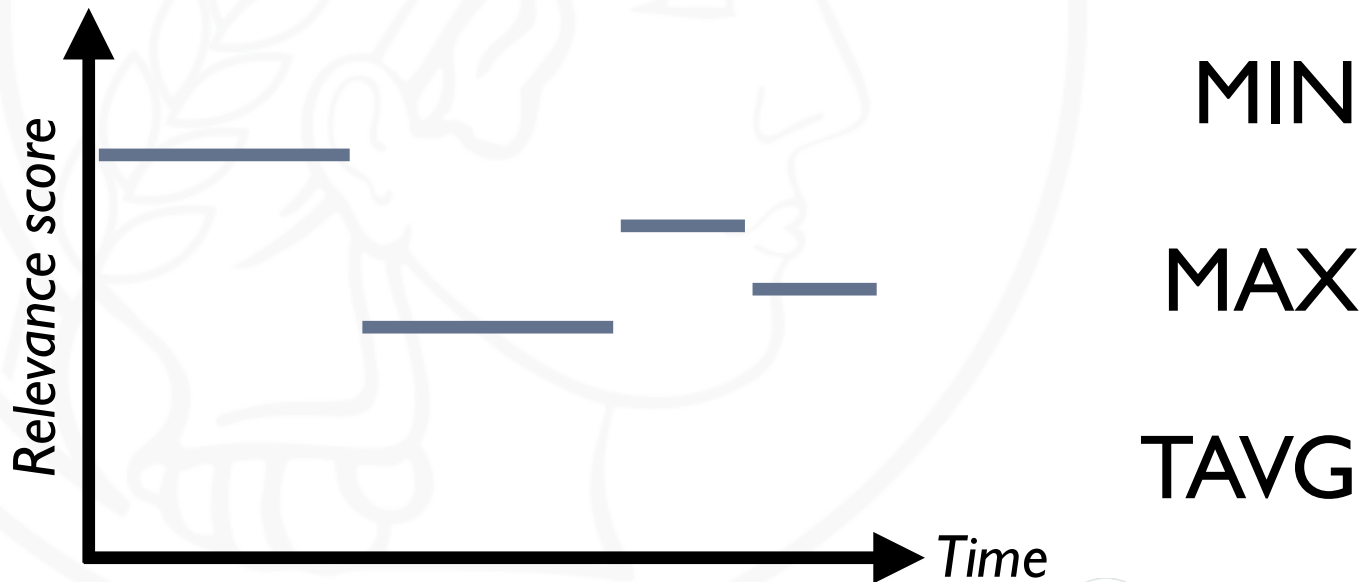
$$w_{tf}(v, d^{t_i}) = \frac{(k_1 + 1) \cdot tf(v, d^{t_i})}{k_1 \cdot \left((1 - b) + b \cdot \frac{dl(d^{t_i})}{avdl(t_i)} \right) + tf(v, d^{t_i})}$$

$$w_{idf}(v, t) = \log \frac{N(t) - df(v, t) + 0.5}{df(v, t) + 0.5}$$



Relevance Model (II)

- Potentially, many nearly-identical versions dilute the query result
- Aggregation of version-level relevance scores at the document level



Relevance Model (II)

- Potentially, many nearly-identical versions dilute the query result
- Aggregation of version-level relevance scores at the document level



Relevance Model (II)

- Potentially, many nearly-identical versions dilute the query result
- Aggregation of version-level relevance scores at the document level



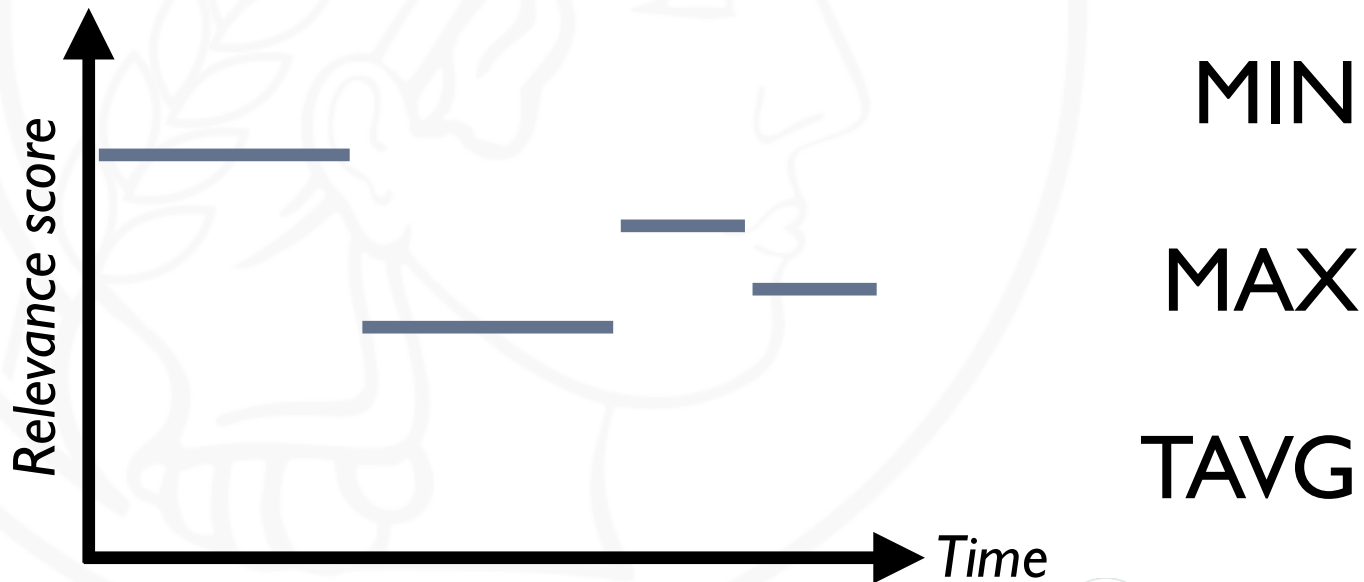
Relevance Model (II)

- Potentially, many nearly-identical versions dilute the query result
- Aggregation of version-level relevance scores at the document level



Relevance Model (II)

- Potentially, many nearly-identical versions dilute the query result
- Aggregation of version-level relevance scores at the document level



Relevance Model (II)

- Potentially, many nearly-identical versions dilute the query result
- Aggregation of version-level relevance scores at the document level



MIN

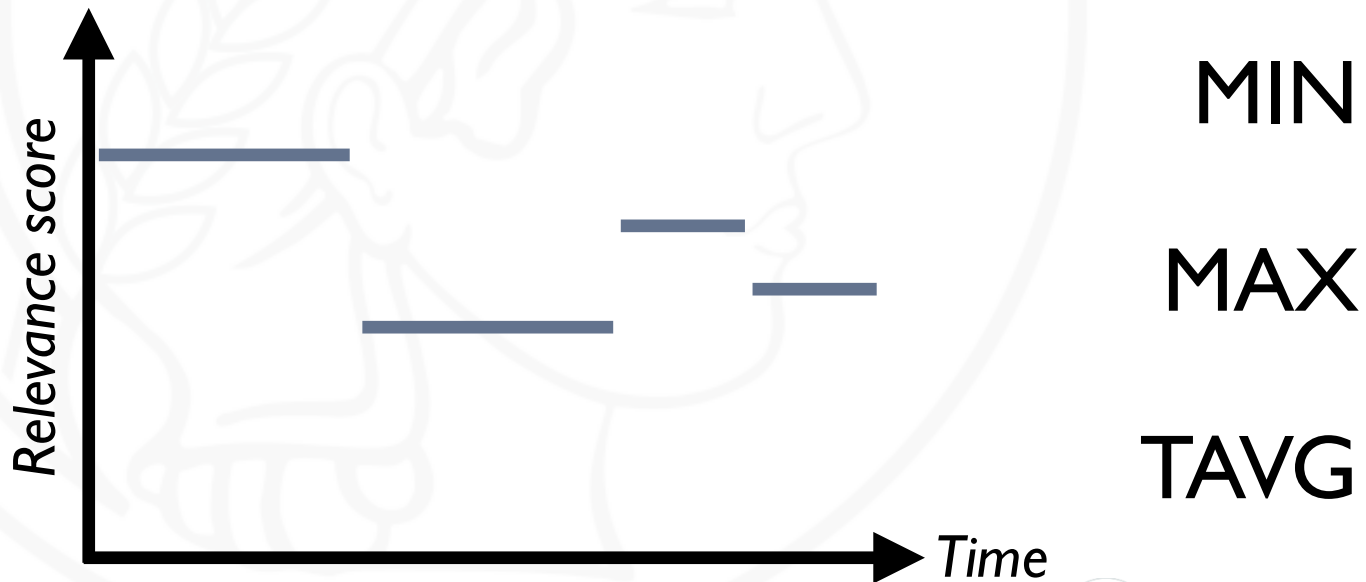
MAX

TAVG



Relevance Model (II)

- Potentially, many nearly-identical versions dilute the query result
- Aggregation of version-level relevance scores at the document level



Outline

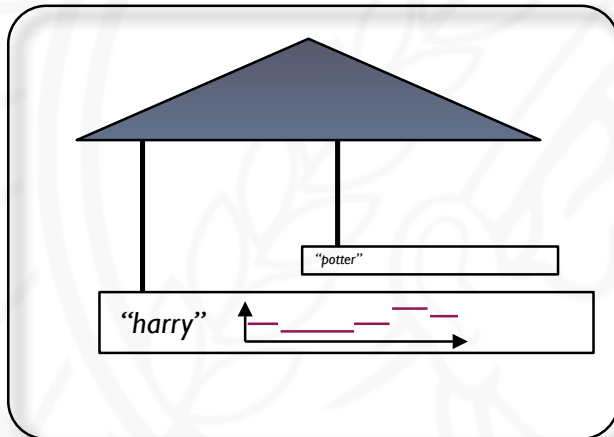
- Motivation
- Data & Query Model
- Relevance Model
- Temporal Text Indexing
- Query Processing
- Experimental Results
- Conclusions



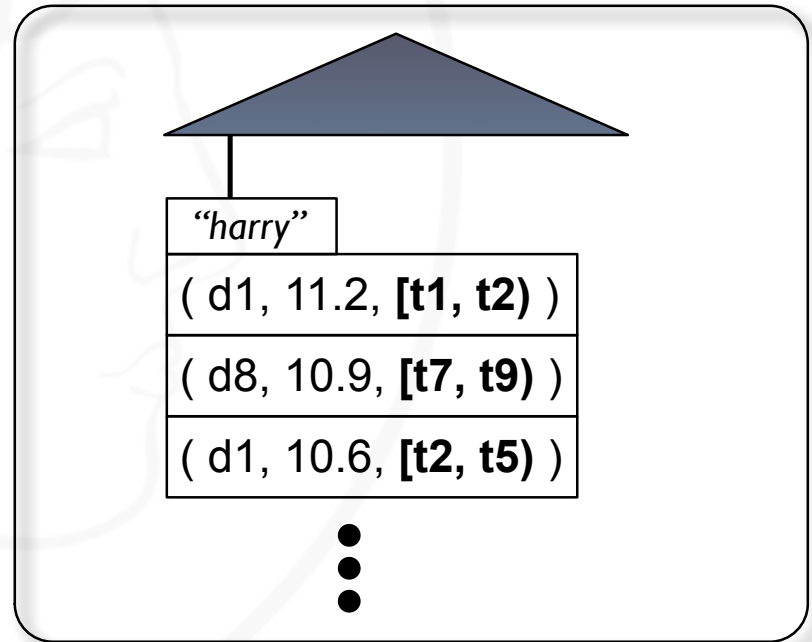
Temporal Text Indexing (I)

- IDF-scores maintained separately in B-Tree
- TF-scores kept in adapted inverted file index

IDF



TF



- One posting per term per version!



Temporal Text Indexing (II)

- Many changes between versions
 - are minor (e.g., corrected typos)
 - have no noticeable effect on the ranked result (e.g., 500 x “harry” vs. 501 x “harry”)
- Approximate temporal coalescing coalesces adjacent postings having similar scores



Temporal Text Indexing (III)

- Approximate temporal coalescing finds a piecewise-constant representation of

$$\langle (t_0, w_{tf}(v, d^{t_0})), \dots, (t_N, w_{tf}(v, d^{t_N})) \rangle$$

- Maximal relative error per segment is upper bounded by a threshold ϵ
- Optimal solution (minimal # of segments) by dynamic programming in time $O(N^3)$
- Approximate solution computable in time $O(N)$ good enough in practice



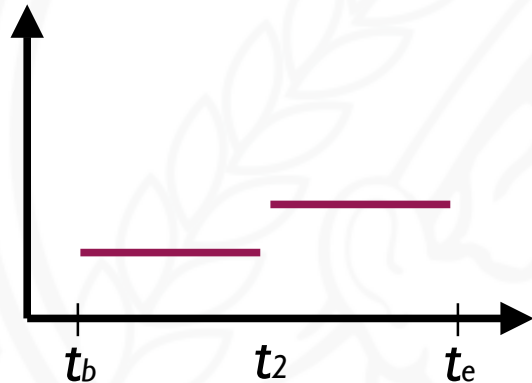
Outline

- Motivation
- Data & Query Model
- Relevance Model
- Temporal Text Indexing
- **Query Processing**
- **Experimental Results**
- **Conclusions**



Query Processing

- Bookkeeping of candidates extended to maintain time series instead of simple scores



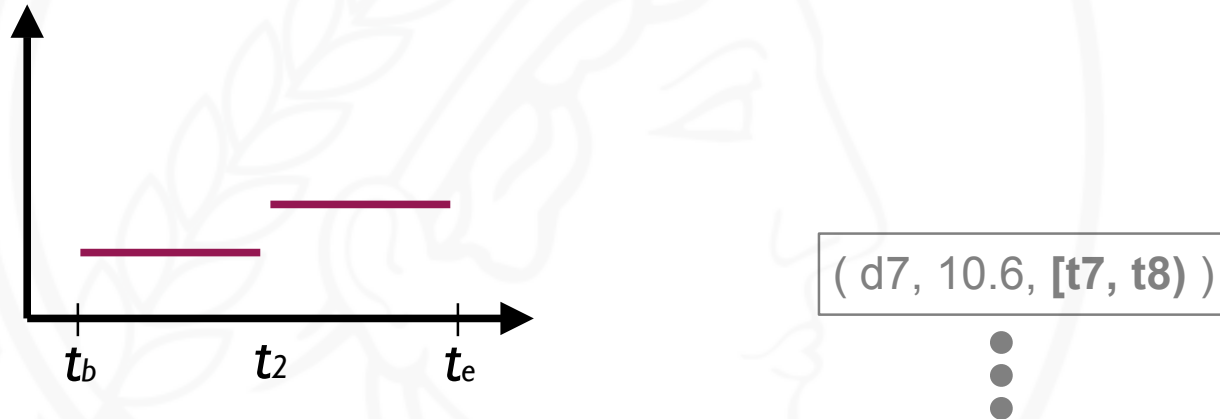
(d1, 11.2, [t1, t3))
(d7, 10.6, [t7, t8))
⋮

- Document-level relevance aggregations ready for efficient top- k query processing



Query Processing

- Bookkeeping of candidates extended to maintain time series instead of simple scores

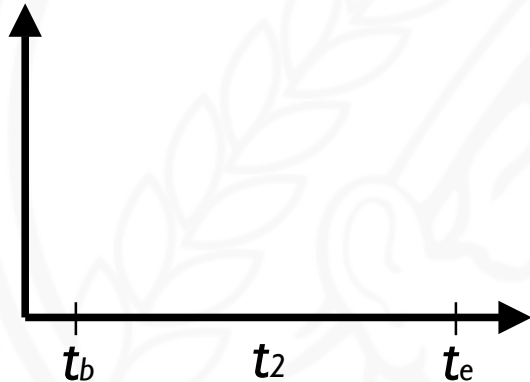


- Document-level relevance aggregations ready for efficient top- k query processing



Query Processing

- Bookkeeping of candidates extended to maintain time series instead of simple scores



(d7, 10.6, [t7, t8))

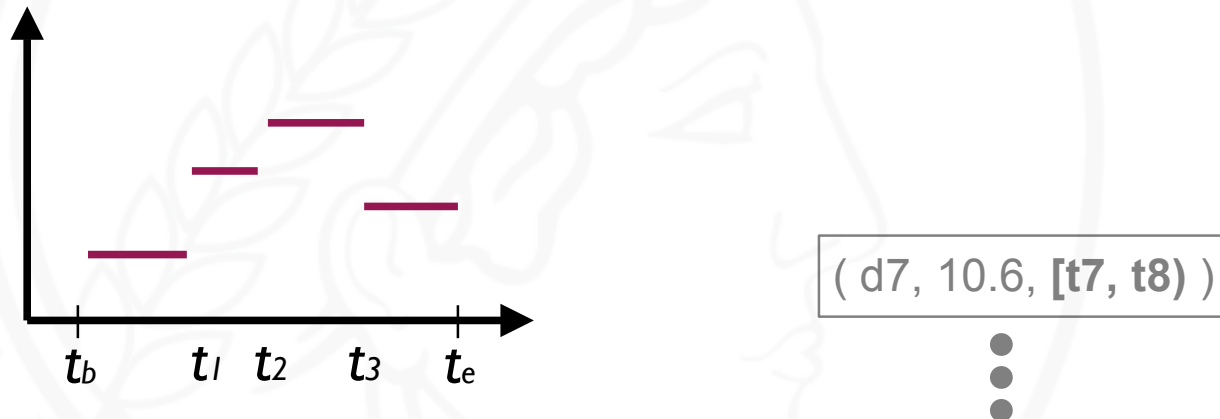


- Document-level relevance aggregations ready for efficient top- k query processing



Query Processing

- Bookkeeping of candidates extended to maintain time series instead of simple scores



- Document-level relevance aggregations ready for efficient top- k query processing



Outline

- Motivation
- Data & Query Model
- Relevance Model
- Temporal Text Indexing
- Query Processing
- **Experimental Results**
- **Conclusions**

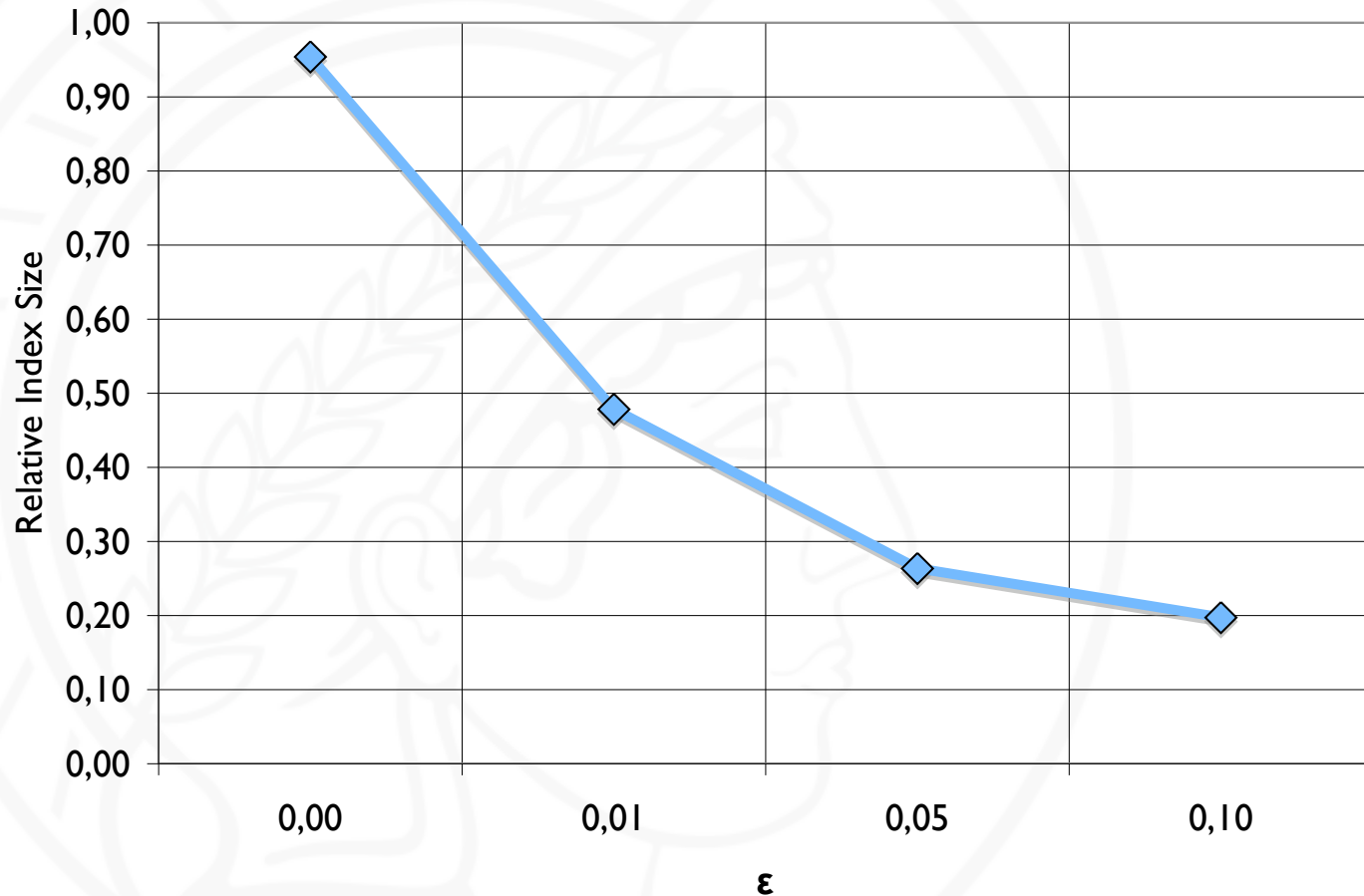


Experimental Setup

- Corpus: Revision history of English Wikipedia
 - 892,255 documents (encyclopedia articles)
 - 2,795,383 versions (20% sample of the corpus)
- Queries: 45 most popular queries, e.g., “*french revolution*”, “*american idol*”, “*da vinci code*” from AOL query log with result clicks on wikipedia.org each combined with 6 temporal contexts
- Prototype implementation: Java, Oracle 10g, 4 CPU 64bit machine, 16GB RAM



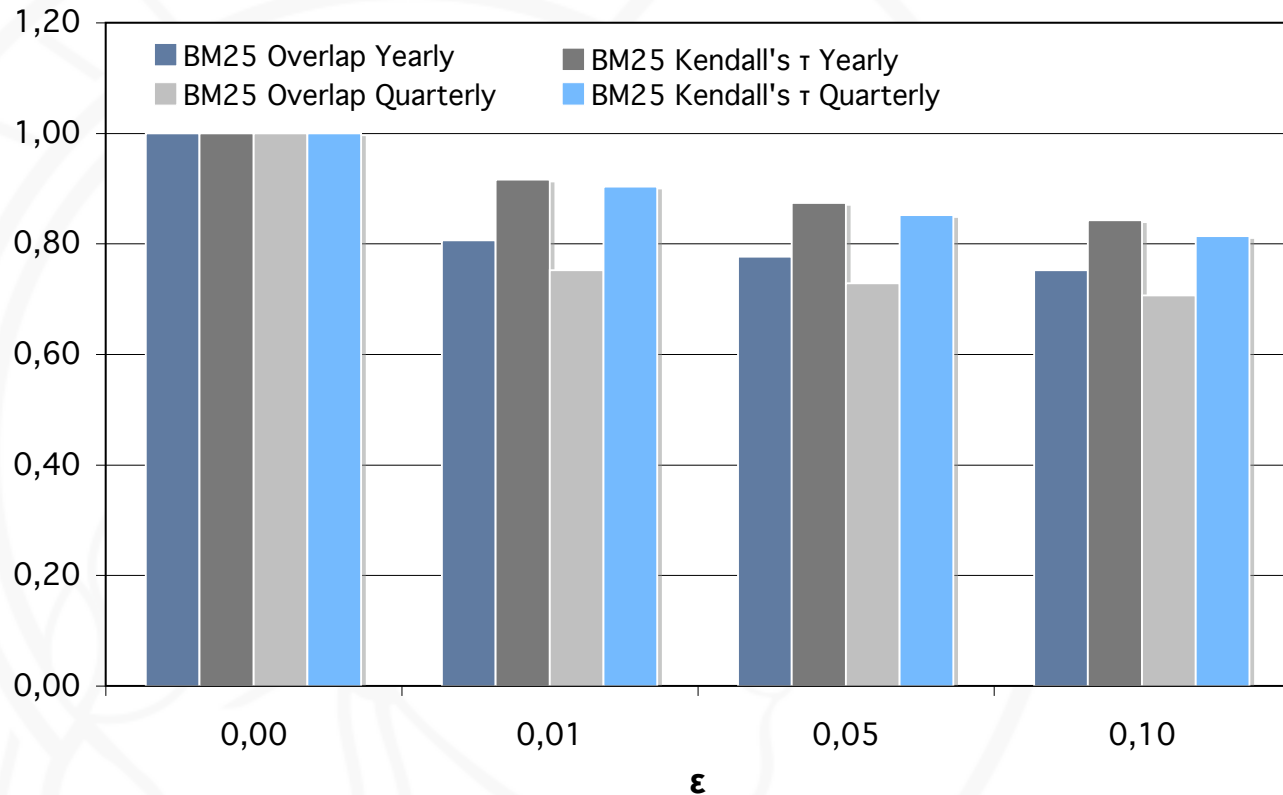
Experimental Results – Index Size



- Original index has 1,244,168,879 postings



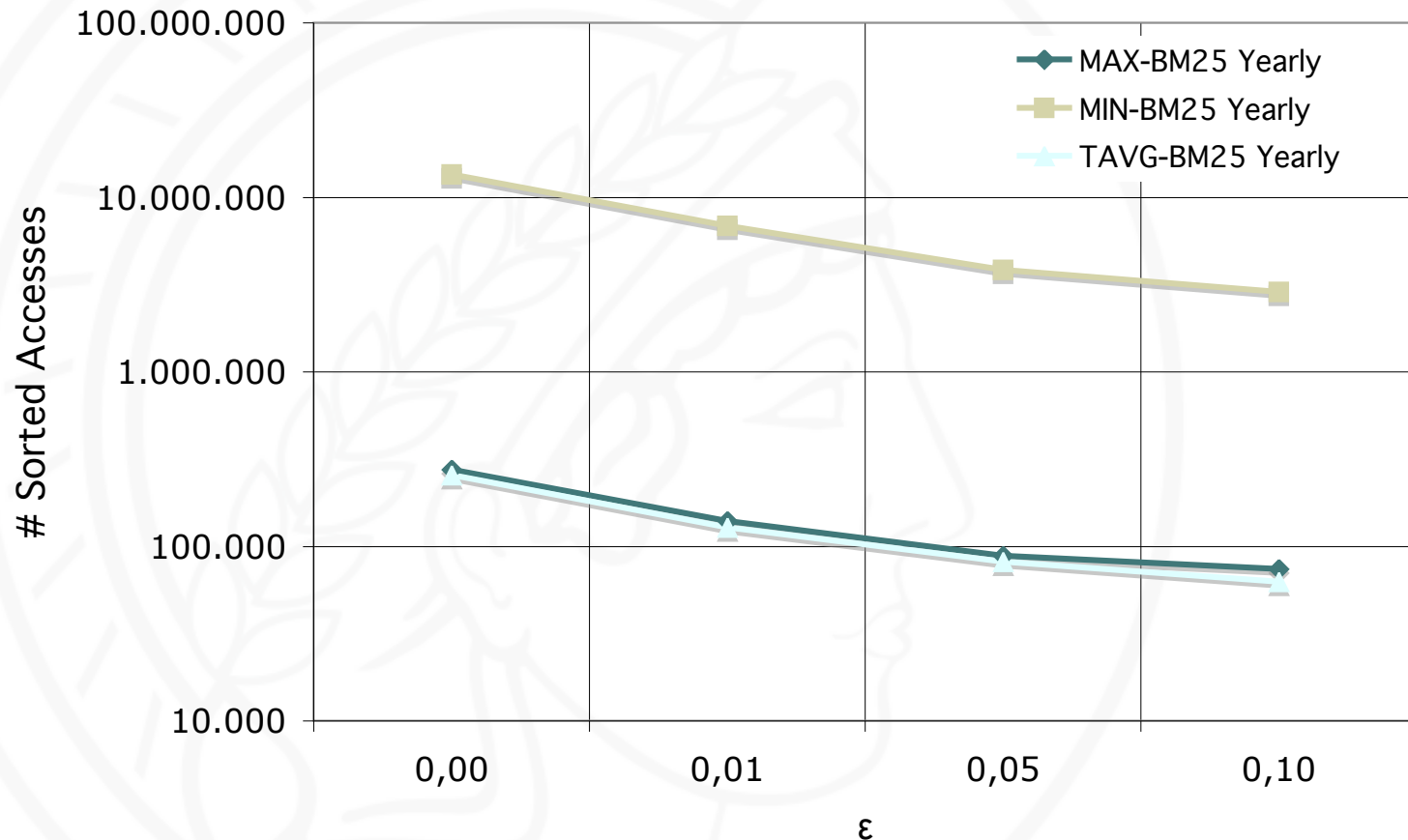
Experimental Results – Precision



- Top-100 relative precision measured using
 - Overlap: fraction of contained original results
 - Kendall's τ computed on overlap



Experimental Results – Performance



- Performance measured as # sorted accesses to process query batch



Outline

- Motivation
- Data & Query Model
- Relevance Model
- Temporal Text Indexing
- Query Processing
- Experimental Results
- **Conclusions**



Conclusions

- Data, query, and relevance model for time-travel search over versioned text collections
- Indexing infrastructure including
 - an adapted inverted file index
 - approximate temporal coalescing as a highly effective technique to reduce index size
- Experiments on large-scale real-world dataset
 - Index size and query-processing time reduced to 20% without sacrificing result quality





Thanks for your attention!

Any questions?

