Temporal Information Retrieval

SIGIR'2016 Tutorial

17 July 2016, Pisa, Italy

Speakers

- Dr. Avishek Anand
  - Senior Researcher
  - L3S Research Center
  - Hannover, Germany

- Dr. Nattiya Kanhabua
  - Assistant Professor
  - Aalborg University
  - Aalborg, Denmark

Web Science @ L3S

- Computer Science and interdisciplinary research on all aspects of the Web
  - Internet: Communication and Networks
  - Information: Accessing information and knowledge on and through the Web
  - Community & society: Supporting communities and groups on the Web, for research, education, production and entertainment

Real-time data processing for finance predictions

Cross-media analysis and interpretation

Retrieval, Exploration and Analytics for Web Archives (ERC Advanced Grant)

ForgetIT: Concise Preservation via Managed Forgetting

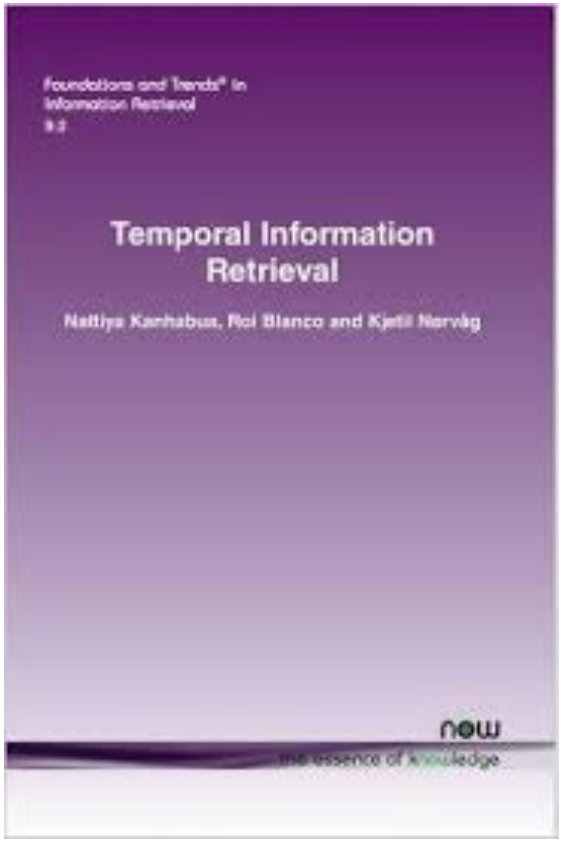
CUbRIK: Searching by computers and humans

Schedule

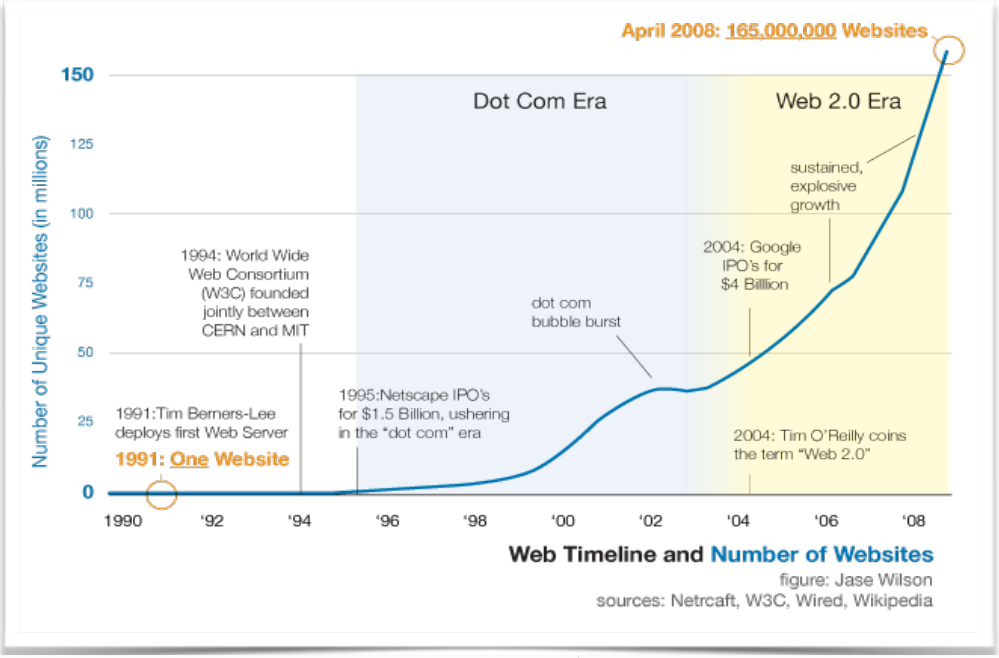
- 9:00 – 10:30 Part I (1.5 hour)
  - Introduction to Temporal IR (*20 minutes*)
  - Temporal Indexing and Query Processing (*35 minutes*)
  - Time-aware Retrieval and Ranking (*35 minutes*)
- 10:30 – 11:00 Coffee break
- 11:00 – 12:30 Part II (1.5 hour)
  - Temporal Query Analysis (*40 minutes*)
  - Applications of Temporal IR (*40 minutes*)
  - Conclusions and Future Directions (*10 minutes*)

Additional Resource

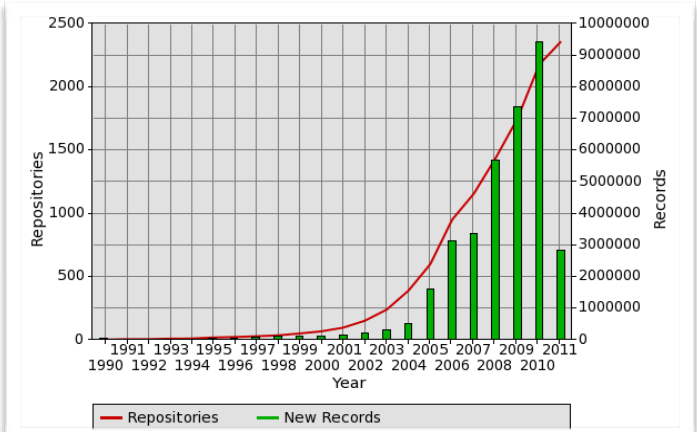
- Recent book: **Temporal Information Retrieval**
  - Authors: N. Kanhabua, R. Blanco, and K. Nørvåg
  - Foundations and Trends® in Information Retrieval
  - Volume 9, Issue 2, pp 91-208, 2015
  - Freely available: https://goo.gl/DUiw5R
  - Download from the authors' home pages

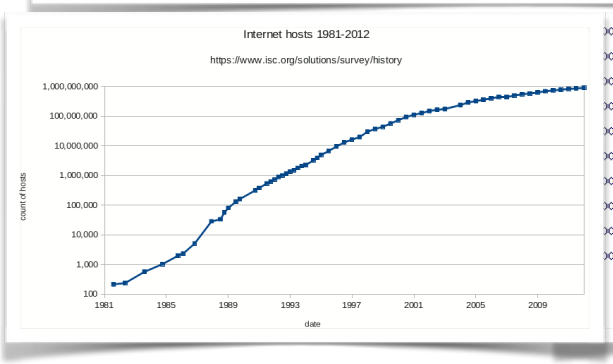


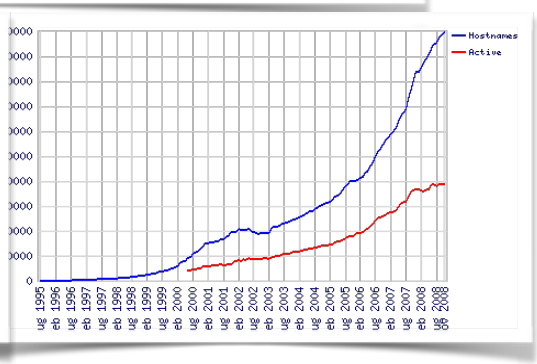
The Opening Line Syndrome

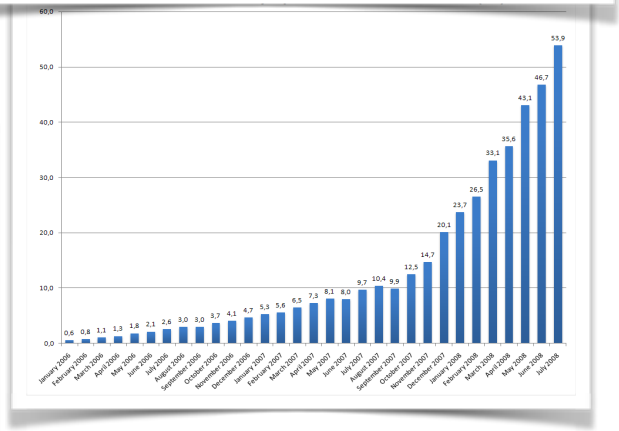


[labeled for non-commercial re-use]









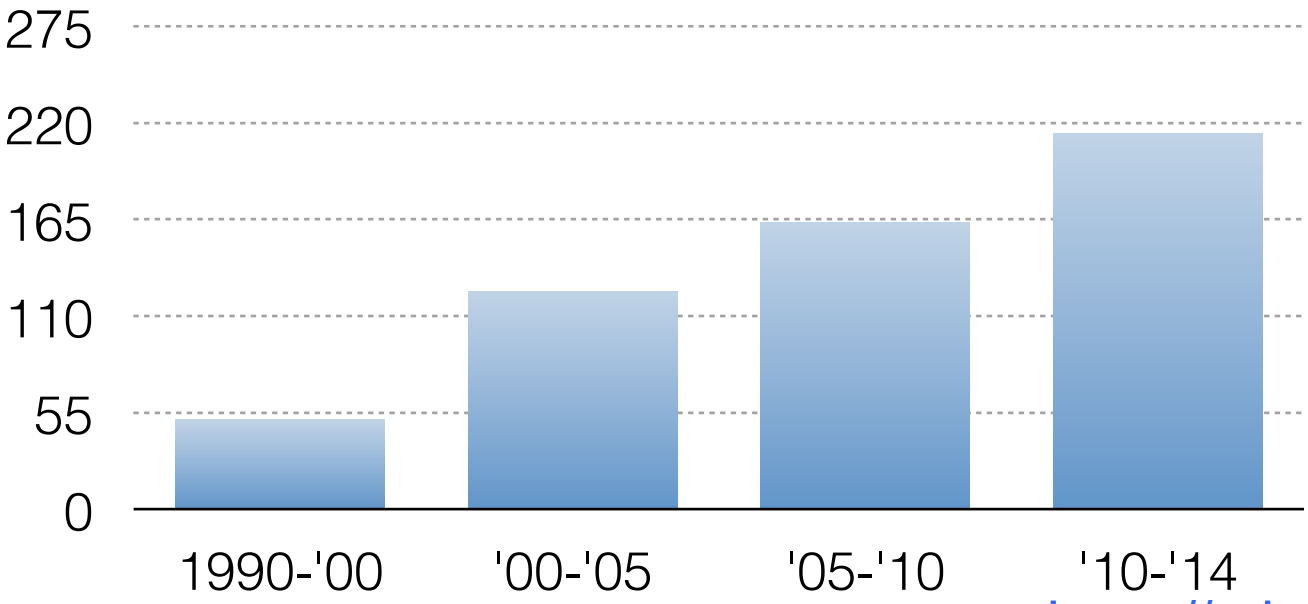
Evidence II

"The Web is * growing"

continuously constantly rapidly

still a...organism ever

explosively dramatically Tremendously



https://scholar.google.com/

Evidence III - Dilbert

The Web is definitely growing

Web Dynamics

How does the Web change and evolve?

- Change in persistent documents
- Change in "real-time" content streams
- Change in Web graphs

What are its applications?

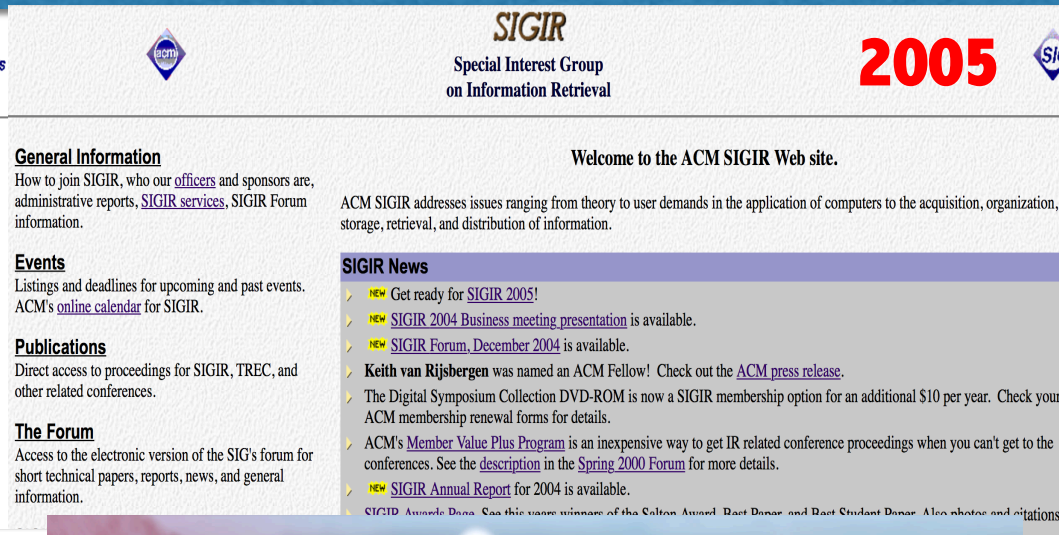
- Change detection useful is designing better crawlers
- Improving freshness in results
- Improving retrieval quality taking time into account
- Resource planning for the future

Research Highlights

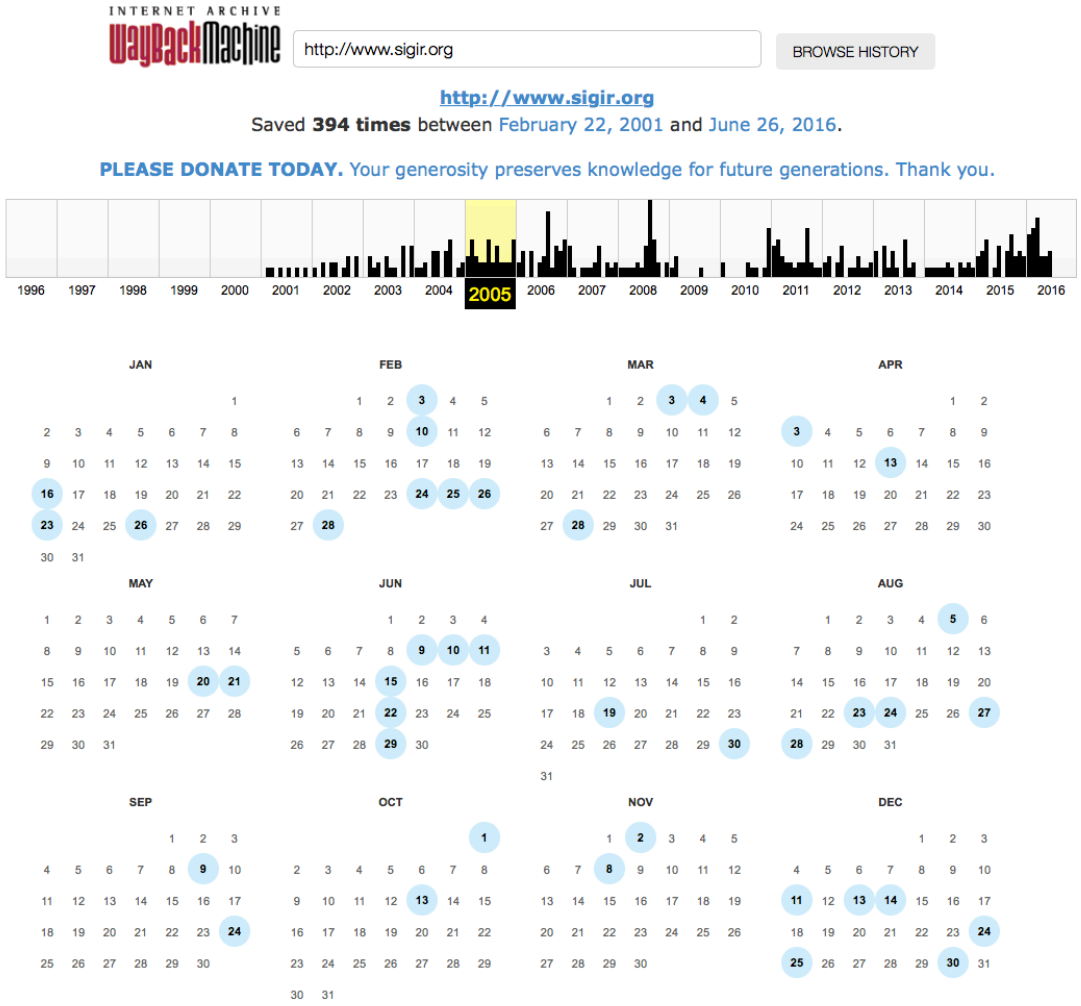
- Studies on dynamics of large scale Web crawls

- Only 20% web pages available today will be accessible after a year
- 1.3% new pages are encountered at every new crawl
- Term-level changes: Popular pages change frequently but not much
- Changes in pages depend on related pages
- Page size doubles every four years, Web gets older slowly (2 years old in 2020)

Web Archives

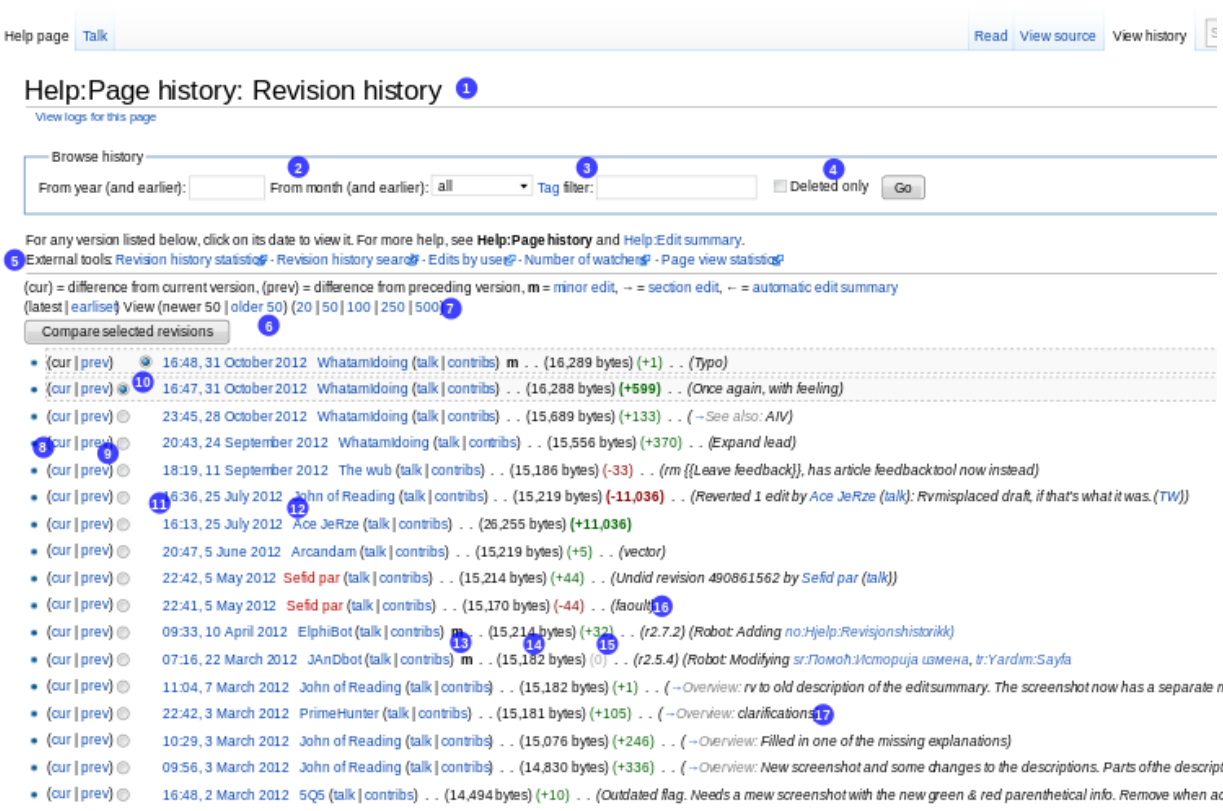


Web Archives – Limited Search



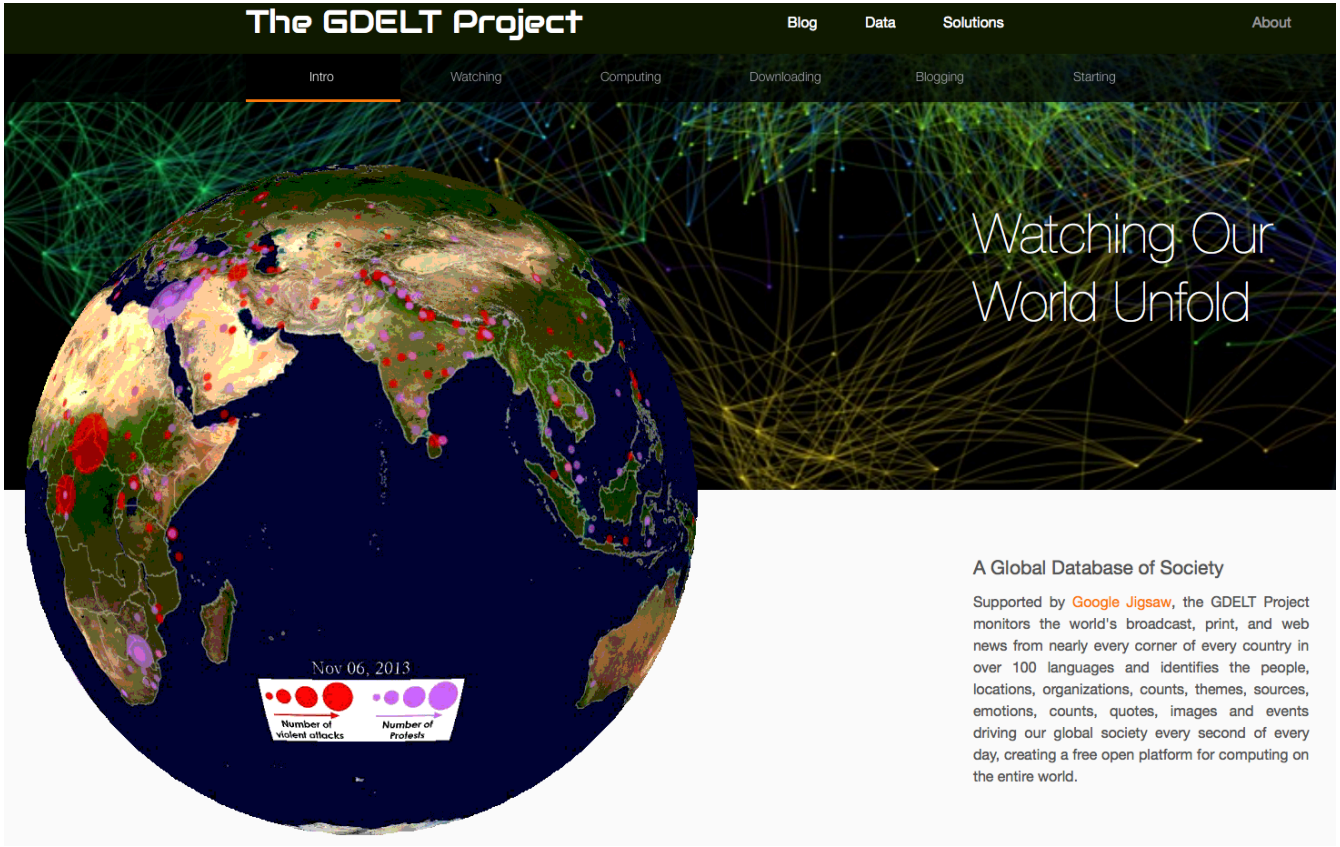
Only lookups allowed, historical queries not supported

Temporal Collections – Versioned Collections



Evolution of knowledge, foundation for many KBs, used in many linking tasks

Temporal Collections – Non versioned



The New York Times

Breaking news queries , event repositories, etc.

Categorization of Content Change

Content Change		
	Non-version	**Version**
Dynamic	Social medias (Twitter, Facebook, Youtube, etc.) News feeds Emails Blogs E-commerce sites	Wikipedia
Static	News archives, e.g., NY Times (20 years), the Times (150 years), and Zeit (17 years) Persistent Web documents Twitter archives	Web archive collections by Internet Archive, Internet Memory Foundation, or British Library Wikipedia history

- Implication:
  - Crawling, Indexing, Ranking, Query Analysis

Motivation

Temporal Collections have great value

- Historical information needs
- Captures evolution

The history NATO interventions in the last 100 years

Tunis Asks NATO Intervention to End French-Algerian Conflict

TUNIS, Feb. 27. (AP)—Tunisian President Habib Bourguiba today called on the NATO powers to intervene to end the conflict between France and the Algerian rebels.

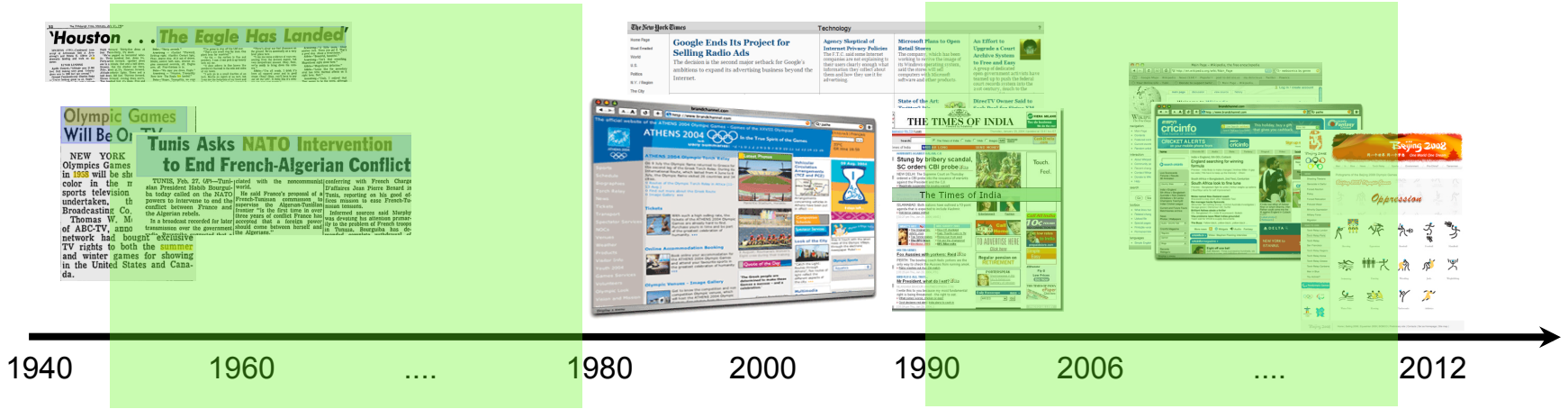
ciated with the noncommunist world.

He said France's proposal of a French-Tunisian commission to supervise the Algerian-Tunisian frontier "is the first time in over three years of conflict France has

conferring with French Charge D'affaires Jean Pierre Benard in Tunis, reporting on his good offices mission to ease French-Tunisian tensions.

Informed sources said Murphy

Time-Travel Text Search



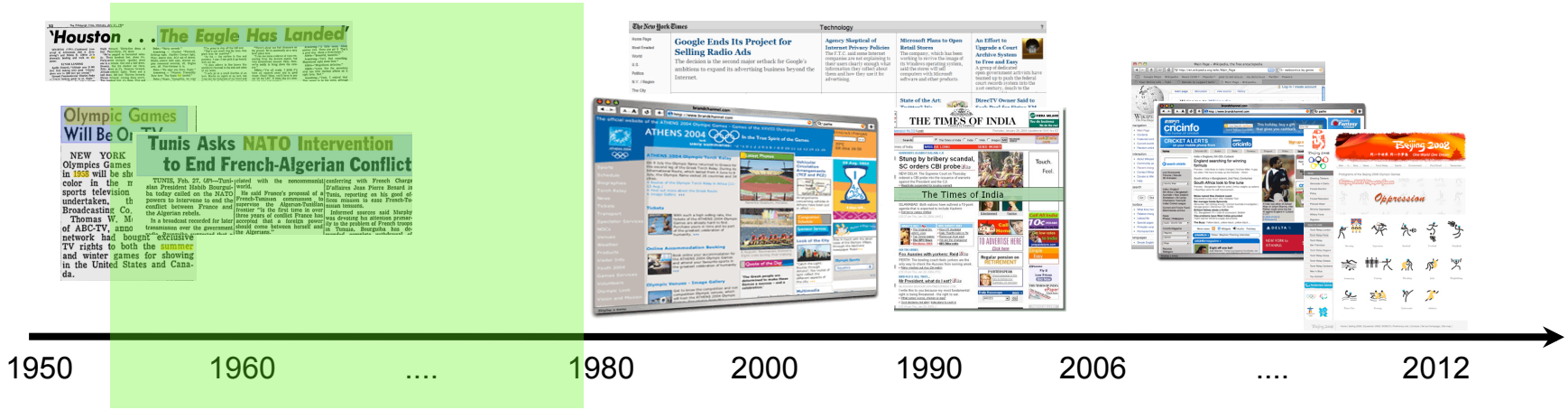
nato intervention @ 1950-1980

nato intervention @ 1990-2012

Tunis Asks **Nato Intervention** To End French-algerian...
Spokane Daily Chronicle - Feb 27, 1958
Habib Bourguiba today called on the **NATO** powers to **intervene** to end the conflict between France and the Algerian rebels. In a broadcast recorded for later ...
Nato .Intervention Ashed In French,... Reading Eagle

Search workloads interesting and important

Time-Travel Text Search

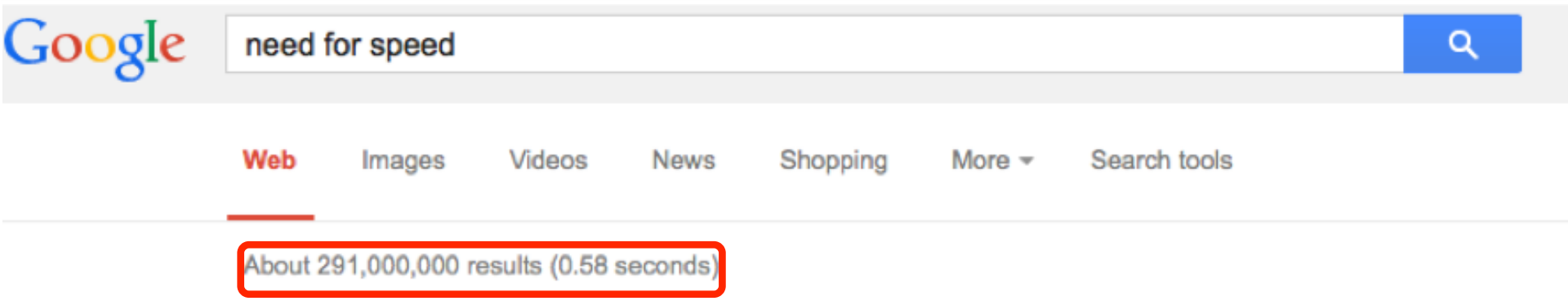


nato intervention @ 1950-1980

Tunis Asks **Nato Intervention** To End French-algerian...
Spokane Daily Chronicle - Feb 27, 1958
Habib Bourguiba today called on the **NATO** powers to **intervene** to end the conflict between France and the Algerian rebels. In a broadcast recorded for later ...
Nato .Intervention Ashed In French,... Reading Eagle

Search as a primitive operation

Challenges in Indexing



- How do we construct indexes for efficient temporal search ?
  - Sub-second retrieval
  - Interactive search
  - Search as a primitive for post-processing
- How do we ensure a small index footprint ?
- How do we maintain indexes for evolving text collections ?

Temporal Issues in Ranking

- Dynamic content also entails dynamic interests of users

- Breaking news queries :
  - "What is the EU ?"
  - "Live updates from France vs Portugal"
  - "presidential debate"
- Temporally ambiguous
  - Prism

- These are Implicit intents with interest in recent information or recency

Temporal Information Need

- Information needs that have a temporal dimension

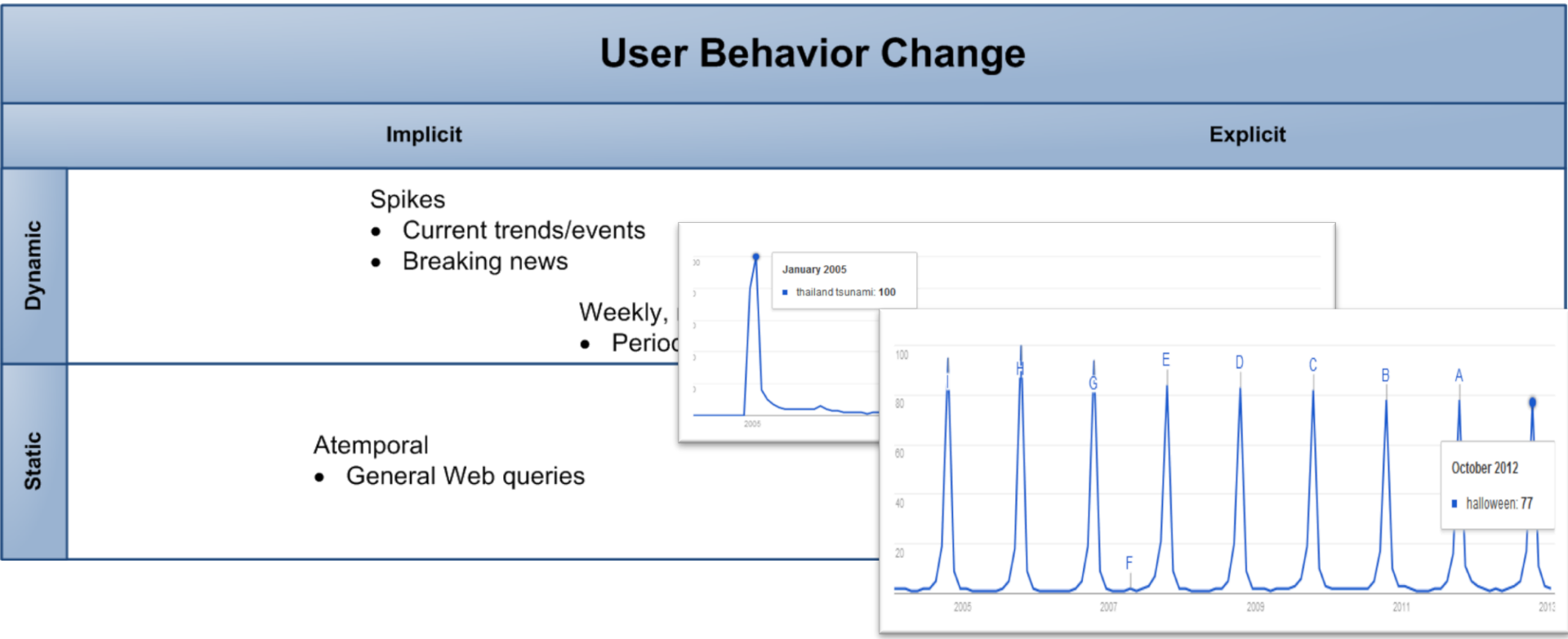
*FIFA World Cup tournaments **of the 1990's***
*Movies that won an Academy Award **in 2007***
*Crusades **of the 12th century***
*London Summer Olympics **2012***

- Queries that contain a temporal expression (e.g., in 1998)
  - indicate an underlying temporal information need
  - account for 1.5% of general web queries
  - are more common for specific domains (e.g., news or sports) and/or specific user groups (e.g., historians or journalists)

- **But:** Not well-supported by existing retrieval models

Categorization of Web Search Queries

- Implications:
  - Query analysis and ranking



http://www.google.com/insights/search

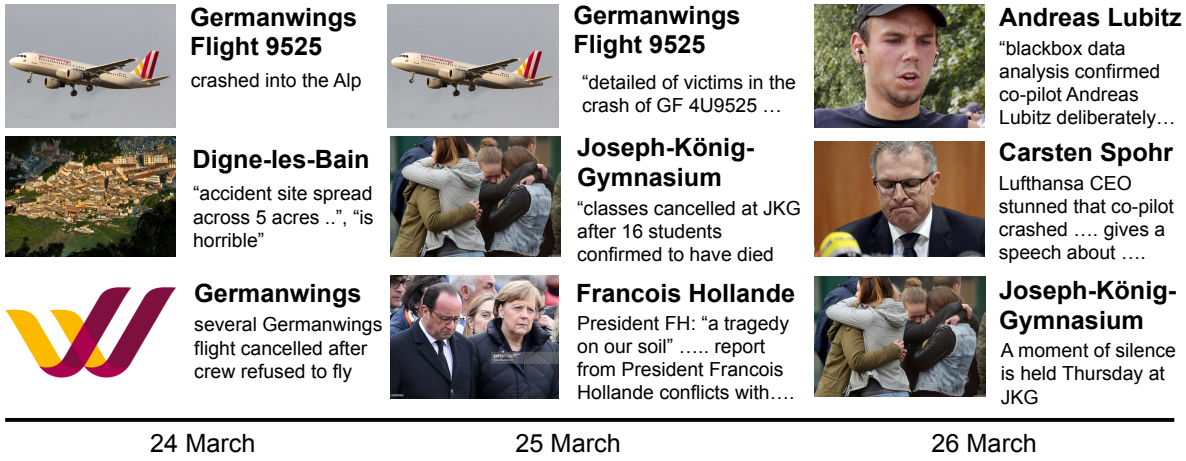
Historical Information Need

- Search on temporal collections, ranking documents at different time points

- Example topics
  - **Evolution** of the underground Art movement in London
  - History of Donald Trump

- Starting point for further exploration
- Needs a broad overview of the topic in the topical and temporal dimension

Challenges in Ranking



- How do we modify result ranking to reflect emerging user interests?

- How do we incorporate temporal query intents?
  - Interest in freshness and recent results
  - Temporal intents
  - Historical intents

Challenges in Query Modelling

- How do we incorporate temporal cues for …
  - Query suggestions
  - Auto-completions
  - Expansions for better retrieval

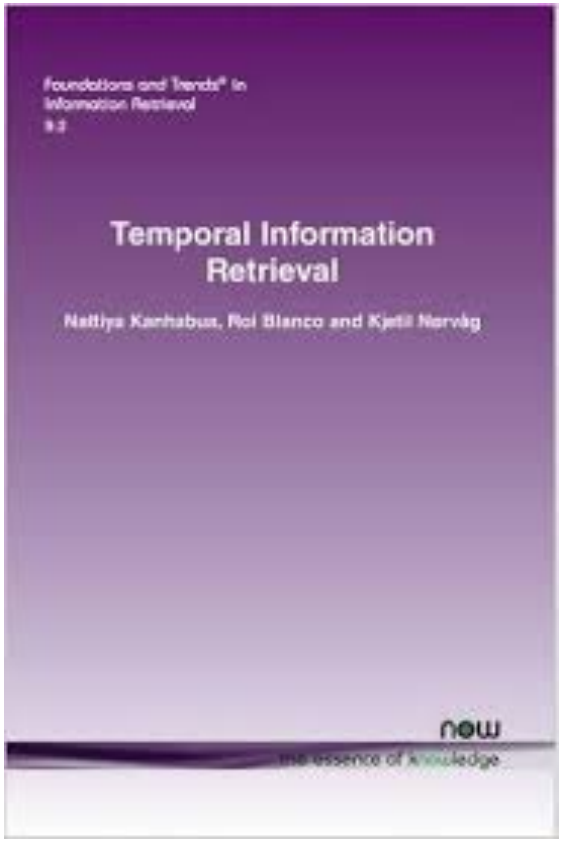
Other Issues in Temporal IR

- How do you extract temporal information?
  - Determining document creation time – document dating
  - Estimating document focus time
- Entity and event evolution detection
  - Entity and event based ranking
  - Name evolutions
- Temporal clustering and mining
- Crawling and Caching

Further readings

Additional Resource

- Book: **Temporal Information Retrieval**
  - Authors: N. Kanhabua, R. Blanco, and K. Nørvåg
  - Foundations and Trends® in Information Retrieval
  - Volume 9, Issue 2, pp 91-208, 2015
  - Freely available: https://goo.gl/DUiw5R
  - Download from the authors' home pages



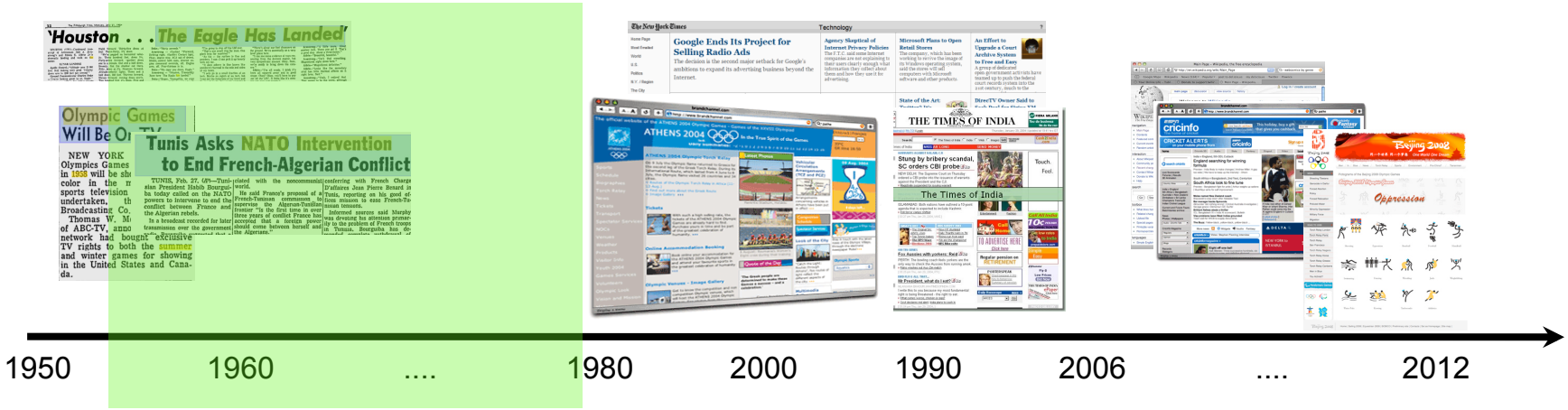
Related Community Efforts

- **Workshops**
  - Temporal Web Analytics Workshop (TempWeb)
  - Temporal, Social, Spatially-aware Information Access (TAIA)
  - Co-located with major conferences, such as, SIGIR and WWW

- **Competitions and Challenges**
  - NTCIR Temporalia: Temporal Diversification, Query Classification
  - TREC Temporal Summarization
  - WebScience'2016 Hackathon: L3S Research Center and the Internet Archive
  - ArchivesUnleashed Hackathons

Temporal Indexing

Temporal Indexing Setup



nato intervention @ 1950-1980

- **Temporal Collection** – Web archive pages, news archives, Wikipedia versions, emails, etc.

- **Temporal Query** – Interested in a time point or interval in the past
- **Results** – Documents/versions satisfying constraints in the query

Collection Model

- Discrete notion of time (i.e., integers as timestamps)
- Special timestamp ***now*** always points to the current time
- Document **d** is a sequence of time-stamped versions

$$\mathbf{d} = \langle \mathbf{d}^{t_1}, \mathbf{d}^{t_2}, \ldots \rangle$$

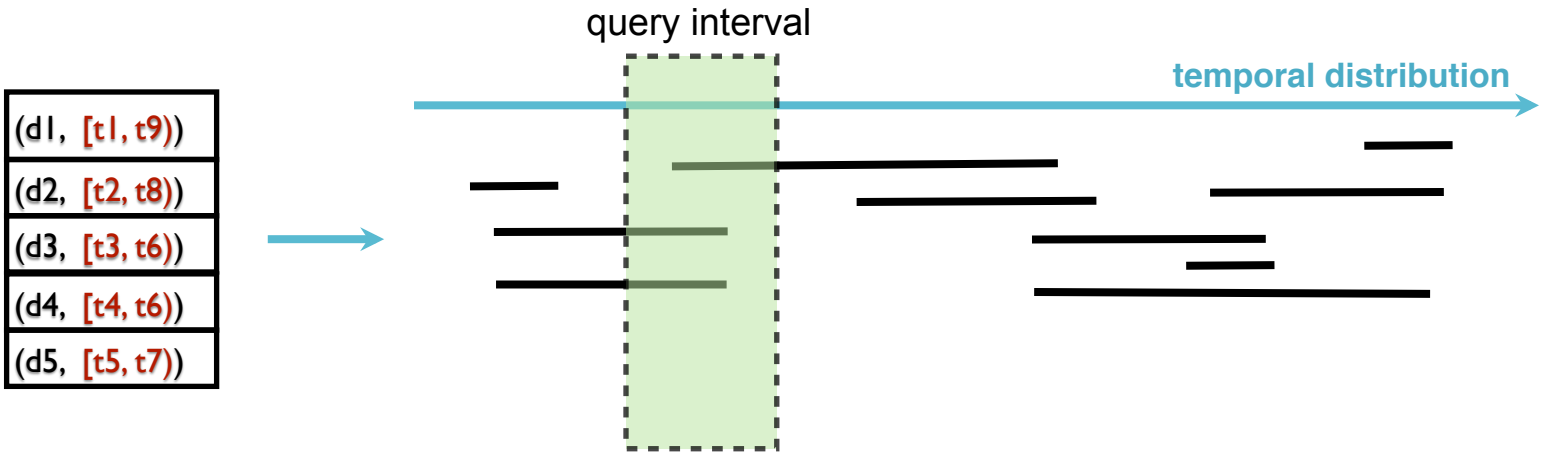
- Document deletion results in tombstone version $\perp$
- Valid-time interval of document version $\mathbf{d}^{t_i}$

$$val(\mathbf{d}^{t_i}) = \begin{cases} [t_i, t_{i+1}) & : \quad \mathbf{d}^{t_{i+1}} \in \mathbf{d} \\ [t_i, now) & : \quad otherwise \end{cases}$$

- State of the document collection **D** during time interval

$$\mathbf{D}^{[t_b, t_e]} = \bigcup_{\mathbf{d} \in \mathbf{D}} \left\{ \mathbf{d}^{t_i} \in \mathbf{d} \;\middle|\; \wedge \begin{array}{l} \mathbf{d}^{t_i} \neq \perp \\ [t_b, t_e] \cap val(\mathbf{d}^{t_i}) \neq \emptyset \end{array} \right\}$$

Data and Query Model



Discrete notion of time

Data Model: Each document is associated with a valid time interval

Query Model: Queries are associated with a time interval [tb, te]

Point in time queries: when begin time = end time

Time interval queries

Problem Statement

- Given a versioned collection of documents with valid time intervals
- Time-travel text queries

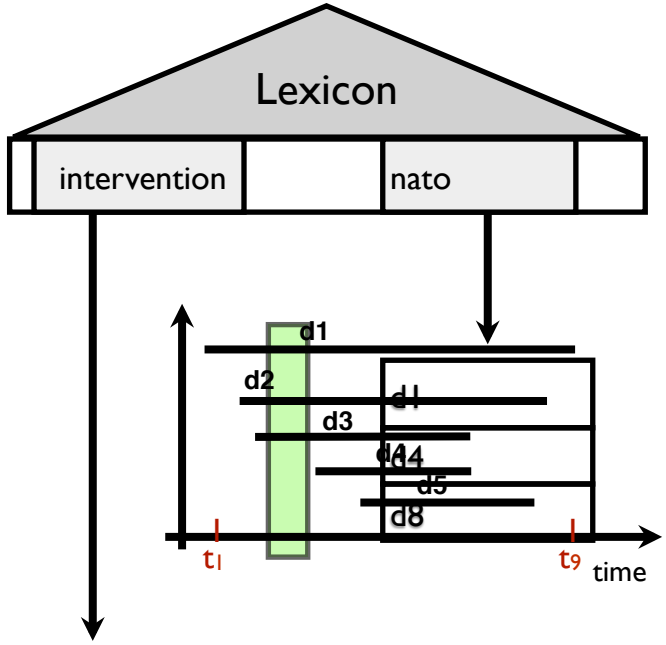
nato intervention @ *2001 - 2004*

- We want to retrieve documents containing terms "nato" ,"intervention" and valid between *2001 - 2004*

How do we efficiently retrieve these documents ?

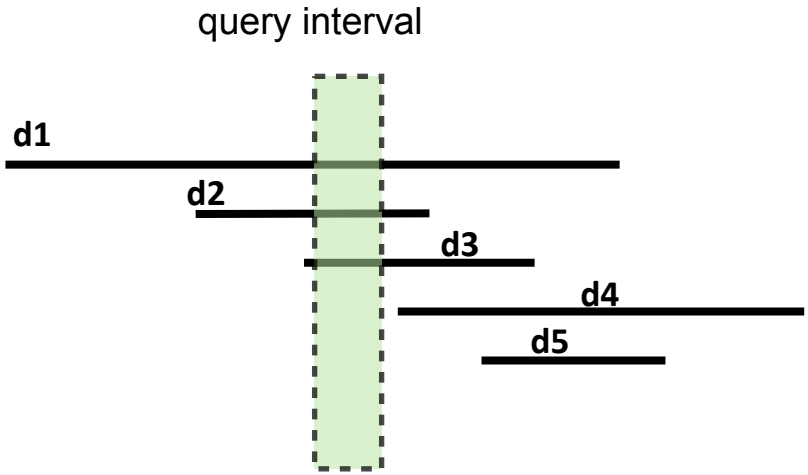
Naiive Solution – Enriched Inverted Index

- Inverted index processes keyword queries
- Intersection of posting lists for processing queries
- Versions have valid time intervals
- Augment postings with valid time intervals
- Post filtering after standard query processing



Challenges in Indexing Time

- We would want to avoid unwanted or wasted access to posting lists
- Typically only access those postings that are relevant or a few more (bounded loss)
- Dealing with time points easy, akin to range queries (sorting acc to begin time and range search)



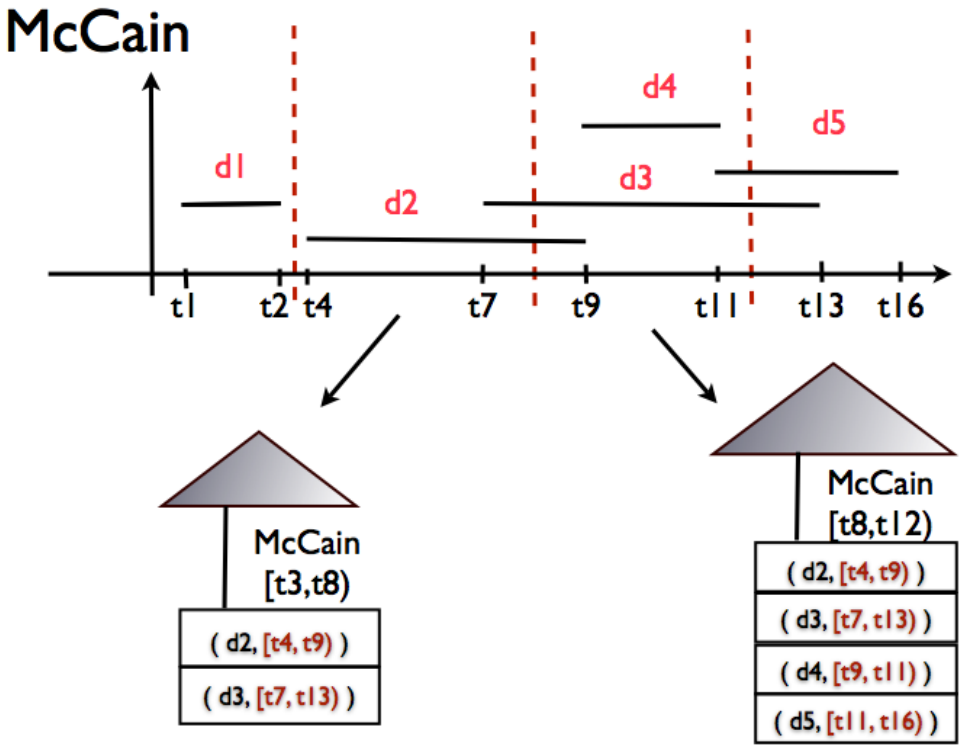
Major Research Directions

- Index Partitioning
  - Can we partition the indexing for faster temporal querying ?
  - What types of query processing and optimizations can we expect ?
  - How do we maintain indexes in case of incremental updates ?

- Index Compression
  - How does one exploit redundancy to reduce index size ?
  - Lossless vs Lossy compression

Time-Travel Index

- Vertically Partition the temporal space and each partition
- Now multiple posting lists per term, each with a valid time interval
- Limits index access, introduces replication



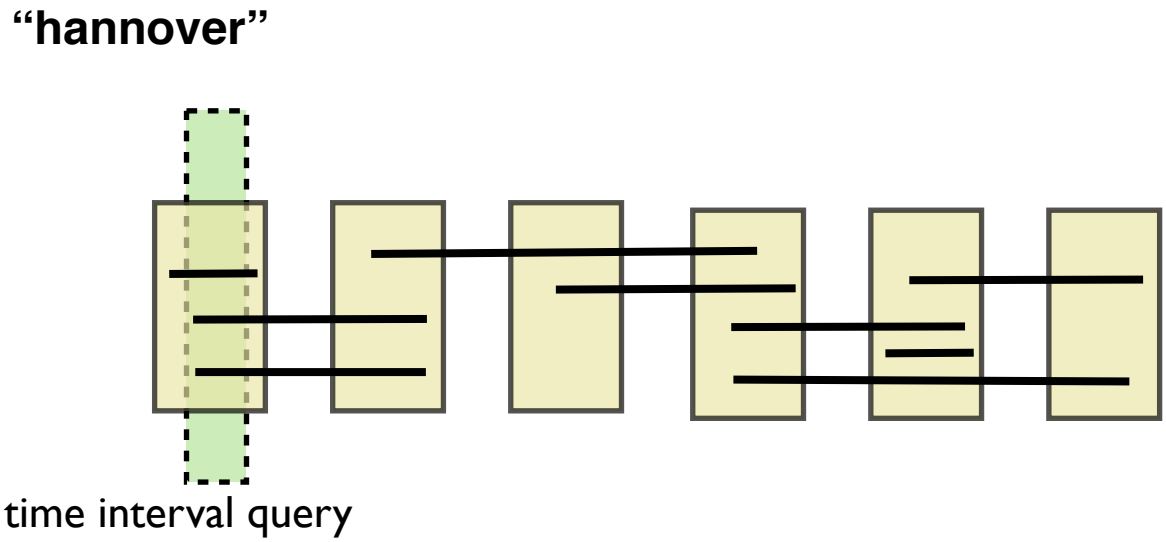
- Posting payloads arbitrary (e.g., scalar or positional)
- Posting-list order arbitrary but consistent for the whole index

[Berberich et al., SIGIR 2007]

Vertical Partitioning - Query Processing

- Dictionary or Lexicon should contain partitioning information
- For each temporal query, select a subset of affected partitions and only read them
- Filter postings which do not overlap with query time interval

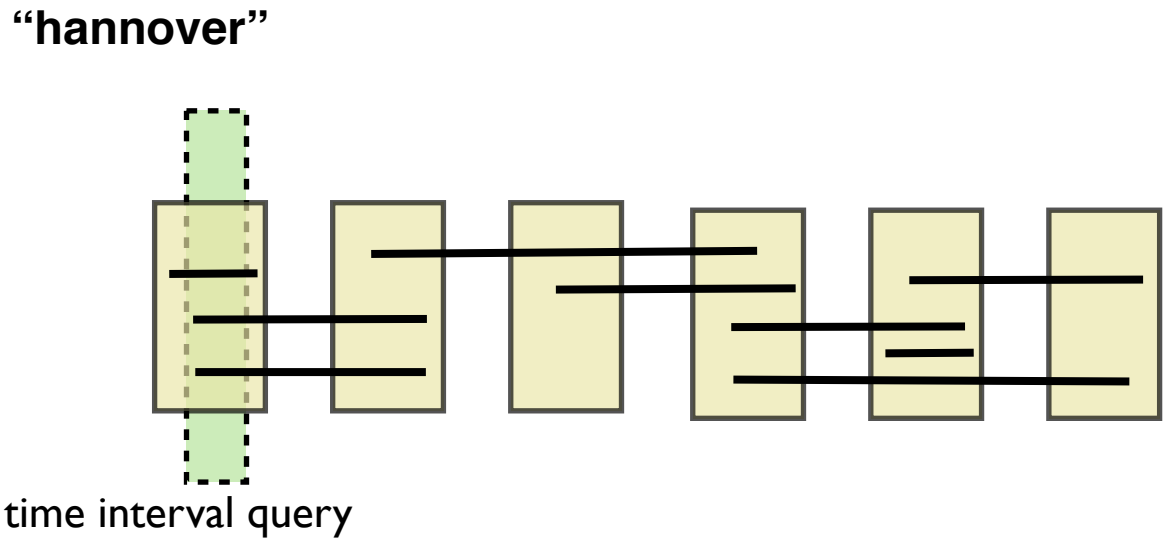
term	partition	offset	
hannover	[t1 - t5)	12646	
hannover	[t5 - t7)	12673	
hannover	[t7 - t25)	13446	
hannover	[t25 - t43)	15324	



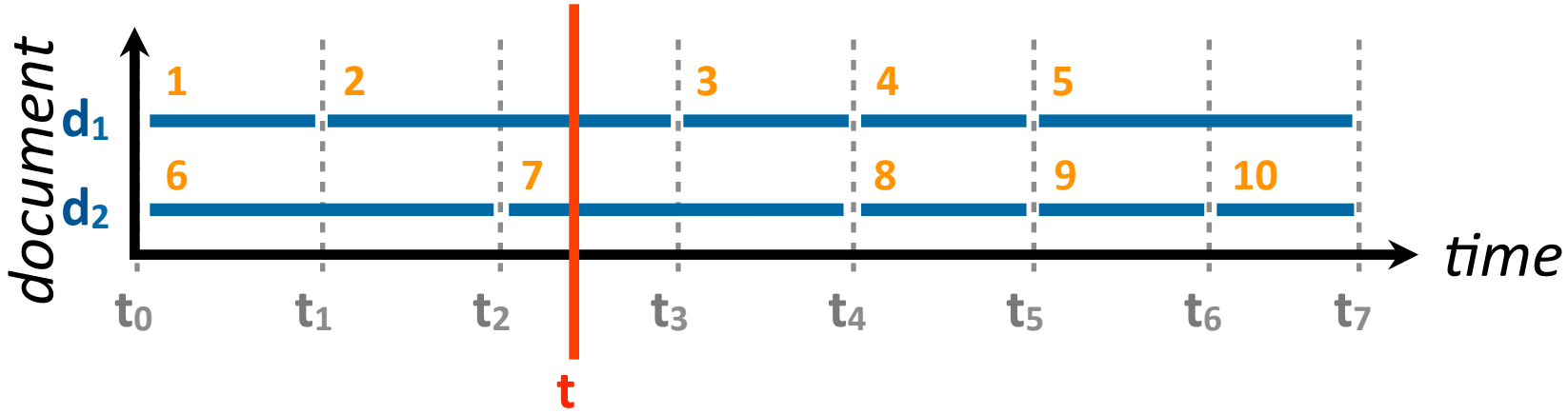
Vertical Partitioning – Replication and Blowup

- Replication of postings of versions with long intervals
- Overall postings associated with a term increases introducing index blowup

term	partition	offset	
hannover	[t1 - t5)	12646	
hannover	[t5 - t7)	12673	
hannover	[t7 - t25)	13446	
hannover	[t25 - t43)	15324	



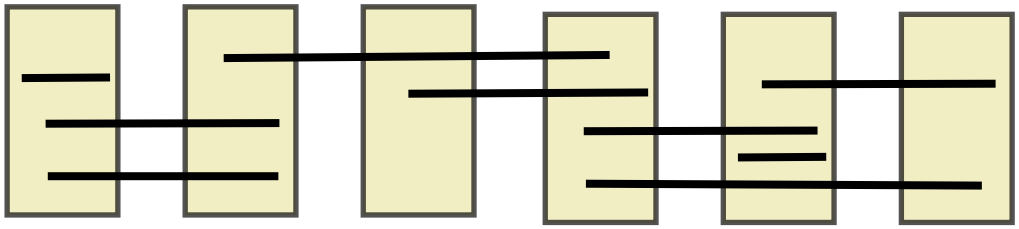
Partitioning Strategies



- **Space Optimal (SOpt)**
  - keeps one posting list L_v per term **v**
  - consumes minimal space
    but achieves only bad query-processing performance
- **Performance Optimal (POpt)**
  - keeps one posting list L_v:$[t_i, t_{i+1})$ for each elementary time interval
  - achieves optimal performance for time-point queries
    but consumes a lot of space (in the worst case $O(|L_v|^2)$)

Partitioning Strategies

- Given an input sequence of intervals how do we partition them into sublists?
- Space Bound Materialization Approach : We have a limited budget for space, need to maximize our performance
- Performance Guarantee Approach: For any query we need a guarantee on the performance loss, need to minimize blowup

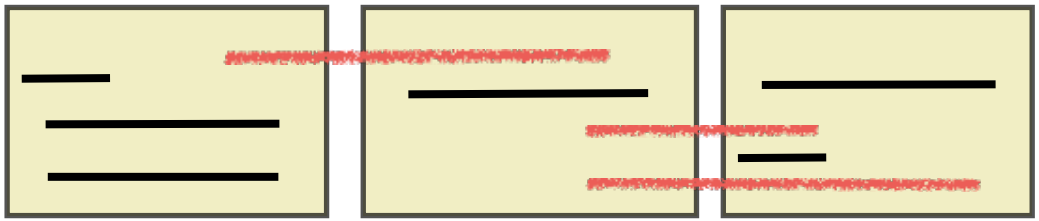


Trade-off size and performance

[Berberich et al., SIGIR 2007]

Space Bound Approach

- Minimize expected number of postings read for a time-point query, while ensuring that the index contains at most κ times the optimal number of postings
- Optimal solution computable in $O(|S| \times n^2)$ time and $O(|S| \times n)$ space using dynamic programming over prefix subproblems $[t_2, t_k)$ and space bounds $s \leq \kappa \cdot |L_v|$



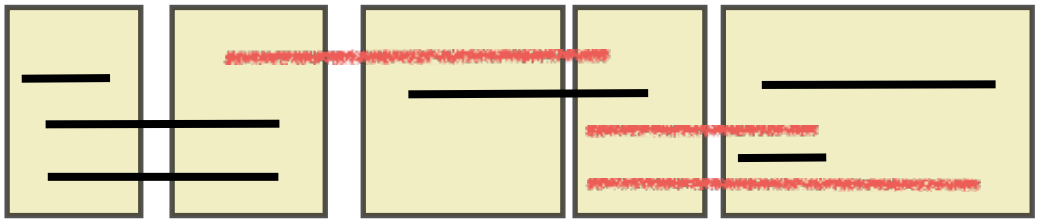
Space budget = 1/3 . (optimal number of postings)

[Berberich et al., SIGIR 2007]

Performance Guarantee

- Minimize total number of postings kept in the index, while guaranteeing that for any time-point query the number of postings read is at most a factor γ worse than optimal

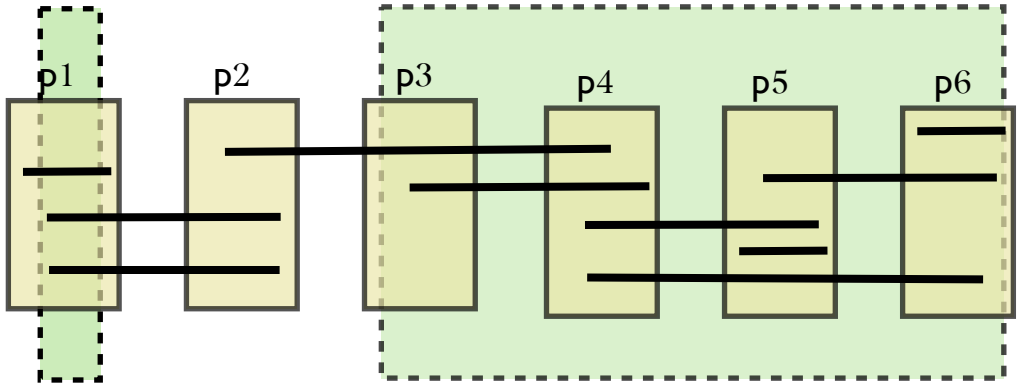
- Optimal solution computable in time $O(|L_v| + n^2)$ and space $O(n^2)$ using dynamic programming over prefix subproblems $[t_2, t_k)$



performance guarantee = γ (times the number of optimal results at that time)

[Berberich et al., SIGIR 2007]

Limitations of Vertical Partitioning



Relevant postings = 3
Postings read : 3

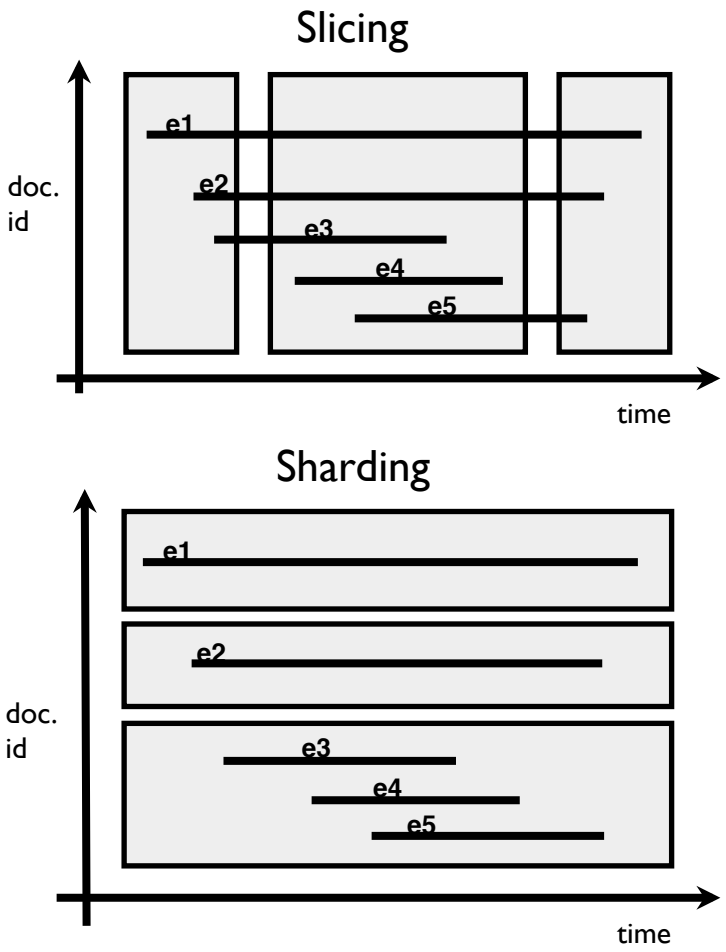
Relevant postings = 7
Postings read : 2 + 4 + 4 + 3 = 13

- Index size blowup due to replication of postings across slices
- Query processing inefficient if replicated postings are accessed multiple times

Can we partition a posting list without replicating postings?

Index Sharding

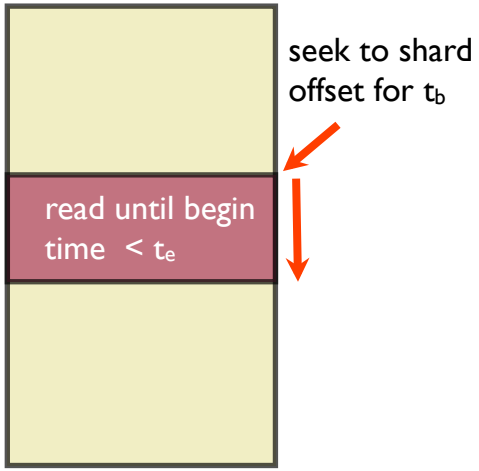
- Partition documents in each posting list into sublists called shards
- Contents of each shard disjoint - no replication, no index blowup
- Postings stored in begin time order
- Access structure over each shard for efficient query processing

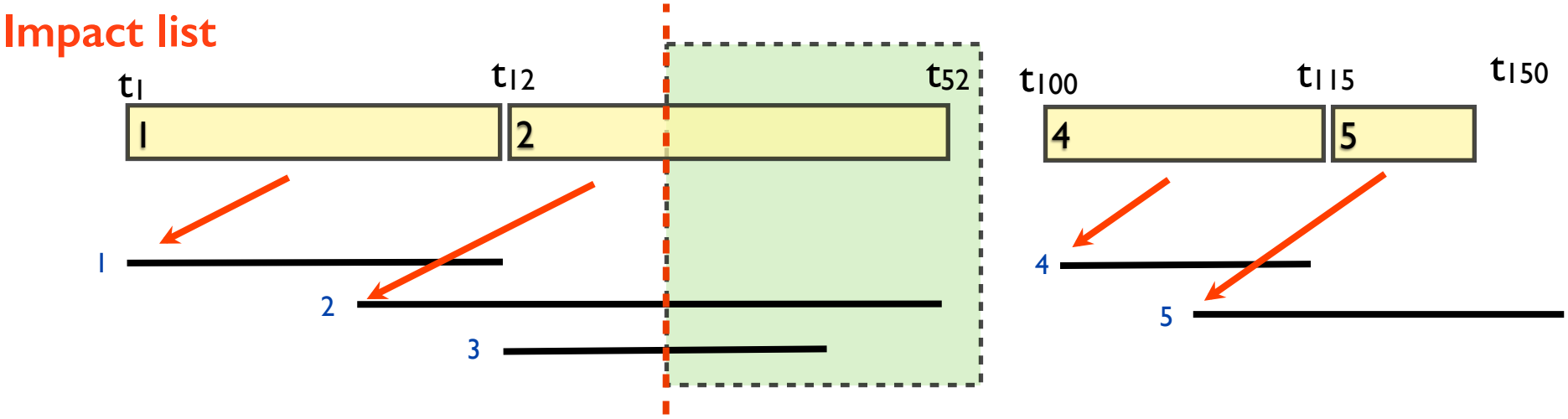


[Anand et al., SIGIR 2011]

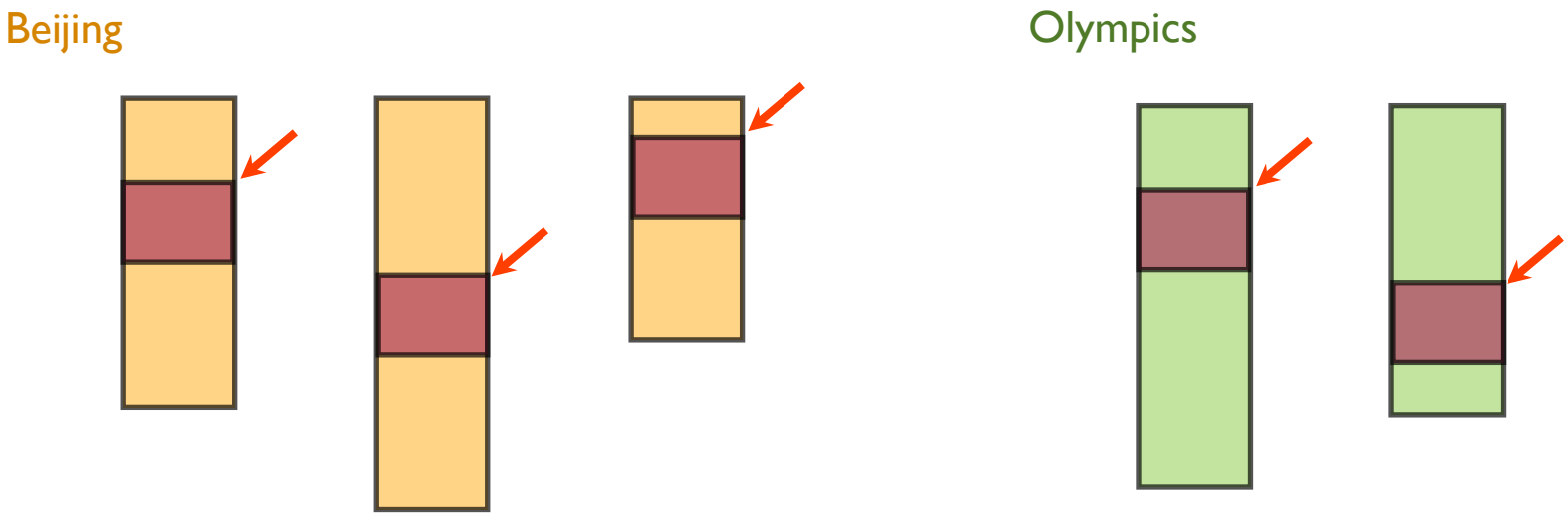
Index Sharding - Impact Lists

- Open - Each shard of a query term opened for access
- Skip - Given a query begin time seek to appropriate offset
- Scan - Read while postings still have overlap with query time interval





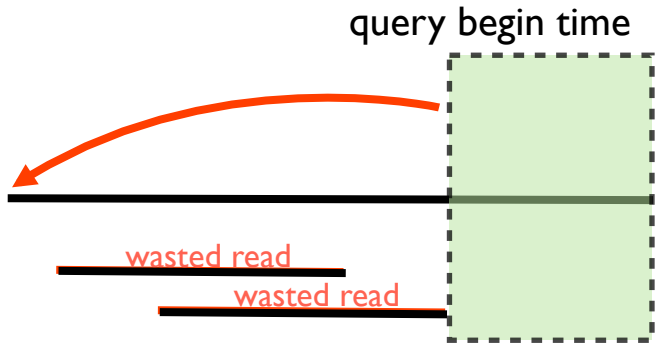
Index Sharding



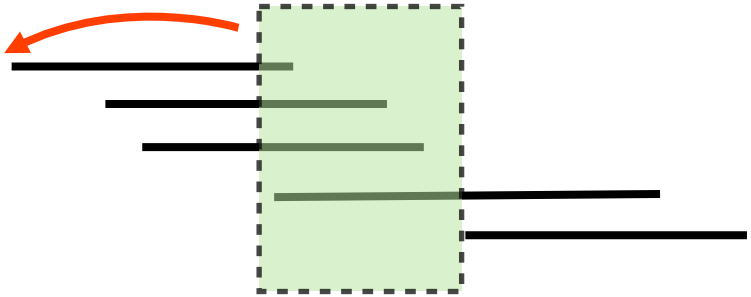
- All shards for a given query term are accessed
- Open-skip-scan on each shard assisted by impact lists
- Result list constructed by merging results from each shard

Index Sharding - Staircase Property

- Wasted reads are processed but do not overlap with the query time interval
- Staircase property in a shard
  - Intervals arranged in begin time order
  - No interval completely subsumes another interval
- Eliminates wasted reads



wasted reads due to subsumption



staircase property

[Anand et al., SIGIR 2011]

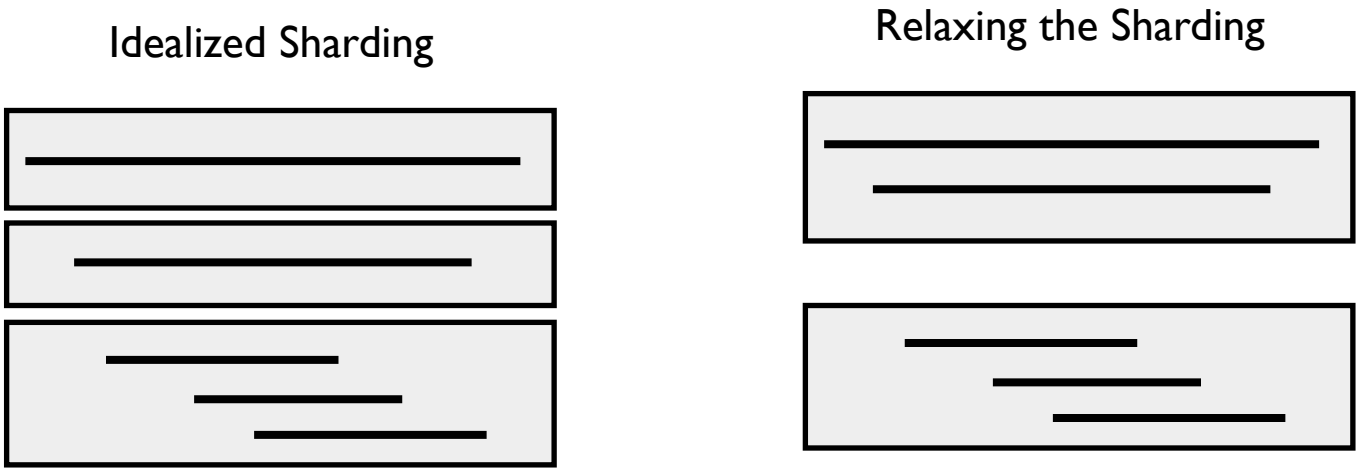
Index Sharding - Idealized Sharding

- Staircase property eliminates sequential accesses of postings non-overlapping with query time interval
- Minimizing number of shards is essential in minimizing number of random accesses
- **Input :** Set of postings/intervals corresponding to a postings list
- **Problem Statement :** Minimize the number of shards where each shard exhibits the staircase property

Greedy Algorithm exists which is proven to be optimal

[Anand et al., SIGIR 2011]

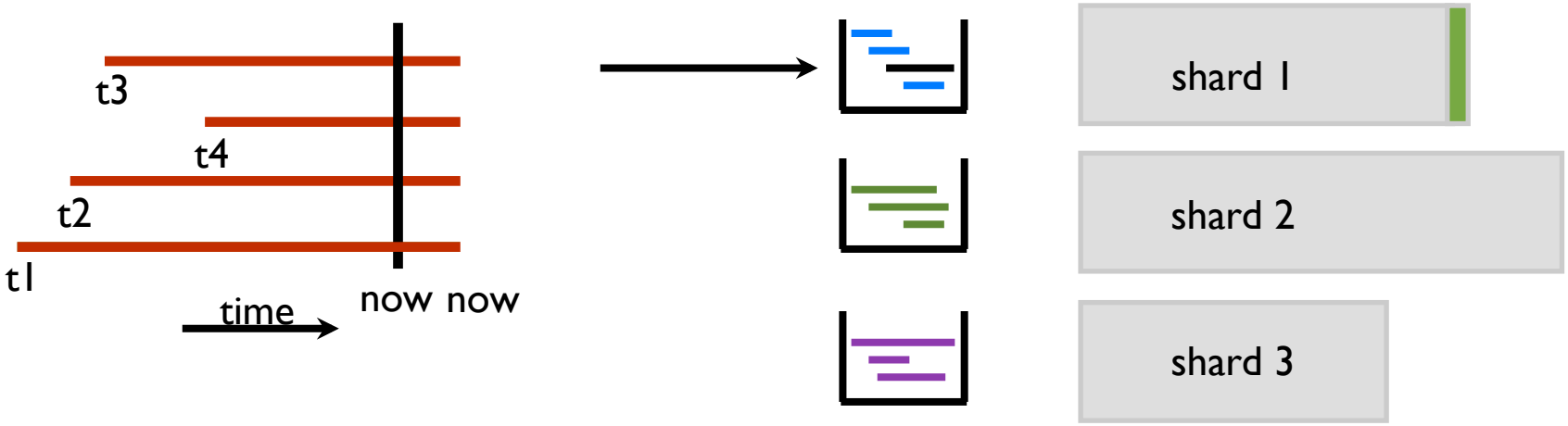
Index Sharding - Challenges



More shards the more the random accesses to disk

- Random accesses (RA) are typically much more expensive than sequential accesses (SA)
- Allow wasted reads to balance SA and RA
- Ensure that index is easily updatable

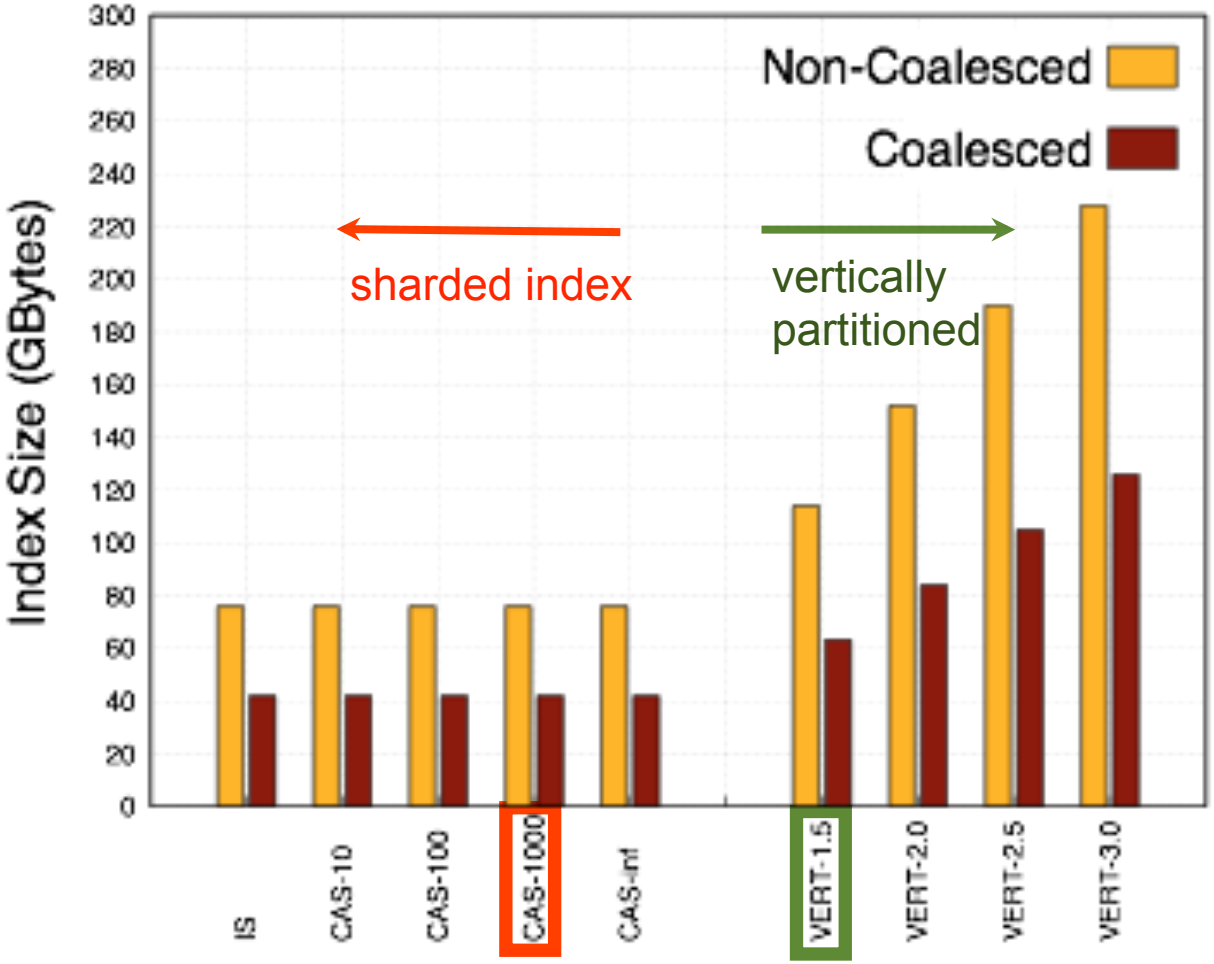
Incremental Sharding



- Algorithm assigns incoming posting to a shard
- Posting inserted into shard buffer maintaining begin time order
- Top posting popped and appended to the shard end

[Anand et al., SIGIR 2012]

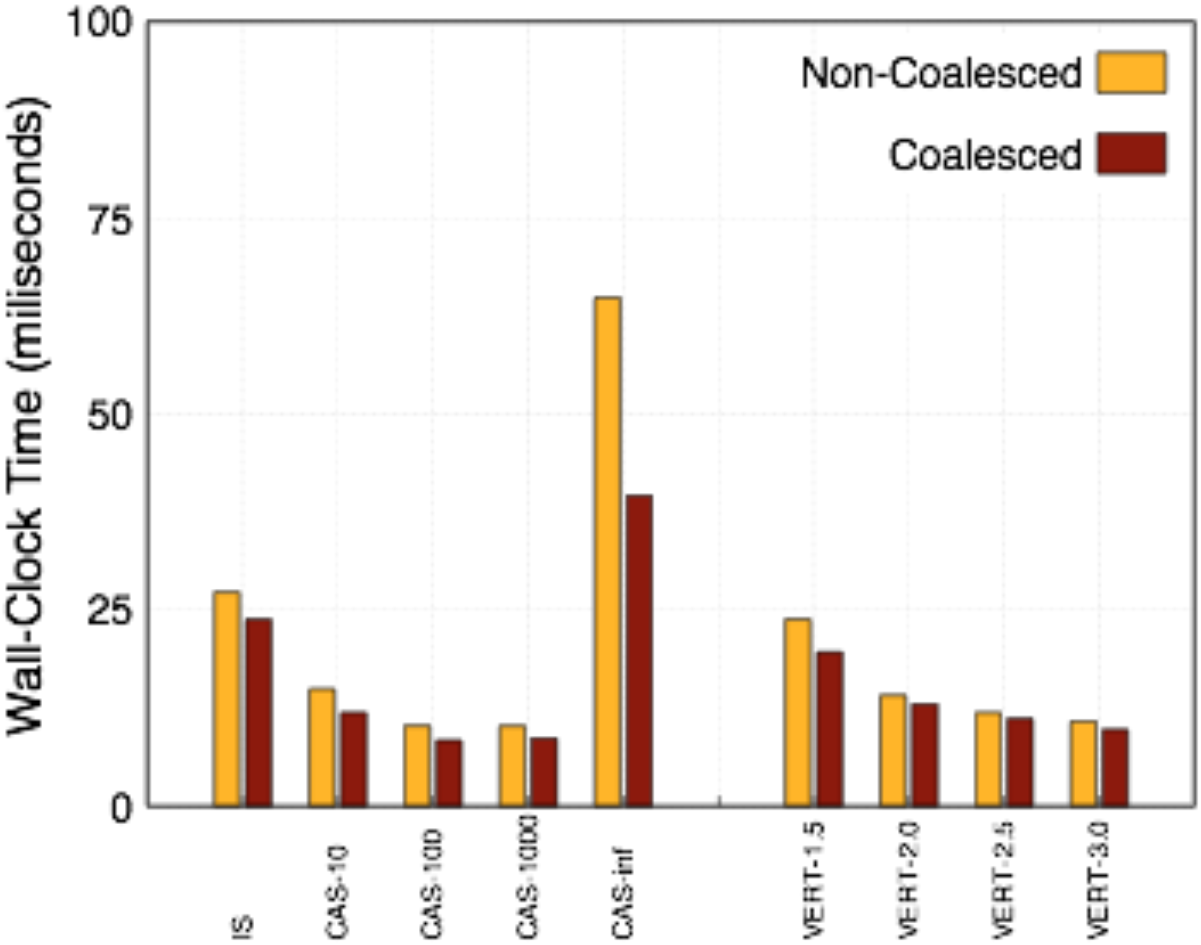
Index Sizes



Dataset: Wikipedia

- No index size blowup in sharded index
- Impact lists relatively small (1-7% of total index size)
- Temporal coalescing possible on sharded lists

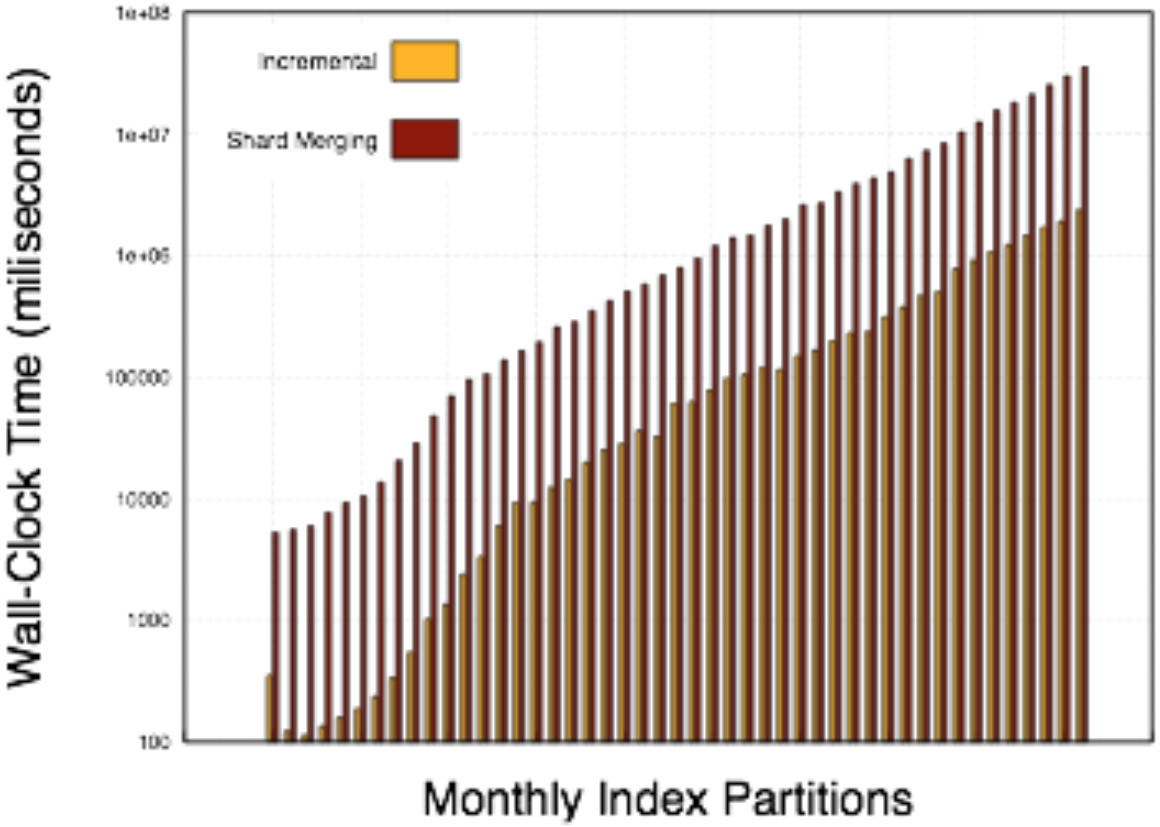
Query Processing



Dataset: Wikipedia

- Day granularity queries
- 62% improvement over IS
- Outperforms CAS-inf
- 22.2% improvement over best VERT for year granularity queries

Index Maintenance

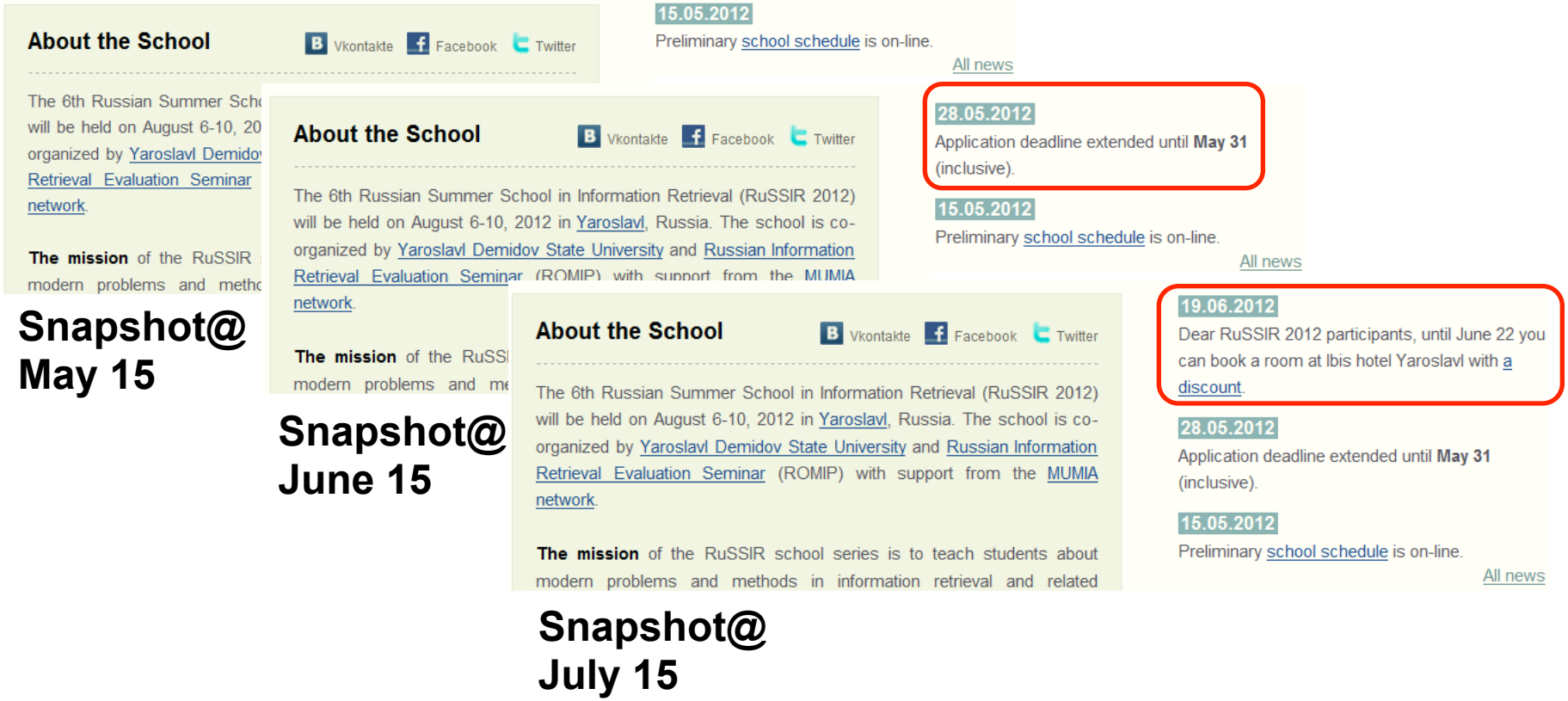


Dataset: Wikipedia

- Month based partial indexes
- Immediate merge employed
- INC outperforms CAS by 4x in UKGOV and 10x in WIKI
- CAS and INC are comparable in query processing performance

Versions Are Similar

- There is a high overlap between versions!



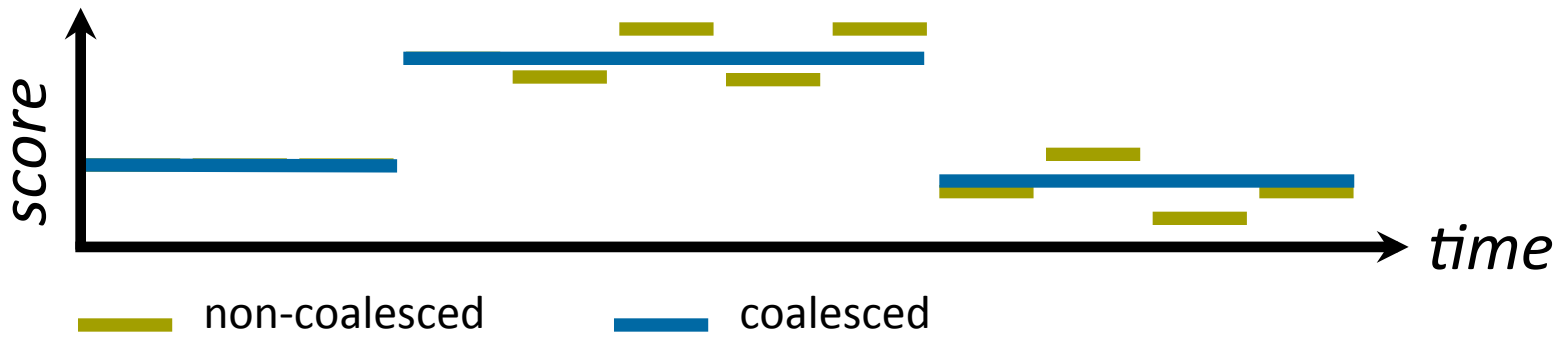
Versioned Collections – Data Characteristics

- **Most changes are small**
  - More than 50% changes between two consecutive versions are less than 5 terms.
- **Term changes are bursty**
  - Terms just appeared are likely to disappear again shortly
- **Change size is bursty**
  - Less than 10% versions makes up more than 50% and 70% changes in wiki and Ireland
- **Terms are dependent**
  - 48.8% terms disappear together if they come together
  - 30.5% terms disappear together otherwise

Temporal Coalescing

d1, [3,7), 4.5

- Scalar payloads represent, e.g., term-frequency information as needed for time-travel keyword queries
- **Observation:** Changes between document versions
  - are often minor (e.g., corrected typos)
  - have little effect on scalar payloads (e.g., 41 vs. 42 x dog)
  - and thus little impact on query results
- **Idea:** Coalesce sequences of postings that belong to the same document and have almost-identical scores



[Berberich et al., SIGIR 2007]

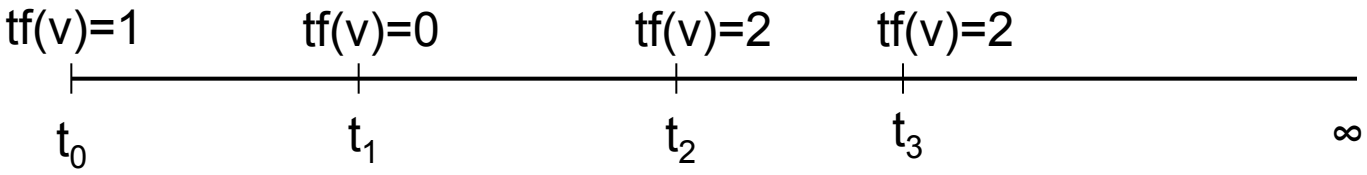
An Alternative Perspective

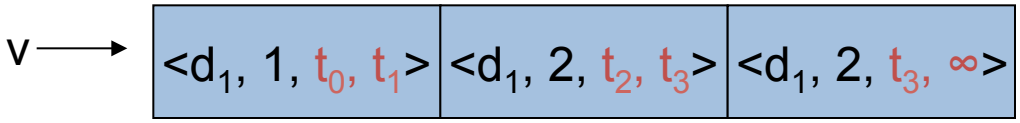
- Focus on the index size

  Approaches up to now consider each version of a document separately: no special attention on the overlap between versions

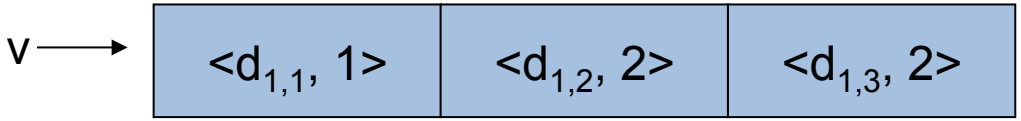
- **Key ideas**
  - Small integers can be represented with smaller codes
  - Doc ids are not so small: instead, compress the gaps between the ids
  - Term frequencies are already small

Indexing Versions





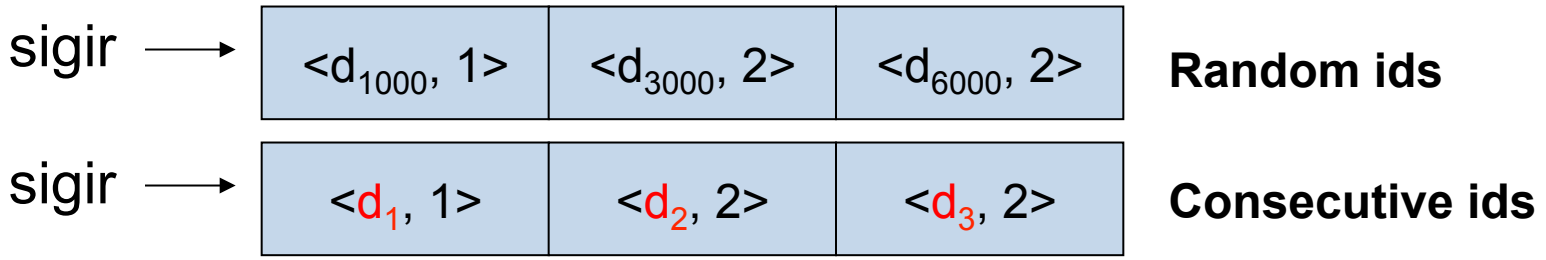
- Assume validity-intervals are stored separately → Time-constraints @ post-processing
- Versions of d_i are represented as $d_{i,j}$



How can we reduce the storage space for such a representation?

Exploit Redundancy between Versions

- **Naive idea (ID re-assignment):**
  - Assign consecutive doc IDs to consecutive versions of the same document
  - Allows small d-gaps for overlapping terms among the versions



d_1: www.sigir.org, [May 15, June 15]
d_2: www.sigir.org, [June 15, July 15]
d_3: www.sigir.org, [July 15, Aug 15]

[He et al., CIKM 2009]
[He et al., CIKM 2010]

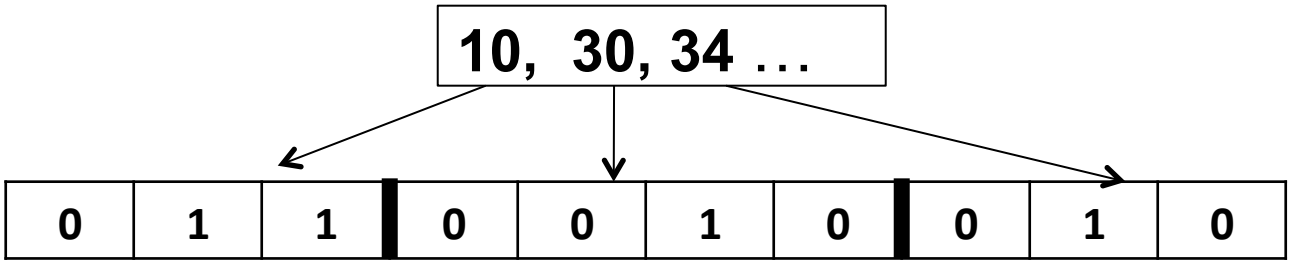
MSA and Diff Approaches

- **Multiple Segment Alignment** [Herscovici et al., ECIR'07]
  - Given *d* with some versions, a virtual document $V_{i,j}$ is all terms occuring in (only) versions i through j
  - Reduces the number of postings but increases the document space! (theoretically, up to N^2 virtual documents!)
- **DIFF approach** [Anick et al., SIGIR'92]
  - For every pair of consecutive versions d_i and d_{i+1}; create a virtual document that is the symmetric difference between these versions.

[He et al., CIKM 2009]
[He et al., CIKM 2010]

Two level Index

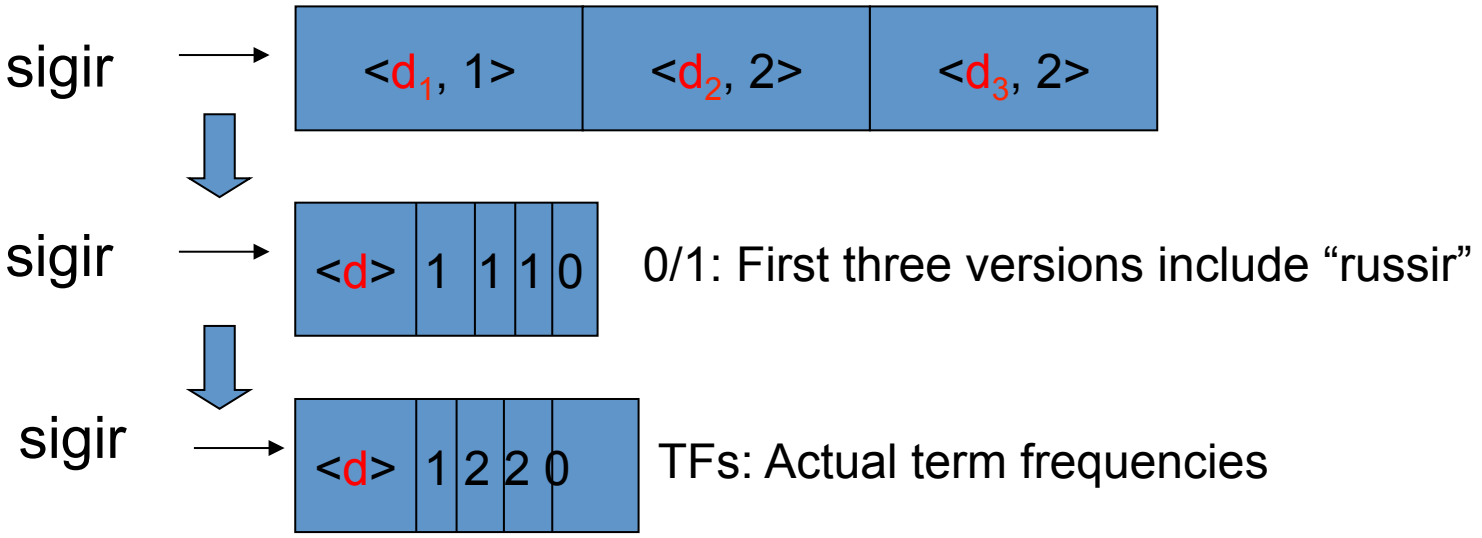
- **Two-level Indexes**
  - Top level index the union of all versions
  - Lower level using bit vectors.



  - The length of each bit vector is the number of versions in the document.
  - For bit vector of term **t**, if t appears in ***i***th version, the ***i***th position of bitmap is set to 1 otherwise, it is set to 0

[He et al., CIKM 2009]
[He et al., CIKM 2010]

Two-level indexing

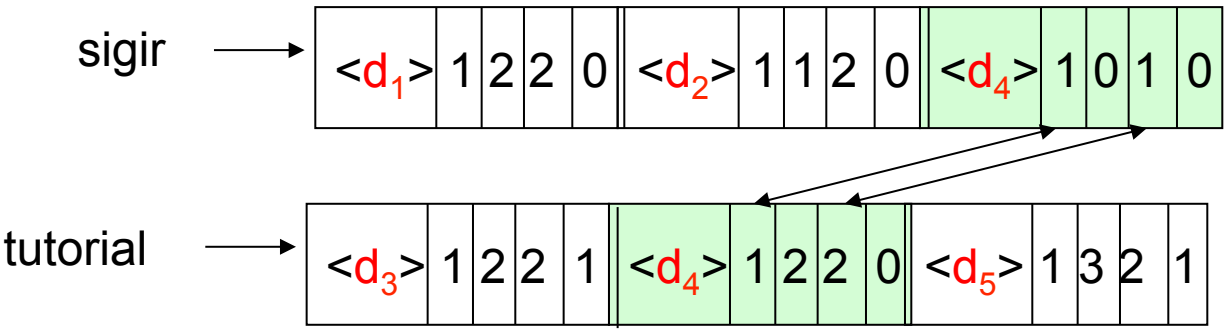


- In actual implementation, group blocks of doc ids and bitvectors of versions
- Bitvectors best compressed by hierarchical **Huffman** coding !

[He et al., CIKM 2009]

Two-level indexing: HUFF

- **Query Processing:**
  - Decompress bit vectors of documents that are in the **intersection** of query terms



Final result: $\mathbf{d_{4,1}}$ and $\mathbf{d_{4,3}}$

Performance

- Two index compression schemes
  - Interpolative coding (IPC)
  - PForDelta with optimizations (PFD)

- Datasets: Wiki, Ireland

Index sizes for doc ids

	Wikipedia	Ireland
Random-IPC	4,957	2,908
Random-PFD	5,499	3,289
Sorted-IPC	570	1086
Sorted-PFD	583	1137
DIFF-IPC	269	751
DIFF-PFD	323	927
MSA-IPC	237	682
MSA-PFD	287	799

Query processing times (in-memory): Sorted is the best!

[He et al., CIKM 2009]
[He et al., CIKM 2010]

Two level indexing

- Same idea can be applied to **MSA** and **DIFF**
  - For virtual documents, create bitvectors as before
  - Compress first level **IPC**, second level **PFD**
- Even better compression performance

Index sizes for doc ids

	One Level		Two Level	
	Wiki	Ireland	Wiki	Ireland
DIFF-IPC	269	751	253	652
DIFF-PFD	323	927	269	757
MSA-IPC	237	682	233	663
MSA-PFD	287	799	243	672
HUFF			213	577

[He et al., CIKM 2009]
[He et al., CIKM 2010]

Temporal Retrieval Models

Major Research Directions

- Retrieval models for temporal search
  - Temporal information needs
  - Language models for temporal information needs
  - Explicit, precision oriented
- Recency based retrieval models recency in queries
  - Balancing relevance and recency
  - Implicit, precision oriented
- Retrieval models for diversity
  - Only temporal diversity
  - Aspect temporal diversity
  - Recall oriented, specialized users

Temporal Information Need

- Existing retrieval models don't perform well for information needs that have a temporal dimension

fifa world cup
1990's

In the last decade the FIFA World Cup was won by Germany (in 1990), Brazil (in 1994), and France (in 1998).

- **Idea:** Leverage temporal expressions (e.g., in 1998) contained in documents and queries
- **Challenges:** Meaning of temporal expressions uncertain, not clear how to seamlessly integrate them into a retrieval model

Temporal Information Need

- Information needs that have a temporal dimension

*FIFA World Cup tournaments **of the 1990's***
*Movies that won an Academy Award **in 2007***
*Crusades **of the 12th century***
*London Summer Olympics **2012***

- Queries that contain a temporal expression (e.g., in 1998)
  - indicate an underlying temporal information need
  - account for 1.5% of general web queries
  - are more common for specific domains (e.g., News or Sports) and/or specific user groups (e.g., historians or journalists)
- **But:** Not well-supported by existing retrieval models

Temporal Expressions

- Temporal expressions can be categorized as:
  - **Explicit** (e.g., July 19th 2010 or September 1872)
  - **Implicit** (e.g, Christmas 2009 or New Year's Eve 2000)
  - **Relative** (e.g., yesterday or last month)

- Meaning of a temporal expression is often uncertain, e.g.,

- **Document and Query Model:** Distinction between the textual part and the temporal part of documents and queries
  - $\mathbf{d}_{\text{text}}$ and $\mathbf{q}_{\text{text}}$ are bags of textual terms
  - $\mathbf{d}_{\text{time}}$ and $\mathbf{q}_{\text{time}}$ are bags of temporal expressions

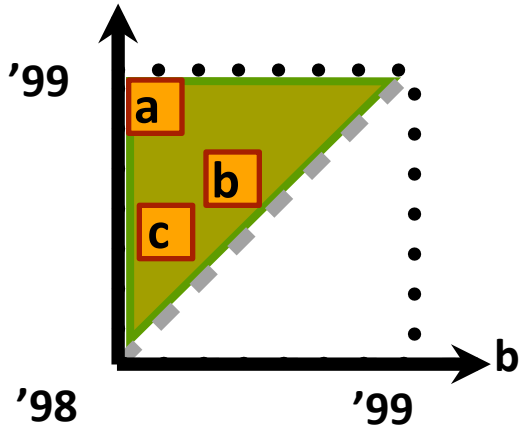
Temporal Expression Model

- We model a temporal expression as a four-tuple

$$T = (\, tb_l,\, tb_u,\, te_l,\, te_u \,)$$

that records the earliest/latest begin/end of any time interval **[b, e]** that **T** may refer to

- The temporal expression in 1998, e.g., is represented as **(1998/01/01, 1998/12/31, 1998/01/01, 1998/12/31)**



(a) [1998/01/01, 1998/12/31]

(b) [1998/07/12, 1998/07/12]

(c) [1998/02/07, 1998/02/22]

[Berberich et al., ECIR 2010]

Language Models in Information Retrieval

- Probabilistic generative model estimated for each document
- Query-likelihood approaches rank documents based on their probability of generating a given query
- Unigram language model estimates probability of generating query **q** from document **d** as

$$P(q \mid d) = \prod_{v \in q} P(v \mid d)$$

- Jelinek-Mercer smoothing estimates the probability **P(v|d)** of generating the term **v** from document **d** as

$$P(v \mid d) = \lambda \cdot \frac{tf(v, d)}{|d|} + (1 - \lambda) \cdot \frac{tf(v, C)}{|C|}$$

Language Model Framework

- Query-likelihood approach assuming that the textual and temporal query part are generated independently

$$P(q \mid d) = P(q_{text} \mid d_{text}) \times P(q_{time} \mid d_{time})$$

- Independent generation of query temporal expressions

$$P(q_{time} \mid d_{time}) = \prod_{Q \in q_{time}} P(Q \mid d_{time})$$

- Two-step generation of temporal expression Q

  (1) Draw a temporal expression **T** at uniform random

$$P(Q \mid d_{time}) = \frac{1}{|d_{time}|} \sum_{T \in d_{time}} P(Q \mid T)$$

  (2) Generate **Q** from **T**

[Berberich et al., ECIR 2010]

Uncertainty-Aware LM

- Intuitively, **P(Q|T)** reflects the probability that the user issuing the query and the author writing the document have the same time interval in mind

- The definition of **P(Q|T)** can be simplified as

$$P(Q \mid T) = \frac{|T \cap Q|}{|T| \cdot |Q|}$$

  treating **Q** and **T** as sets of time intervals

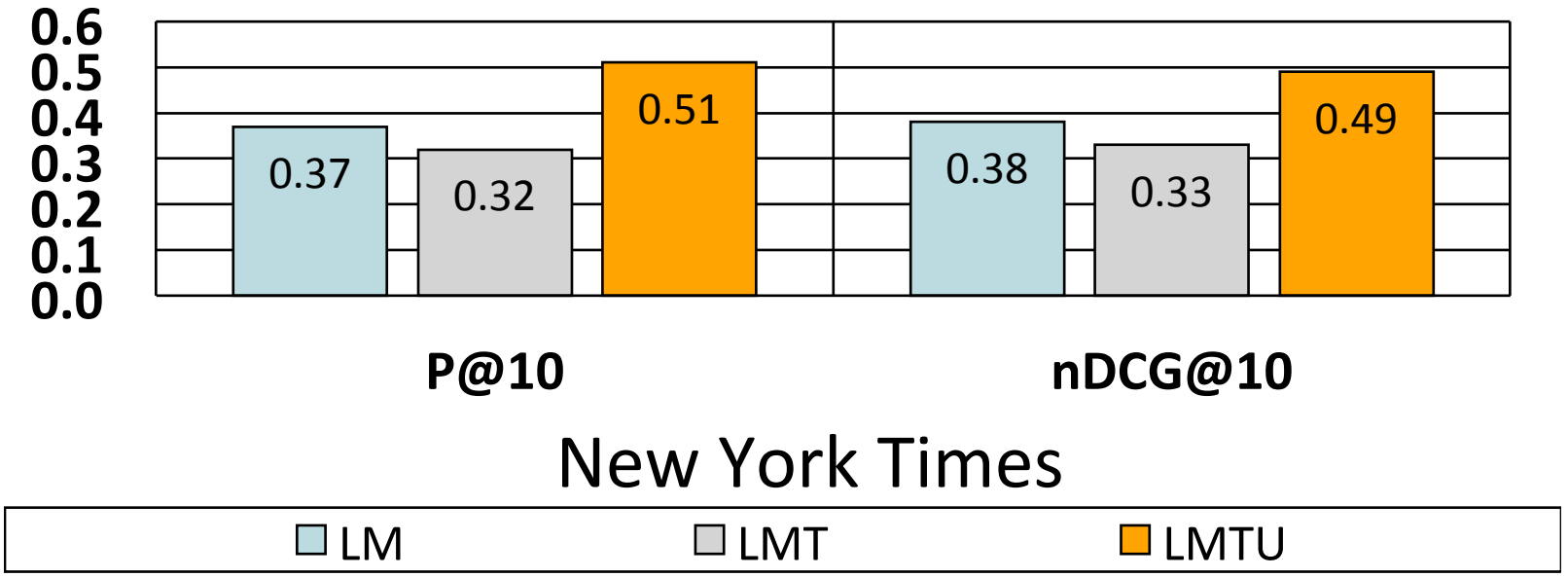
- **|T|, |Q|, and |T ∩ Q|** efficiently computable based on the four-tuple representation, i.e., no need to enumerate the huge but finite number of time intervals in **Q** and **T**

[Berberich et al., ECIR 2010]

Retrieval Effectiveness

- Queries and binary relevance assessments collected using the crowdsourcing platform Amazon Mechanical Turk

- Query workload consists of 40 queries, e.g.:
  boston red sox [october 27, 2004], pink floyd [march 1973],
  wright brothers [1905], siemens [19th century], babe ruth [1921]

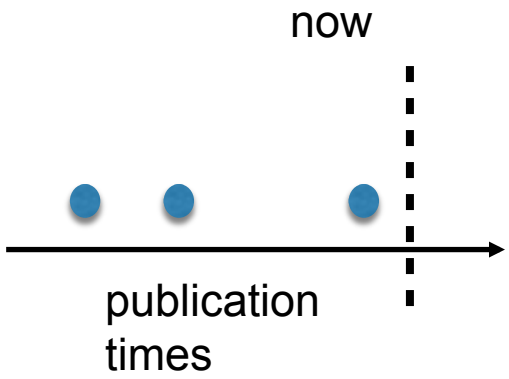


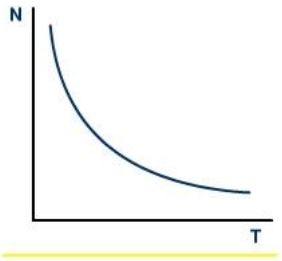
Recency Aware Ranking

- Freshness of documents could be utilised for improving ranking
- How do we include freshness ?

$$P(D|Q) \propto P(Q|D).P(D)$$

- Querying time or hitting time : Time when the query was issued
- Recently published documents are more relevant that older documents
- Exponential decay to model old documents





$$P(D) = \lambda e^{-\lambda.(t_{now}-t_D)}$$

freshness param.

pub. time

[Li and Croft, CIKM 2003]

Timeliness

- Equal decay to all queries

$$P(D|Q) \propto P(Q|D).P(D) \rightarrow P(D) = \lambda e^{-\lambda.(t_{now}-t_D)}$$

- Some queries are more timely than others
  - "hairstyle fashion trends", "olympics", "house of cards"
- Estimate different parameters for different queries from user assessments

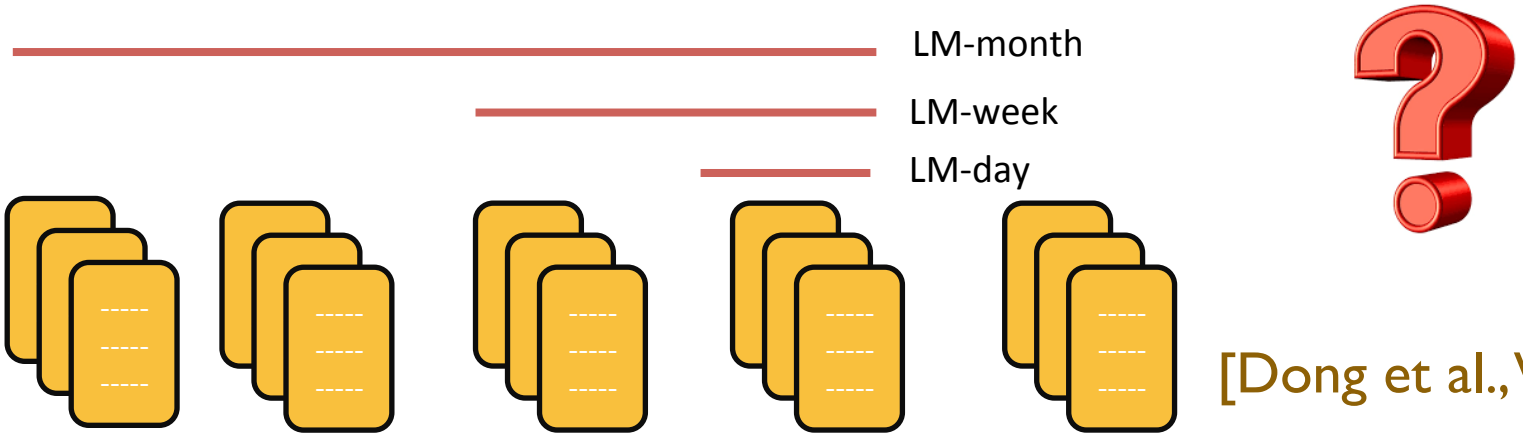
$$P(D) = \lambda_Q e^{-\lambda_Q.(t_{now}-t_D)}$$

[Efron and Golovchinsky, SIGIR 2011]

Recency Classification

- Not all queries are recency sensitive
- Classification of recency sensitive queries
  - Buzziness of queries: present popularity vs past background reference model of popularity
  - Affinity LM at different granularities
    - LM-day
    - LM-Week
    - LM-Month

$$\text{buzz}(q,t,Q) = \max_i \log \hat{P}(q|M_{Q,t}) - \log P(q|M_{Q,t-r_i})$$



[Dong et al., WSDM 2010]

Recency Ranking

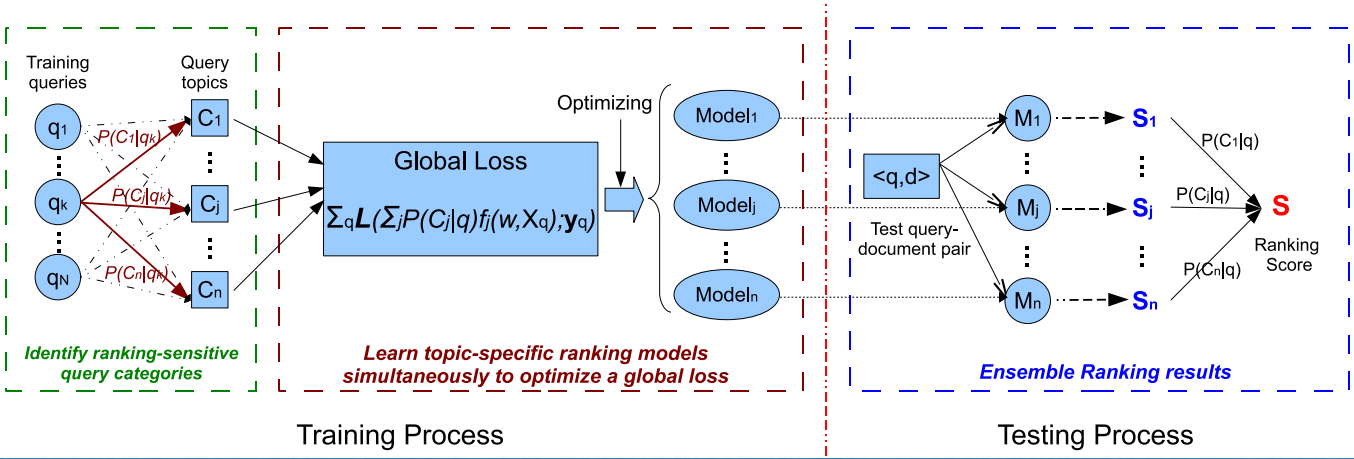
- Learning to Rank framework
  - **Features**:
    - Age of page based on timestamps
    - Age of page based on link times
    - Temporal authority
    - Type of page : news, web, journal, …
  - Pairwise Learning to Rank [GBRANK]
- Transformation to training labels
  - Collect relevance assessments and freshness assessments separately
  - Demotion of relevance labels based on freshness labels

[Dong et al., WSDM 2010]

Relevance vs Freshness

- Misclassification can significantly degrade the performance
- Optimize for freshness and relevance where the trade-off is adaptive to query characteristics
- Divide and Conquer Paradigm
  - Multiple rankers with different objectives
  - Query is subject to multiple rankers and rankings are aggregated

[Dai et al., SIGIR 2011]
[Bian et al.,WWW 2010]



Relevance vs Freshness

- Labels are composed of both freshness and relevance judgements

$$\widetilde{y}_{q,d,i} = \frac{(1+\beta_i^2)\cdot y_{q,d}^R \cdot y_{q,d}^F}{y_{q,d}^R + \beta_i^2 \cdot y_{q,d}^F}$$

- For each ranker a different mixing parameter value is learned [flexibility]
- RankSVM incorporating query importance and document importance
  - Temporal Feature: STL decomposition of TS of tf-idf for each document (versions)

- Evaluation Measure: Hybrid NDCG

$$hybrid\ NDCG(n) = Z_n \sum_{j=1}^{n} \frac{2^{(\gamma \mathbf{y}_R + (1-\gamma)\mathbf{y}_F)} - 1}{\log_2(j+1)}$$

[Dai et al., SIGIR 2011]

Experiments

	Temporal Queries (Google Trends)			
	NDCG1	NDCG3	NDCG5	NDCG10
SepR	0.373	0.359	0.375	0.411
TopicalSVM	0.342	0.354	0.365	0.408
Over-weighting	0.355	0.351	0.368	0.411
CS-DAC	0.385	0.365	0.377	0.417
CS-DAC($\mathcal{U}$)	0.401†‡	0.375	0.389	0.426†
	Non-Temporal Queries (MSN logs)			
	NDCG1	NDCG3	NDCG5	NDCG10
SepR	0.481	0.517	0.532	0.562
TopicalSVM	0.490	0.508	0.521	0.566
Over-weighting	0.476	0.510	0.538	0.570
CS-DAC	0.493	0.520	0.541	0.574
CS-DAC($\mathcal{U}$)	0.509	0.522	0.541	0.574

Relevance measure

	Temporal Queries (Google Trends)			
	NDCF1	NDCF3	NDCF5	NDCF10
SepR	0.378	0.360	0.372	0.408
TopicalSVM	0.365	0.355	0.365	0.402
Over-weighting	0.340	0.348	0.363	0.404
CS-DAC	0.398‡	0.364	0.376	0.411
CS-DAC($\mathcal{U}$)	0.416†§‡	0.379‡	0.388	0.400
	Non-Temporal Queries (MSN logs)			
	NDCF1	NDCF3	NDCF5	NDCF10
SepR	0.348	0.411	0.434	0.475
TopicalSVM	0.355	0.408	0.430	0.485
Over-weighting	0.335	0.408	0.434	0.480
CS-DAC	0.427†§‡	0.454†§‡	0.473†§‡	0.510§‡
CS-DAC($\mathcal{U}$)	0.452†§‡	0.466†§‡	0.488†§‡	0.527†§‡

Freshness measure

Dataset : 158 million unique URLs and 12 bi. links from the .ie domain, covering the time span from Jan 2000 to Dec 2007 (one snapshot per month and 88 in total).

Historical Query Intent

- Give the user a historical overview.
- Good starting point for further exploration.

- Historical Query Intent: I want a historical overview of Rudolph Giuliani

[Singh et al., CHIIR 2016]

Temporal Diversity

- Consider time points as query intents, as opposed to topics from a taxonomy in their case.

$$\arg\max_{S} \sum_{t \in T} P(t|q) \cdot (1 - \prod_{d_i^{t_i} \in S} (1 - P(R|t, t_i) \cdot P(R|q, d_i)))$$

$$P(R|t, t_i) = \frac{1}{1 + e^{-\mathbf{w} + |t - t_i|}}$$

$$P(R|q, d_i) = \frac{P(q|d_i)}{\max_{d_j \in D} P(q|d_j)}$$

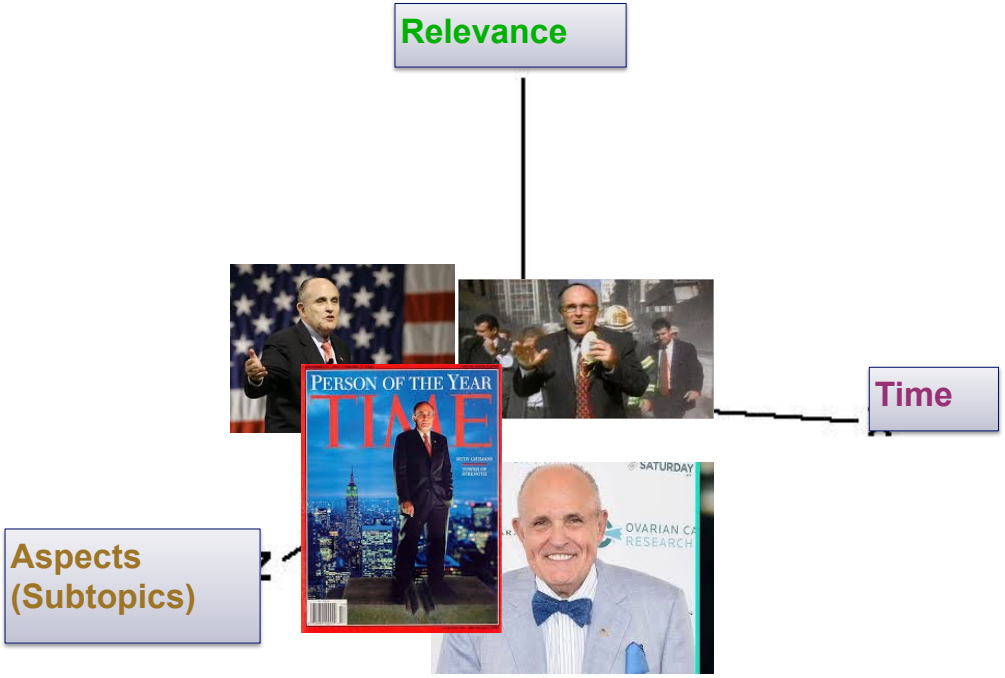
- **Temporal discounting:** User interested in time point t is satisfied with a document published at time t_i

- Modified greedy algorithm for diversification

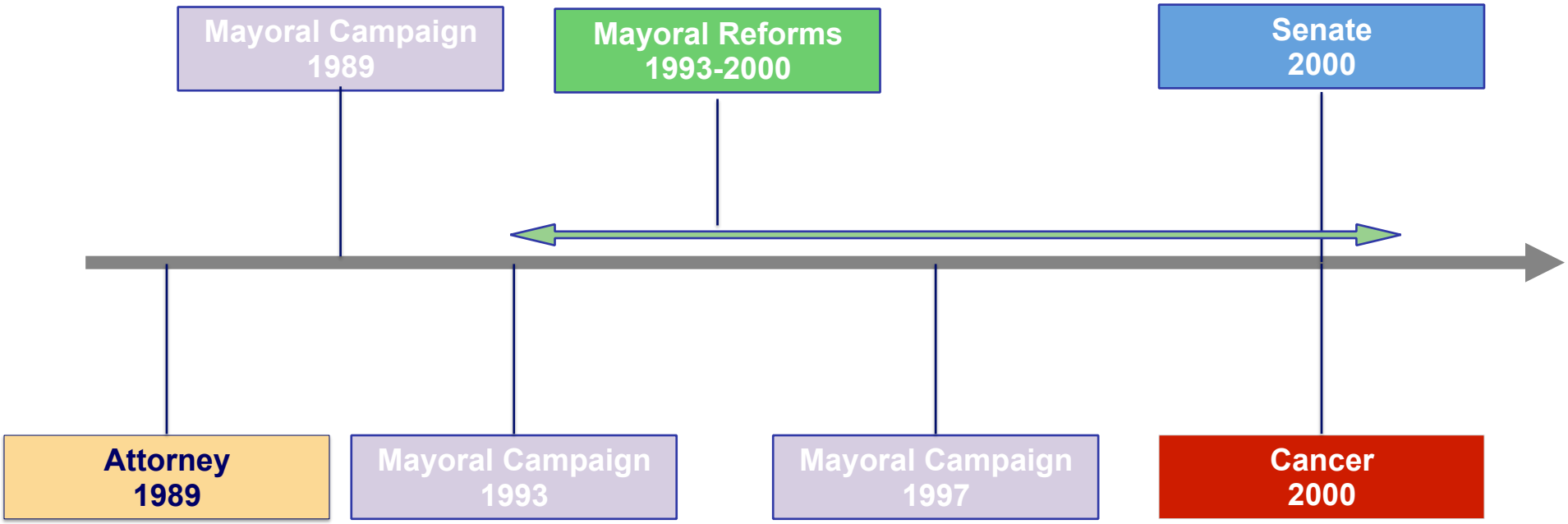
[Berberich and Bedathur, TAIA 2013]

Problem Statement

*"I am looking for **relevant** documents regarding the most important **aspects** from the **time** period when these aspects were relevant."*



Topics and Time

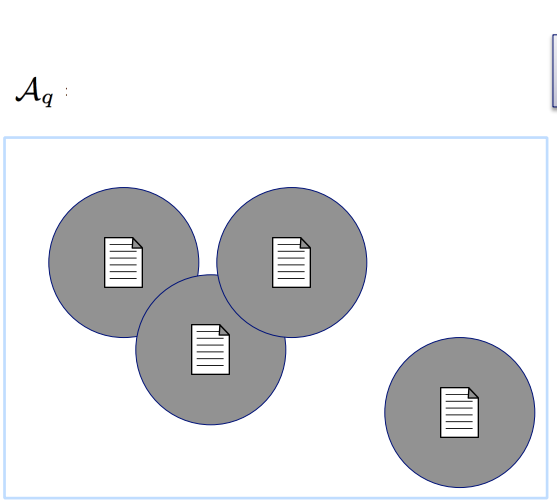


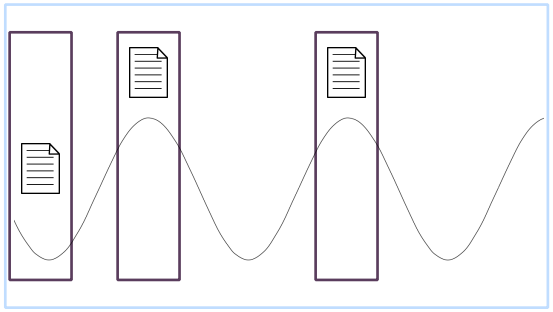
- Aspects are diverse across time
- Time periods are aspect diverse

[Singh et al., CHIIR 2016]

History by Diversity

2 Dimensional Coverage Problem → **Diversification based approaches**





Relevance

Time

Aspects (Subtopics)

IA-Select (Agrawal WSDM'09)
PM2 (Dang SIGIR'12)
xQuad (Santos WWW'10)

OnlyTime
(Berberich TAIA'13)

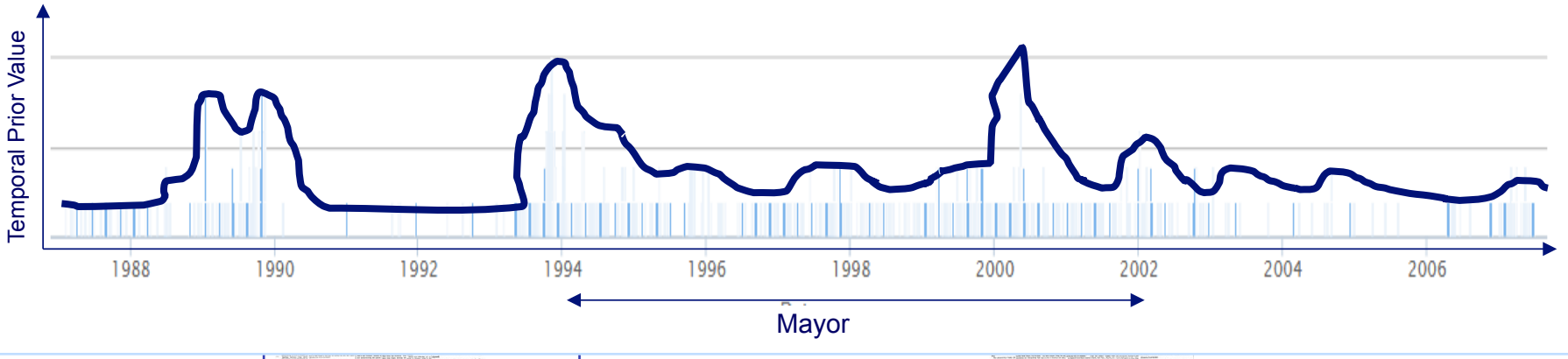
Multi-Dimensional Diversity (MDiv)
(Dou WSDM'11)
n-arbitrary dimensions

[Singh et al., CHIIR 2016]

Modelling Time

$$P(\delta_i|q) = \theta.P_{pub}(\delta_i|q) + (1-\theta).P_{ref}(\delta_i|q)$$

Primary sources — Secondary sources



Time

diagram with doc - primary secondary sources

from top k

[Singh et al., CHIIR 2016]

Modelling Aspects

CNN politics

Rudy Giuliani (sort of) endorses Donald Trump

By Julia Manchester, CNN

Updated 2114 GMT (0514 HKT) April 19, 2016

Entities as Aspects

Story highlights

Giuliani said he would endorse Trump, but not have a role on the campaign

He thinks Hillary Clinton would easily dispatch with Ted Cruz in a general election but have no idea how to handle Trump

Washington (CNN) — Former New York City Mayor Rudy Giuliani endorsed Republican front-runner Donald Trump Tuesday, the day of the New York primary.

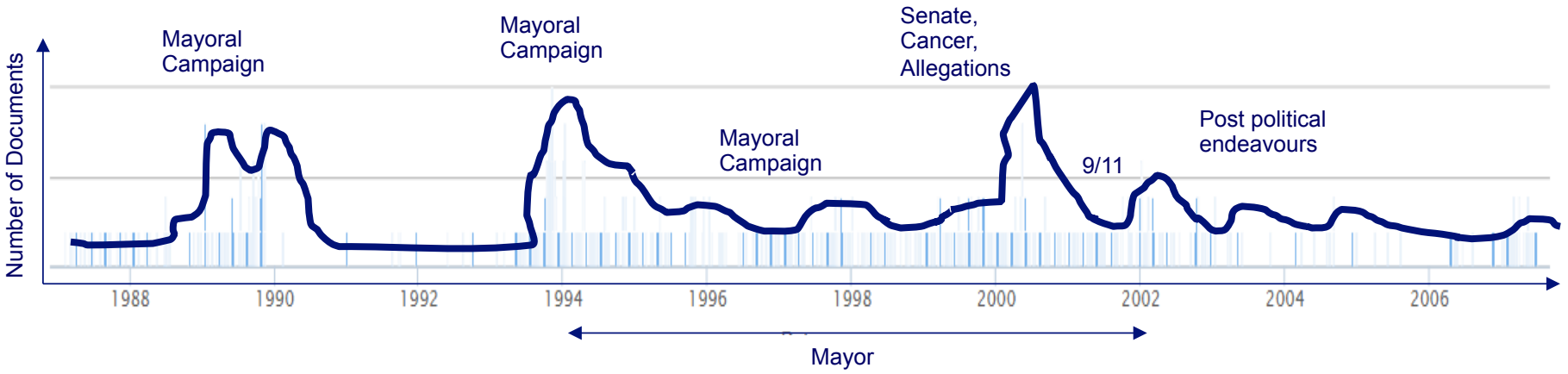
"I'll endorse, but I'm not a part of the campaign," Giuliani told CNN's Chris Cuomo on "New Day."

When pressed by Cuomo to clarify what he meant, Giuliani repeated that he would endorse Trump, but not have a role on the campaign.

Canonicalisation of entities

[Singh et al., CHIIR 2016]

The HistDiv Approach



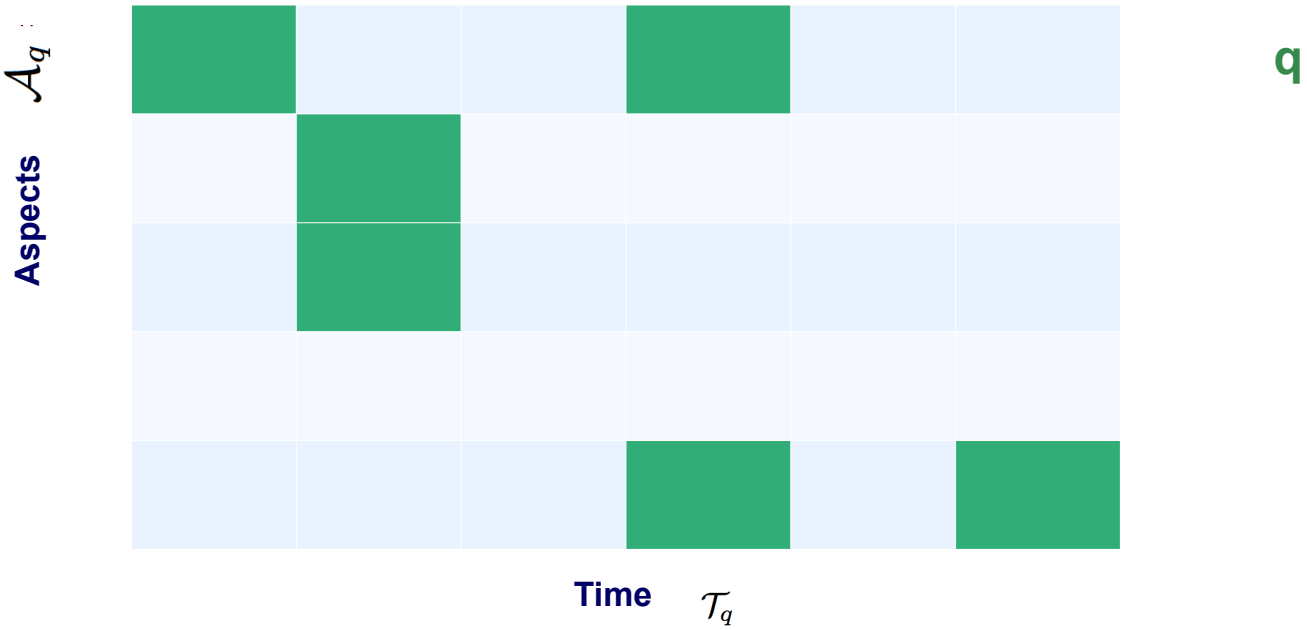
- Aspects are temporal in nature -
  - Aspect utility is updated using an **exponential decay function**.
- Time windows are aspect diverse -
  - Time window utility is updated based on **aspect coverage**
- We represent aspects as **entities** - *David Dinkins, Republican Party, Manhattan*

[Singh et al., CHIIR 2016]

Evaluating Historical Search

- Standard diversity: coverage of aspect space
- Temporal diversity: coverage of time space

- **Historical diversity: coverage of aspect-time space**



- SBR@k using the Aspect-Time Space

Conclusions

- Information needs could be based on
  - Explicit temporal information needs
  - Historical intents
  - Recency and freshness

- Temporal language models leverage temporal expressions along with publication times

- Temporal diversity used for historical intents
- Balancing relevance and recency is best suited for freshness in search result quality

Demo this SIGIR – Try it out

Expedition: A Time-Aware Exploratory Search System Designed for Scholars in SIGIR '16: J. Singh, W. Nejdl, A. Anand.

Demo:
bit.ly/archive-search

Temporal Query Analysis

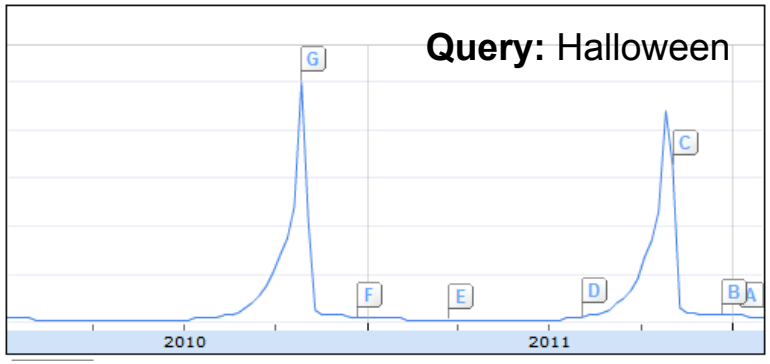
Temporal Queries

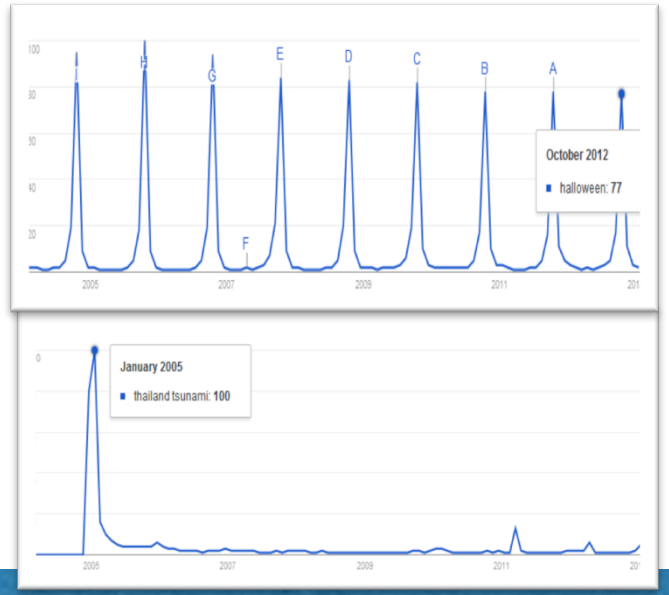
- Temporal information needs exist in both standard test collections, such as, TREC and the Web
  - E.g., users' timelines, web/news archives, or digital libraries
  - Users: social scientist, journalists, historians, or librarians

- Two main classes of temporal queries:
  - 1) Temporal search patterns observable in query streams
  - 2) No temporal search patterns, but relevance is time-dependent

Temporal Patterns in Query Streams

1. Periodic (weekly/monthly) and seasonal queries
   - *E.g., recurring and annual events*
2. Trending queries
   - *E.g., anticipated and ongoing events*
3. Sporadic or spiky queries
   - *E.g., breaking news, celebrities, and unplanned events*





Google Insights for Search
http://www.google.com/insights/search/

Temporal Patterns in QRel

Recency query

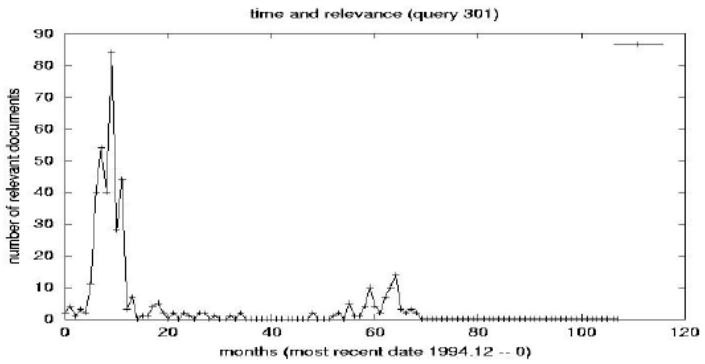


Figure 2.2: Query 301 "International Organized Crime" – A "recency" query.

Time-sensitive query

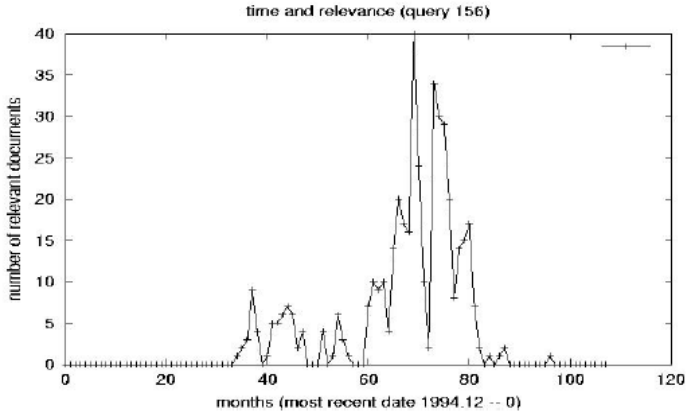


Figure 2.3: Query 156 "Efforts to Enact Gun Control Legislation"- Relevant documents mostly in the past.

Time-insensitive query

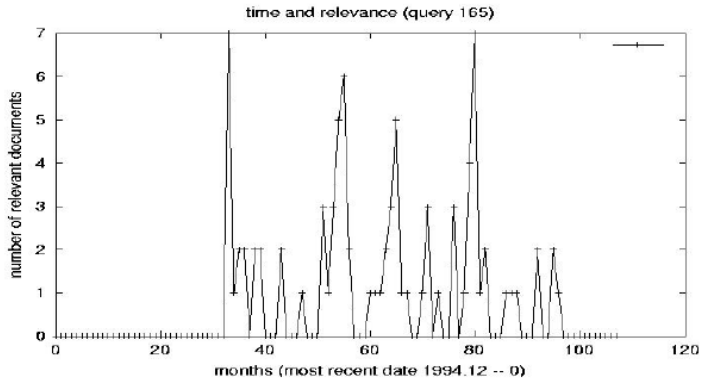


Figure 2.4: Query 165 "Tobacco Company Advertising and the Young" - More uniform distribution

[Li and Croft, CIKM 2003]

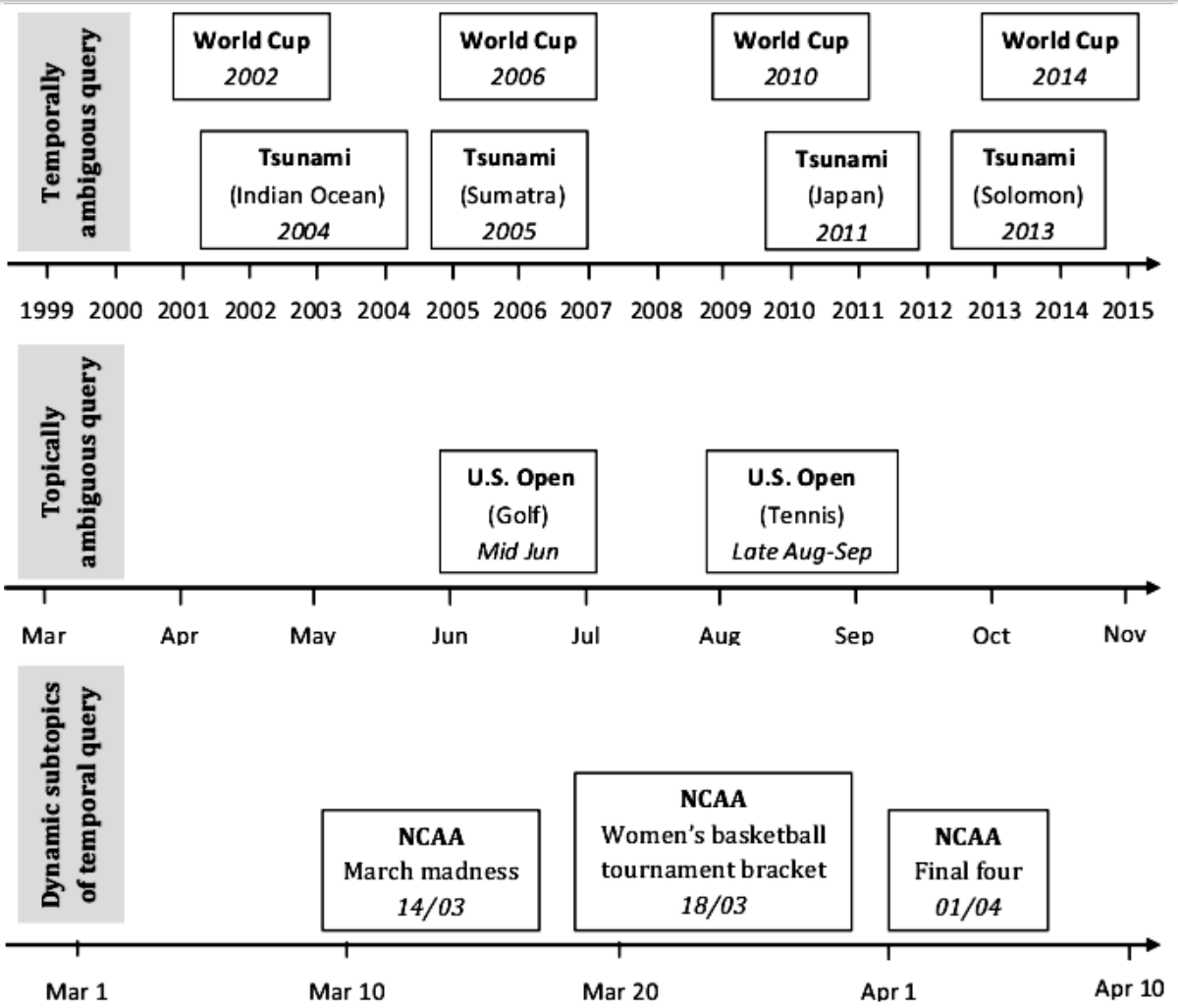
Time Constraints of Queries

Temporal query = *term* + *time*

1. Explicit queries: *time* is provided, "Presidential election 2016"
2. Implicit queries: *time* is *not* provided, "Brazil World Cup"
   - Temporal intent can be implicitly inferred
   - I.e., can refer to FIFA World Cup events in 2014 or 1950

- Previous studies show a significant fraction of temporal queries
  - 1.5% of web queries are *explicit* [Nunes et al., ECIR 2008]
  - 7% of web queries are *implicit* [Metzler et al., SIGIR 2009]
  - 13.8% of web queries contain *explicit* time; 17.1% of queries have implicit temporal intent [Zhang et al., EMNLP 2010]

Temporal Query Dynamics



[Kanhabua et al., FnTIR 2015]

Overview of Research Topics

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

2. Dynamic Query Subtopics

2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

3. Query Enhancement

3.1 Temporal Relevance Feedback
3.2 Time-aware Query Reformulation

Overview of Research Topics

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

2. Dynamic Query Subtopics

2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

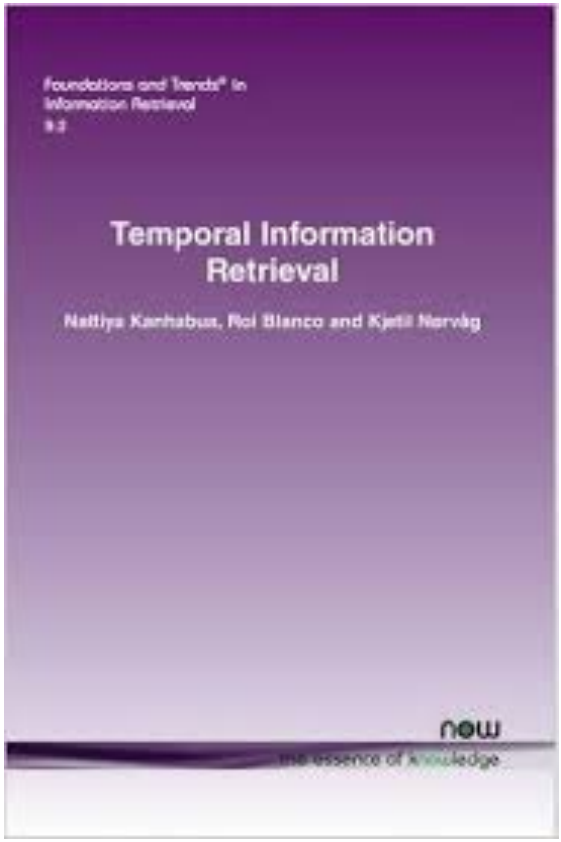
3. Query Enhancement

Further readings

3.1 Temporal Relevance Feedback
3.2 Temporal Query Performance Prediction
3.3 Time-aware Query Reformulation

Additional Resource

- Book: **Temporal Information Retrieval**
  - Authors: N. Kanhabua, R. Blanco, and K. Nørvåg
  - Foundations and Trends® in Information Retrieval
  - Volume 9, Issue 2, pp 91-208, 2015
  - Freely available: https://goo.gl/DUiw5R
  - Download from the authors' home pages



Overview of Research Topics

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

2. Dynamic Query Subtopics

2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

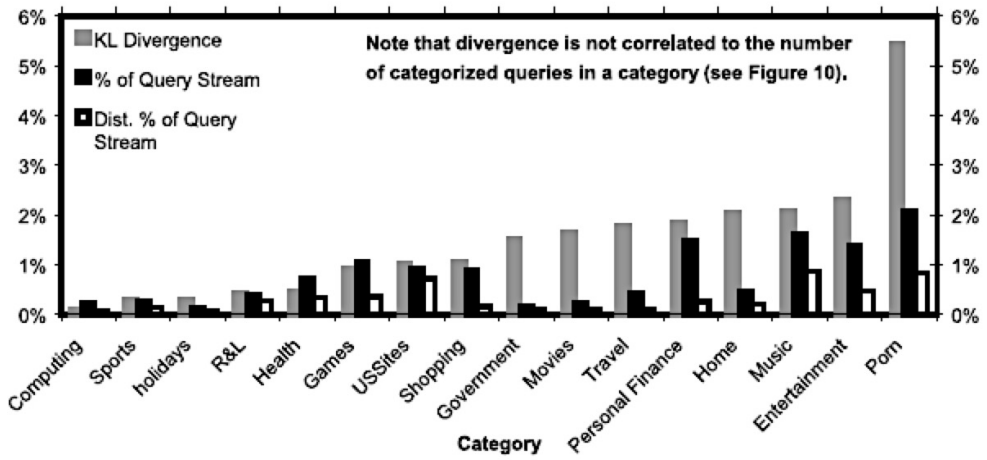
Analyzing Changes in Query Popularity

- Measuring query fluctuations over time using KL-divergence

$$KL(P(q|t)||P(q|c,t)) = \sum_q P(q|t) \times \log \frac{P(q|t)}{P(q|c,t)},$$

  - Non-symmetric measure of the difference of two distributions
  - Computed over hours and compared among query categories

Intuitive, but *unable* to model and detect popularity changes



[Beitzel et al., SIGIR 2004]

Overview of Research Topics

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

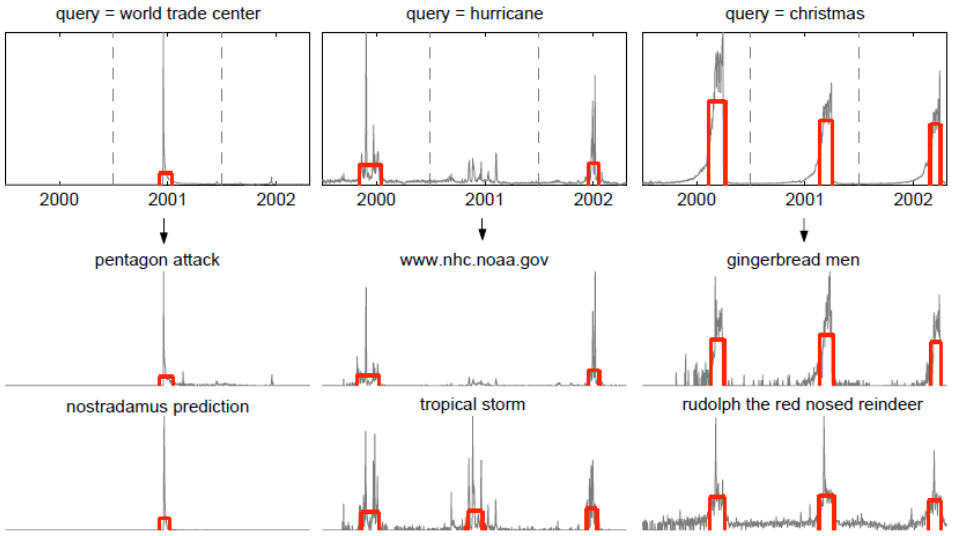
1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

2. Dynamic Query Subtopics

2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

Detecting and Categorizing Temporal Queries

- Efficient and effective method for detecting (short, long) bursts
- Classifying queries based on burst shapes and duration
  1) *Periodic*: events happening in a weekly basis
  2) *Seasonal*: events recurring monthly or yearly
  3) *large peak*: unplanned events or breaking news
- Identifying queries with similar patterns, called query-by-burst



[Vlachos et al., SIGMOD 2004]

Detecting and Categorizing Temporal Queries

- Temporal query categorization based on two dimensions:
  1. Popularity changes
     - Number of spikes
     - Shape of spikes
     - Query periodicity
     - Overall trend
  2. Content changes
     - Query-dependent (TF-IDF)
     - Query-independent (Dice)

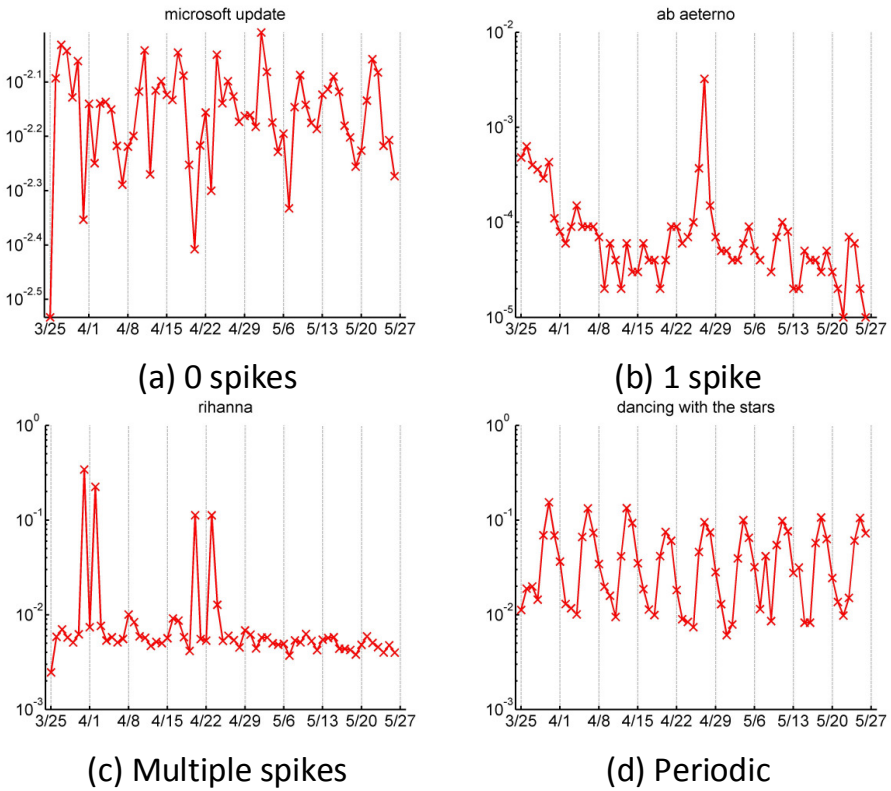


Figure 1. Different queries had different numbers of spikes in query popularity during the study period.

[Adar et al., WSDM 2009]
[Kulkarni et al., WSDM 2011]

Detecting and Categorizing Temporal Queries

- Temporal query categorization based on two dimensions:
  1. Popularity changes
     - Number of spikes
     - Shape of spikes
     - Query periodicity
     - Overall trend
  2. Content changes
     - Query-dependent (TF-IDF)
     - Query-independent (Dice)

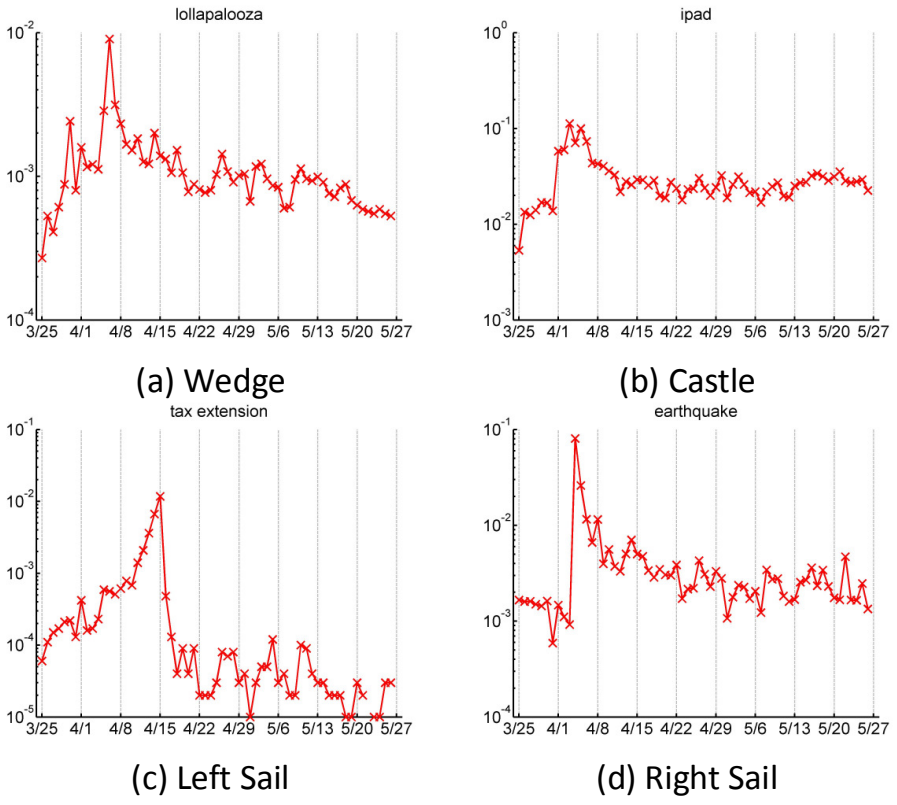
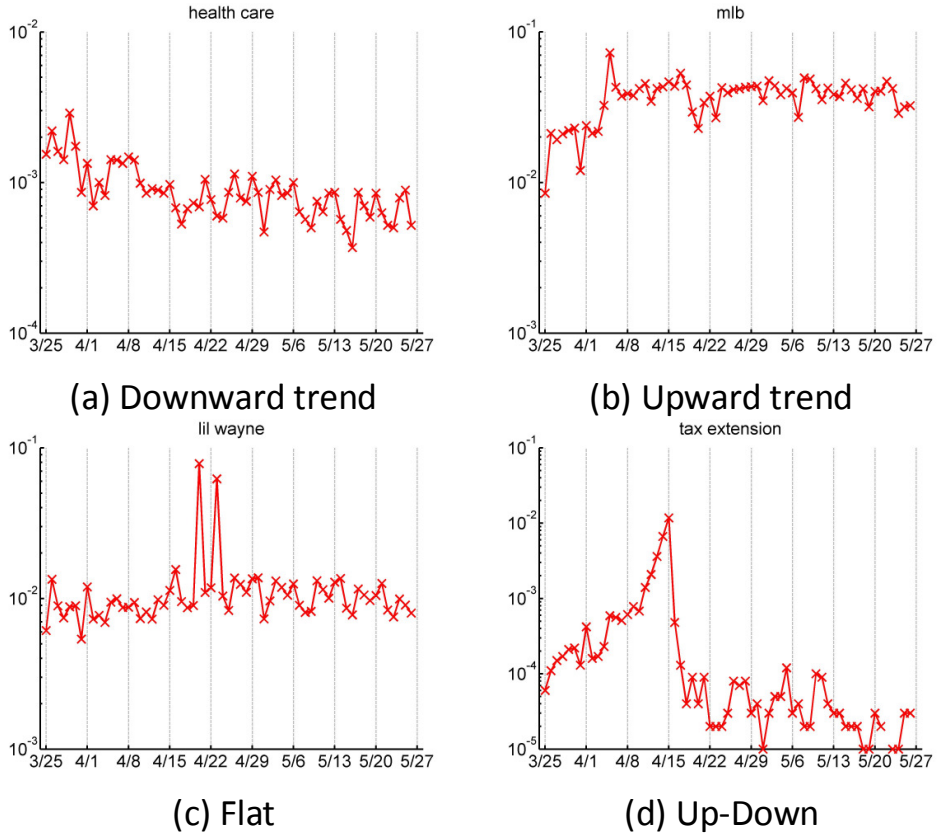


Figure 2. When a query spiked in popularity, the spike could occur in a variety of different shapes.

[Adar et al., WSDM 2009]
[Kulkarni et al., WSDM 2011]

Detecting and Categorizing Temporal Queries

- Temporal query categorization based on two dimensions:
  1. Popularity changes
     - Number of spikes
     - Shape of spikes
     - Query periodicity
     - Overall trend
  2. Content changes
     - Query-dependent (TF-ID
     - Query-independent (Dice



(a) Downward trend (b) Upward trend

(c) Flat (d) Up-Down

Figure 3. Different trends in query popularity. (y-axis in log scale)

[Adar et al., WSDM 2009]
[Kulkarni et al., WSDM 2011]

Detecting and Categorizing Temporal Queries

- Temporal query categorization based on two dimensions:
  1. Popularity changes
     - Number of spikes
     - Shape of spikes
     - Query periodicity
     - Overall trend
  2. Content changes
     - Query-dependent (TF-IDF)
     - Query-independent (Dice)

$$Dice(W_i, W_j) = 2\frac{W_i \cap W_j}{|W_i| + |W_j|},$$

[Adar et al., WSDM 2009]
[Kulkarni et al., WSDM 2011]

Key Findings

- *High popularity change, Low content change.* Annual events like Easter (hard boiled eggs) or April 15 US tax day (taxes online)

- *High popularity change, High content change.* Ongoing events, e.g., the query mlb, exhibited a lot of change in both measurements because the baseball season began

- *Periodic/aperiodic, High content change.* Periodic queries (e.g., TV shows or sports events) change more than aperiodic queries (such as, celebrity queries)

Overview of Research Topics

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

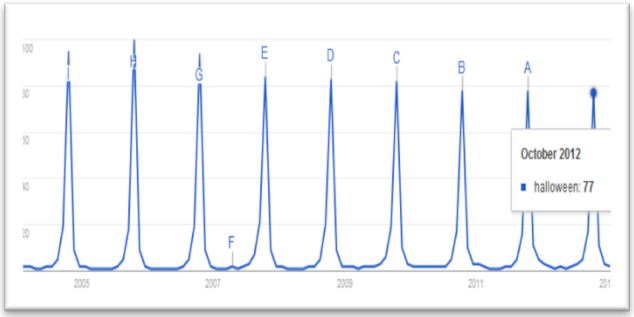
1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

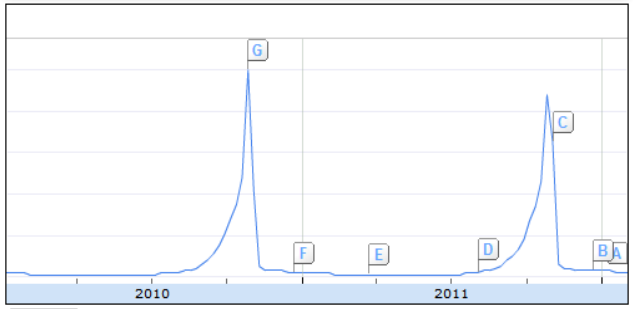
2. Dynamic Query Subtopics

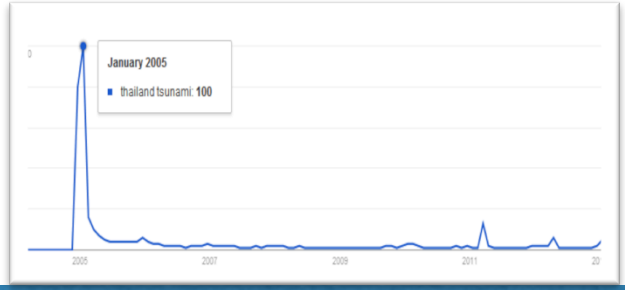
2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

Modeling and Predicting Changes in Query Popularity

- Using time-series analyses to capture temporal dynamics
- Three features extracted from time-series data
  1. Seasonality
  2. Trend
  3. Surprise







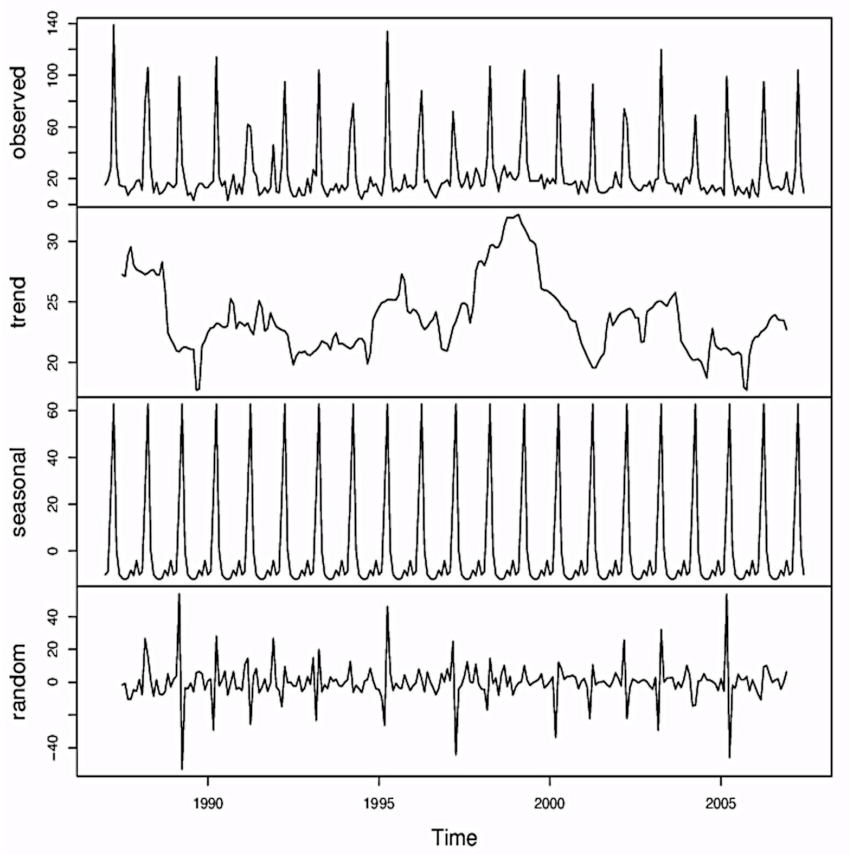
[Diaz and Jones, SIGIR 2004]
[Shokouhi, SIGIR 2011]
[Radinsky et al., WWW 2012]

Seasonality

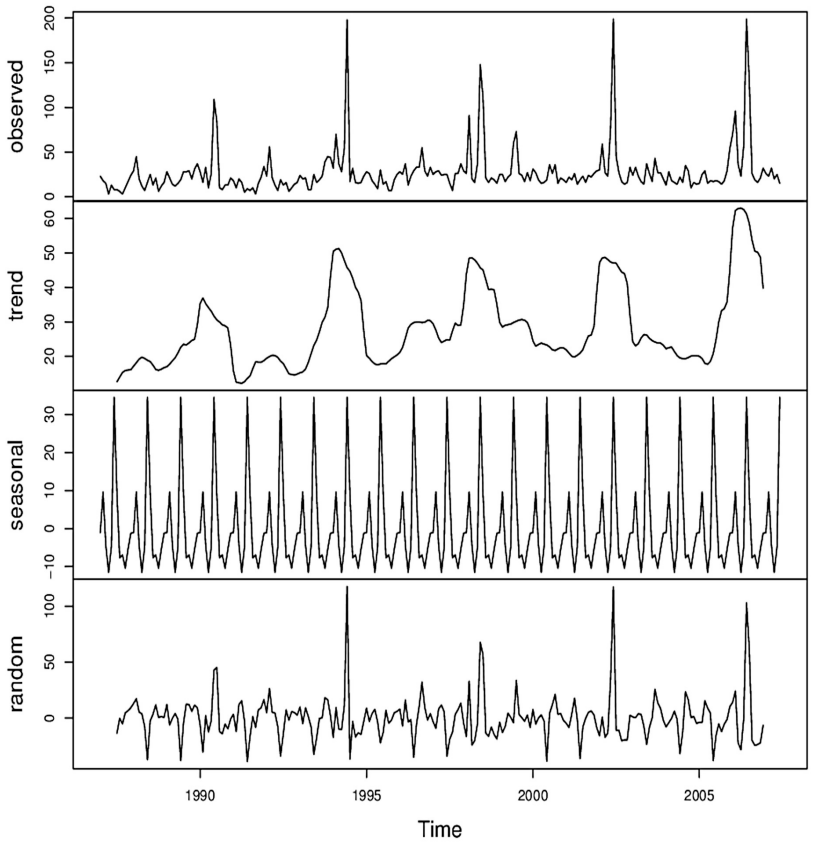
- Detecting seasonal queries [Shokouhi, SIGIR 2011]
  - Annual events (e.g., US Open and Easter)
  - Recurring events (e.g., FIFA World Cup and Olympic Games)

- Time-series decomposition based on Holt-Winters adaptive exponential smoothing
  - Input: time-series data, Y
  - Output: 3 main components (level L, trend T and seasonal S)

Time-series Decomposition



Query: Easter



Query: World cup

Seasonality Score

- Compute a cosine similarity as *seasonality*
  - Y is the original time-series data
  - S is the seasonal component

$$Seasonality(Y, S) = \frac{\vec{Y}\vec{S}}{\| \vec{Y} \| \cdot \| \vec{S} \|}$$

Autocorrelation

- Detecting *trends* or *emerging events* by their predictability
- Autocorrelation, *a cross correlation* with itself or between its past and future values at different time lags

$$Autocorrelation(Y, l) = \frac{\sum_{i=1}^{N-k} (y_i - \bar{y})(y_{i+l} - \bar{y})}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

  - where *lag*=1, shifting the second time series by one day, called first-order autocorrelation
  - The stronger inter-day dependencies, the higher value for autocorrelation

Surprise

- Detecting *unseen events* or *surprisingly popular* queries [Radinsky et al., WWW 2012]
  - Assume an unplanned event happening when there is a significant prediction error
  - Compute the sum of squared errors of prediction (SSE) using a simple linear regression model

$$SSE(Y, \hat{Y}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Overview of Research Topics

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

2. Dynamic Query Subtopics

2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

Analyzing Top-k Search Results

- When no long-term query logs available, temporal query intent can be determined by analyzing search results

- [Diaz and Jones, SIGIR 2004] introduced *temporal profiles*, referring to the temporal characteristics of a query

$$P(t|q) = \sum_{d \in D_q} P(t|d) \frac{P(q|d)}{\sum_{d' \in D_q} P(q|d')},$$

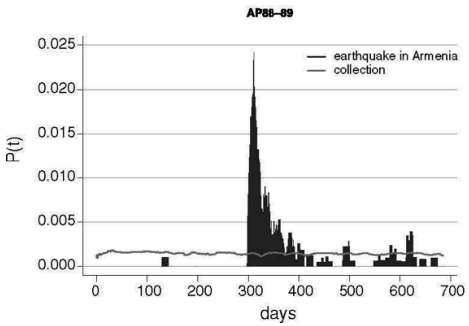
  - where P(t|q) is the probability of a publication date t given q
  - P(q|d) is a document relevance score obtained by *relevance language modeling* [Lavrenko and Croft, SIGIR 2001]

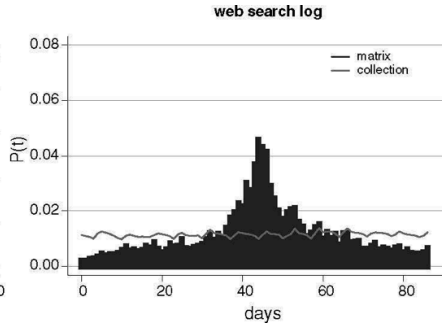
$$P(t|d) = \begin{cases} 0 & \text{if } PubTime(d) \neq t, \\ 1 & \text{if } PubTime(d) = t. \end{cases}$$

Temporal Profiles of Queries

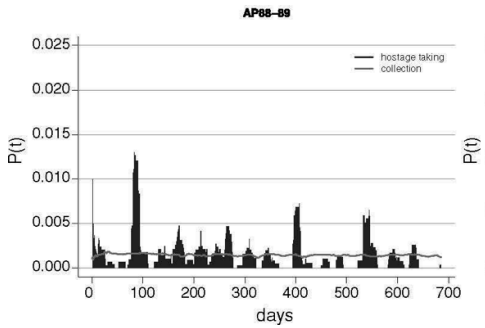
- Three class of queries categorized by temporal profiles

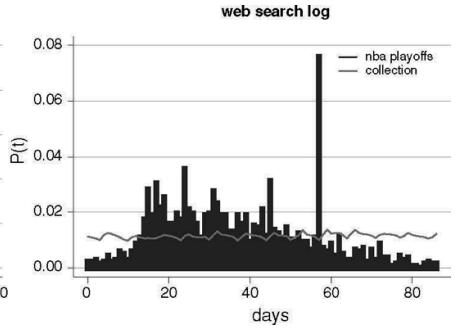
Temporally unambiguous



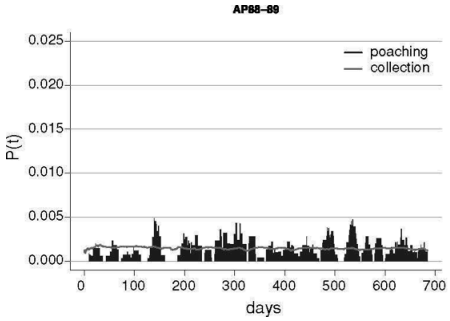


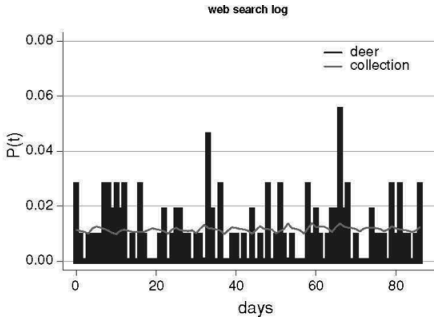
Temporally ambiguous





Atemporal





[Diaz and Jones, SIGIR 2004]

Overview of Research Topics

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

2. Dynamic Query Subtopics

2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

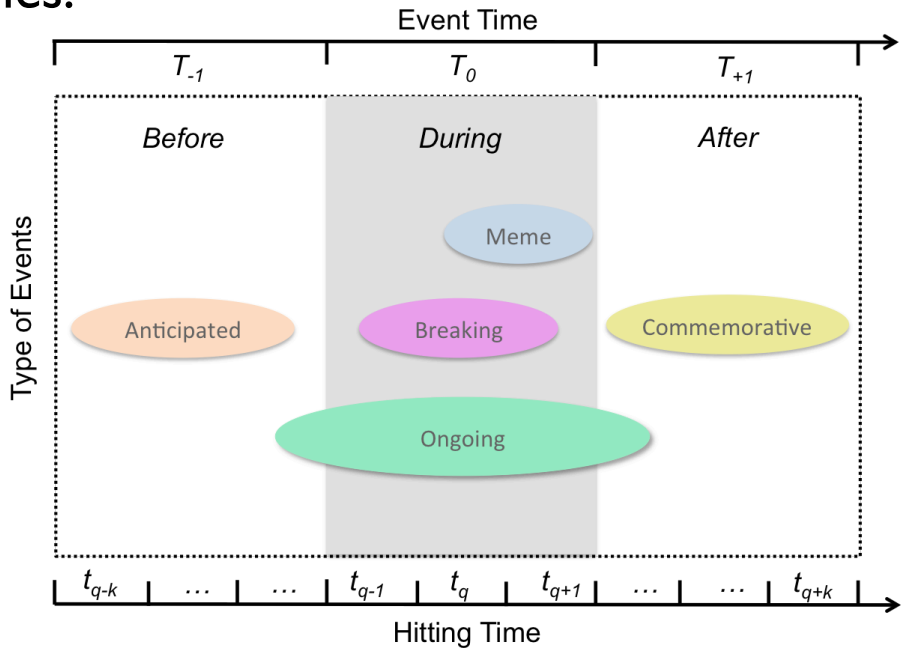
Learning to Classify Temporal Queries

- Using Temporal Profiles of Queries for Precision Prediction [Diaz and Jones, SIGIR 2004]
  - Three classes: temporally unambiguous, temporally ambiguous, and atemporal
  - Method: state-of-the-art classification algorithms
  - Features: temporal profiles and other features (temporal KL-divergence, autocorrelation, kurtosis, burst-related features)

- Temporalia NTCIR-11 [Joho et al., TempWeb 2014]
  - Temporal Query Intent Categorization (TQIC) Task
  - Four classes: past, recency, future and atemporal
  - Participants: 6 teams (17 formal runs)

Learning Dynamic Classes of Event-related Queries

- Taking into account *hitting time* and *event aspects*
- Novel taxonomy of temporal queries:
  1. Anticipated
  2. Breaking
  3. Commemorative
  4. Meme
  5. Ongoing
  6. Atemporal



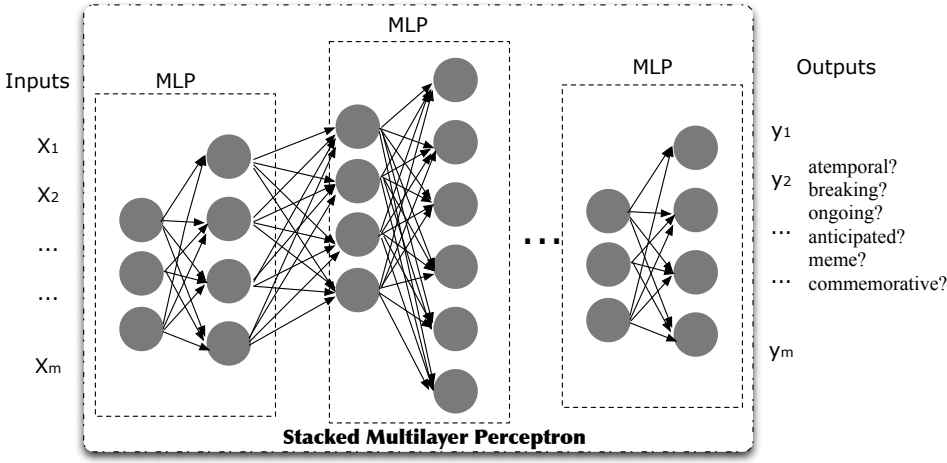
- Method: a novel deep learning model, called Stacked Multilayer Perceptron Networks (S-MLPs)

[Kanhabua et al., Neu-IR 2016]

Stacked Multilayer Perceptron Networks (S-MLP)

- Deep learning architecture
  - Use a multi-layer perceptron as a basic learning unit
  - Assemble stacked units to learn *complex relationships*
  - Handle *imbalanced data* from various dynamic classes

- Multi-class classification problem
  - Input: Xi is a feature vector
  - Output: Yi∈{0,..., m} is a label
  - Goal: to minimize δ(Yi, yi)



[Kanhabua et al., Neu-IR 2016]

Classification Features

Table 1: List of features used in our event-related query classification task: *long-span* denoting features obtained from a long time-span temporal document collection, and *short-span* referring features from a query log with a short time-span.

Feature	Description	Feature	Description
long_span_acf	long-span autocorrelation	*short_span_acf*	short-span autocorrelation
long_span_seasonal	long-span seasonality	*short_span_seasonal*	short-span seasonality
long_span_kurtosis	long-span kurtosis	*short_span_kurtosis*	short-span kurtosis
long_span_KL_PT	long-span KL divergence	*prediction_sse*	prediction error
burstLength	longest burst duration	*t_scope*	trending scope
burstWeight	maximum burst weight	*t_level*	trending amplitude
noOfBursts	number of bursts	*avgFreq*	average frequency
isPer	if a query contains person entities	*maxFreq*	maximum frequency
isLoc	if a query contains location entities	*CElong*	click entropy for 14 days
isOrg	if a query contains organization entities	*CEshort*	click entropy for 3 days
isTempEx	if a query contains temporal expressions	*CEper*	ratio of CEshort to CElong
noOfQueries	number of queries in a query cluster (C)	*sumCFreq*	sum of query frequency in C
burstDistM	distance from the max burst	*avgCFreq*	average of query frequency in C
burstDistL	distance from the longest duration burst	*maxCFreq*	maximum of query frequency in C

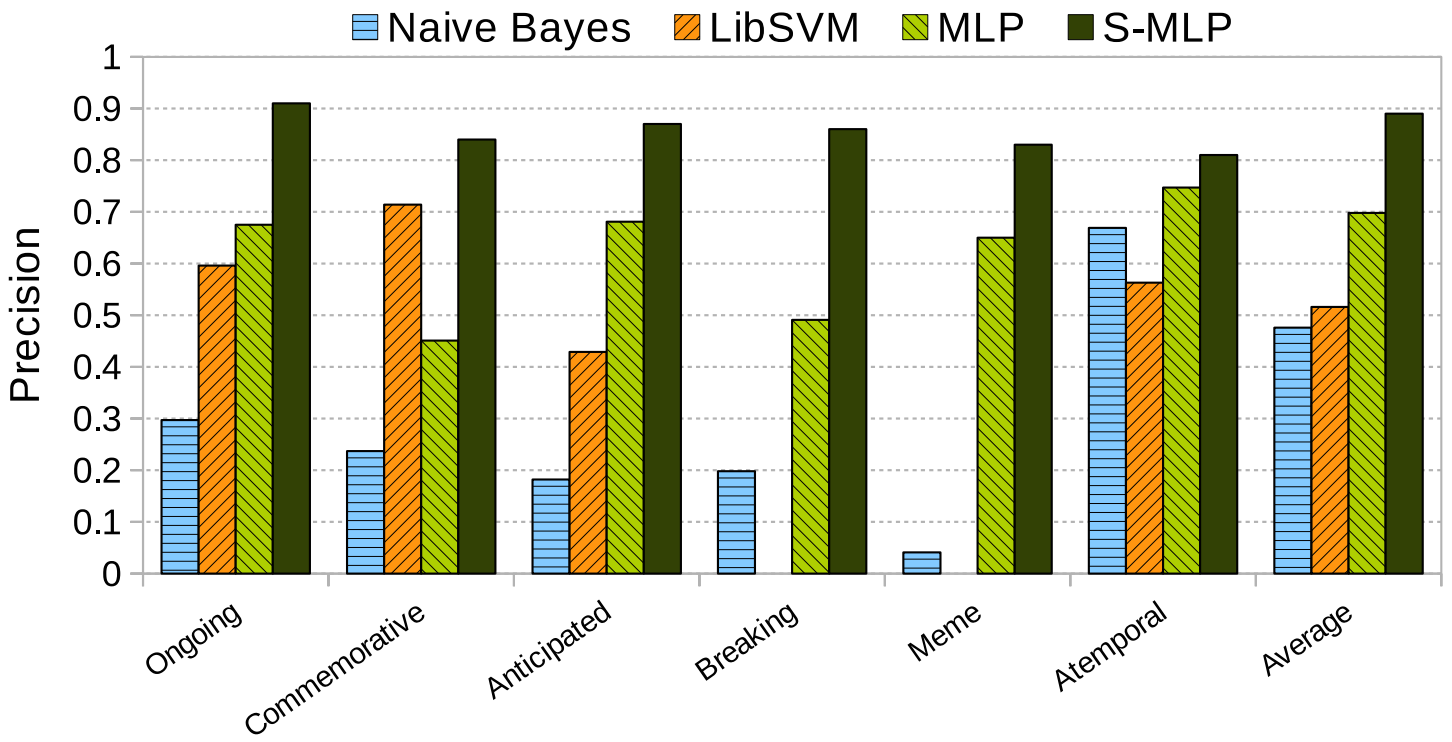
[Kanhabua et al., Neu-IR 2016]

Experiments

- Query logs:
  - AOL: 30M queries March 1 - May 31, 2006
  - MSN: 15M queries from May 2006
- Temporal collection:
  - The New York Times Annotated Corpus
  - 1.8M documents from 1987 – 2007
- Relevant judgement:
  - Label(q,Te,tq), a triple of a query q, an event date Te, and hitting time tq,
  - Totally 10,370 triples labelled (988 of anticipated, 531 of breaking, 304 of commemorative, 315 of meme, 2,520 of ongoing, and 5,712 of atemporal)

[Kanhabua et al., Neu-IR 2016]

Results



[Kanhabua et al., Neu-IR 2016]

Overview of Research Topics

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

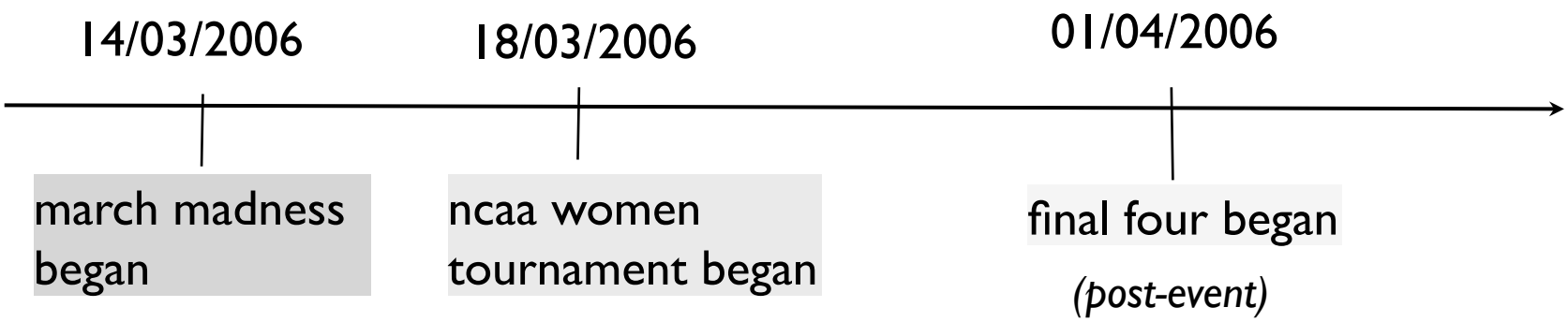
2. Dynamic Query Subtopics

2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

Dynamic Query Subtopics

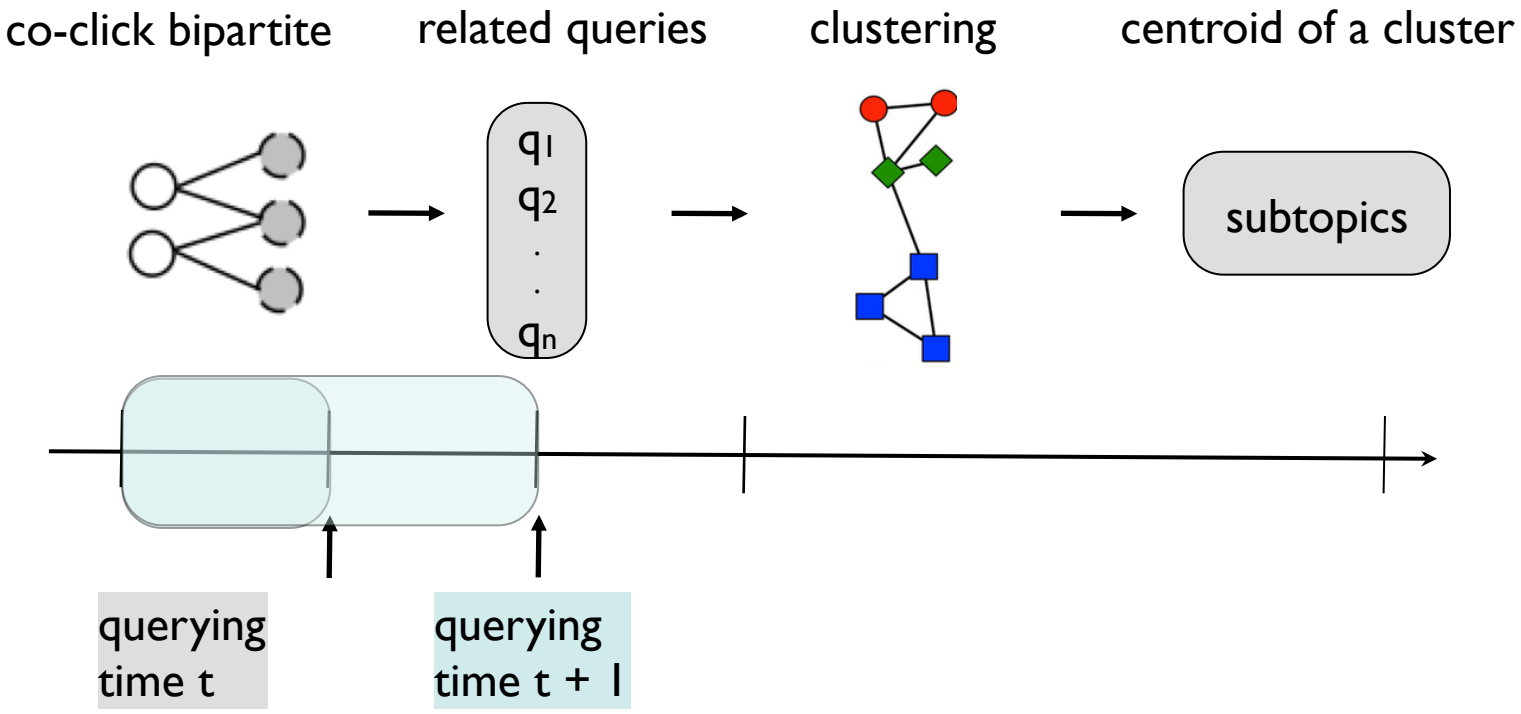
- Event-driven queries contain highly variable subtopics
  - Impact on search, e.g., ranking and result diversification
  - Methods: extracting Wikipedia (section hierarchy), query logs, and external document collections

- Examples of dynamic subtopics for a given query: *ncaa*



[Whiting et al., SIGIR 2013]
[Nguyen et al., ECIR 2014]

Mining Subtopics from Historical Query Logs



[Nguyen et al., ECIR 2014]

Dynamic Subtopics Mined from Query Logs

Table 4.2: Top-10 dynamic query subtopics ordered by weights for the query ncaa.

March 14		March 31		April 07	
subtopic c	$w(c)$	subtopic c	$w(c)$	subtopic c	$w(c)$
march madness schedule	0.0132	oakland raiders	0.0100	*ncaa women's basketball tournament*	0.0122
ncaa basketball tournament	0.0117	ncaaw	0.0090	ncaa basketball tournament	0.0053
nfl draft	0.0068	tito francona	0.0042	cbs sports line	0.0049
selection sunday	0.0048	ncaa brackets	0.0031	*ncaaw*	0.0033
oakland raiders	0.0037	ncaa division ii	0.0029	*ncaa final four*	0.0031
2006 ncaa tournament bracket	0.0032	andy goram	0.0024	ncaa wrestling	0.0029
brad hopkins released nfl	0.0026	lakers	0.0024	march madness bracket	0.0028
roger clemens	0.0023	*ncaa women's basketball tournament bracket*	0.0024	*ncaa basketball results*	0.0019
ncaa division ii	0.0021	ncaa basketball brackets	0.0021	andy goram	0.0009
college basketball	0.0014	nit brackets	0.0021	ncaa division ii	0.0009

[Nguyen et al., ECIR 2014]

Mining Subtopics from a Document Collection

- Using Latent Dirichlet Allocation (LDA) [Blei et al., *J. Mach. Learn. Res.* 2003]
  - An unsupervised model of latent query subtopics
  - Directly integrate into probabilistic subtopic models

- TREC Blog08 Collection
  - High-quality data with aspects underlying a target collection
  - Input: a set of relevant documents at different time

[Nguyen et al., ECIR 2014]

Dynamic Subtopics Mined from a Document Collection

Table 4.3: Subtopics associated to the query *apple* using LDA-based subtopic mining from TREC Blog08.

IPHONE		MACBOOK		FRUIT		PIE	
word w	$P(w\|c)$	word w	$P(w\|c)$	word w	$P(w\|c)$	word w	$P(w\|c)$
iphone	0.652	macbook	0.999	apple	0.397	apple	0.480
software	0.276	apple	0.511	tree	0.192	pie	0.220
developers	0.220	pro	0.404	trees	0.118	butter	0.185
app	0.212	nvidia	0.280	fruit	0.110	recipe	0.181
nda	0.192	graphics	0.252	garden	0.092	sugar	0.134
store	0.143	air	0.137	varieties	0.056	juice	0.125
application	0.120	ghz	0.127	growing	0.045	cup	0.116

[Nguyen et al., ECIR 2014]

Summary

1. Temporal Query Intent

1.1 Mining Temporal Patterns in Query Streams
- Analyzing Changes in Query Popularity
- Detecting and Categorizing Temporal Queries
- Modeling and Predicting Popularity Changes

1.2 Analyzing Top-k Search Results
- Learning to Classify Temporal Queries
- Determining Relevant Time for Queries

2. Dynamic Query Subtopics

2.1 Mining Subtopics from Query Logs
2.2 Mining Subtopics from Documents

Applications of Temporal IR

Temporal IR Users

- **Big data analysts and social informatics researchers**, who want to enhance their algorithms to deal with social data, gain multi-disciplinary research skills, harmonise existing data and analytics infrastructures, and engage other research communities in the development of these key enabling technologies for the future digital economy and society
- **Economists, social science and humanities researchers, journalists, policy and law makers**, who have to analyse the avalanche of (big) social data, in order to gain insight and actionable knowledge
- **Researchers in related communities**, who would like to use the algorithms, the analytical competences and data infrastructure
- **Industrial innovators & startuppers**, who would like to create rapid proof-of-concepts of data-driven innovative ideas and services
- **The public as a whole**, who would like to understand their role in the production, consumption and value-creating of social data

Application Areas

(1)
Web Archive Search
News Archive Search

(2)
Temporal Analytics and Exploration

(3)
Temporal Clustering and Summarization

(4)
Search the Past
Future Event Retrieval

Application Areas

(1)
Web Archive Search
News Archive Search

(2)
Temporal Analytics and Exploration

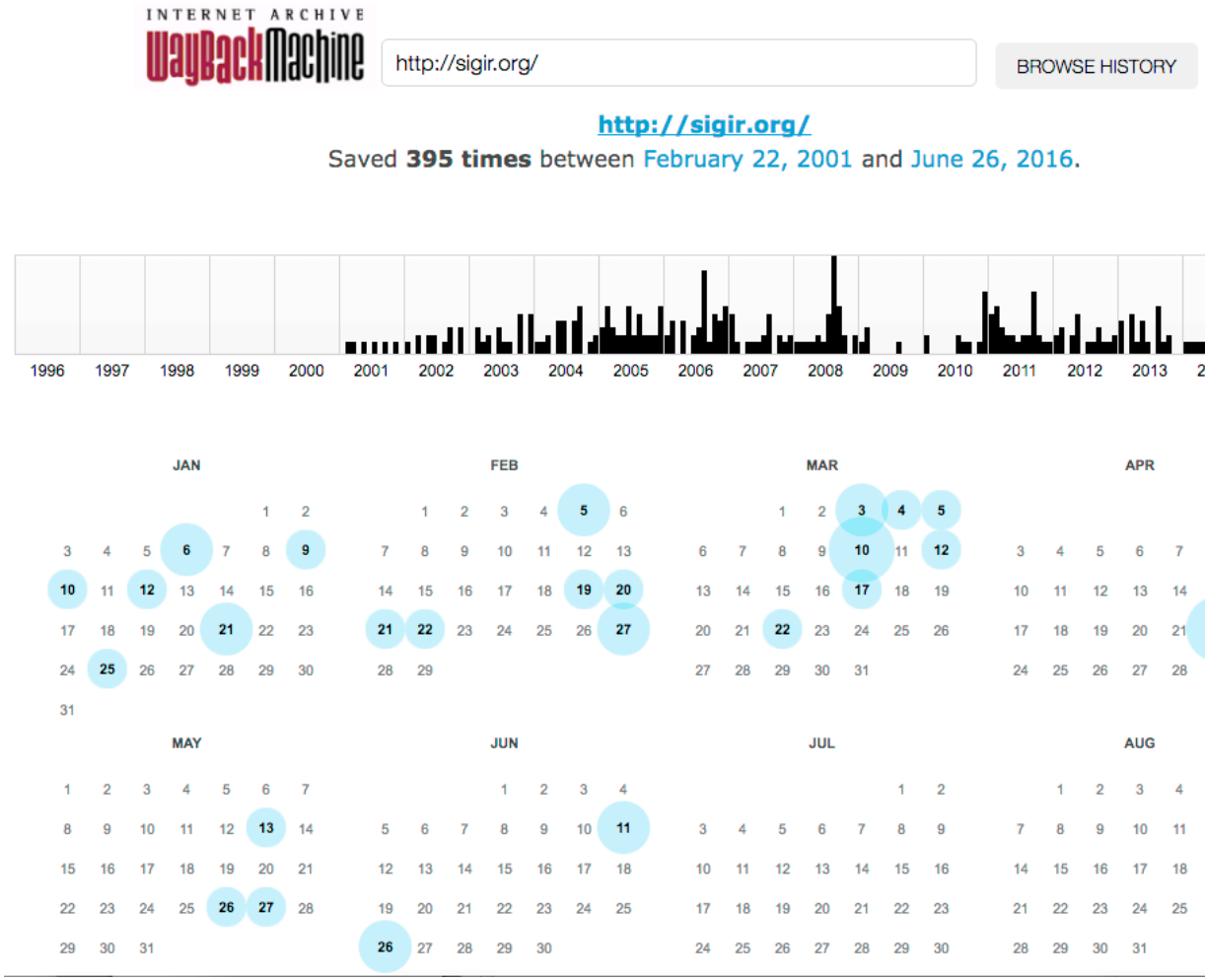
(3)
Temporal Clustering and Summarization

(4)
Search the Past
Future Event Retrieval

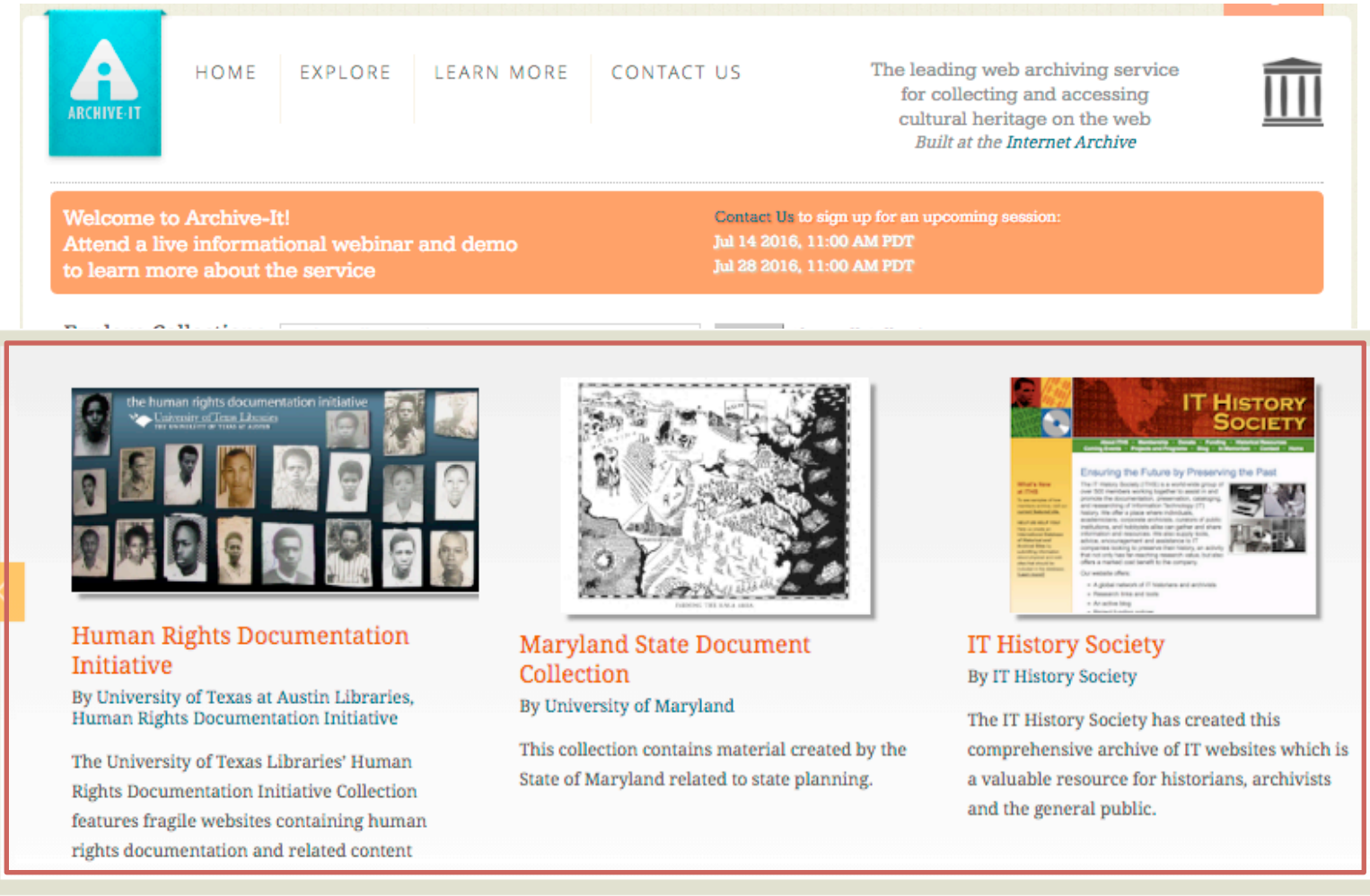
Web Archives as Scholarly Source

- Users from various disciplines:
  - Social scientists and political scientists
  - Historians, librarians, and journalists
- Search intent:
  - Complex, information seeking behaviors
  - Required creating subsets of a collection
- Existing archive search engines
  - WayBack Machine, and Archive-IT
  - Google news archive
  - ALEXANDRIA entity-oriented archive search

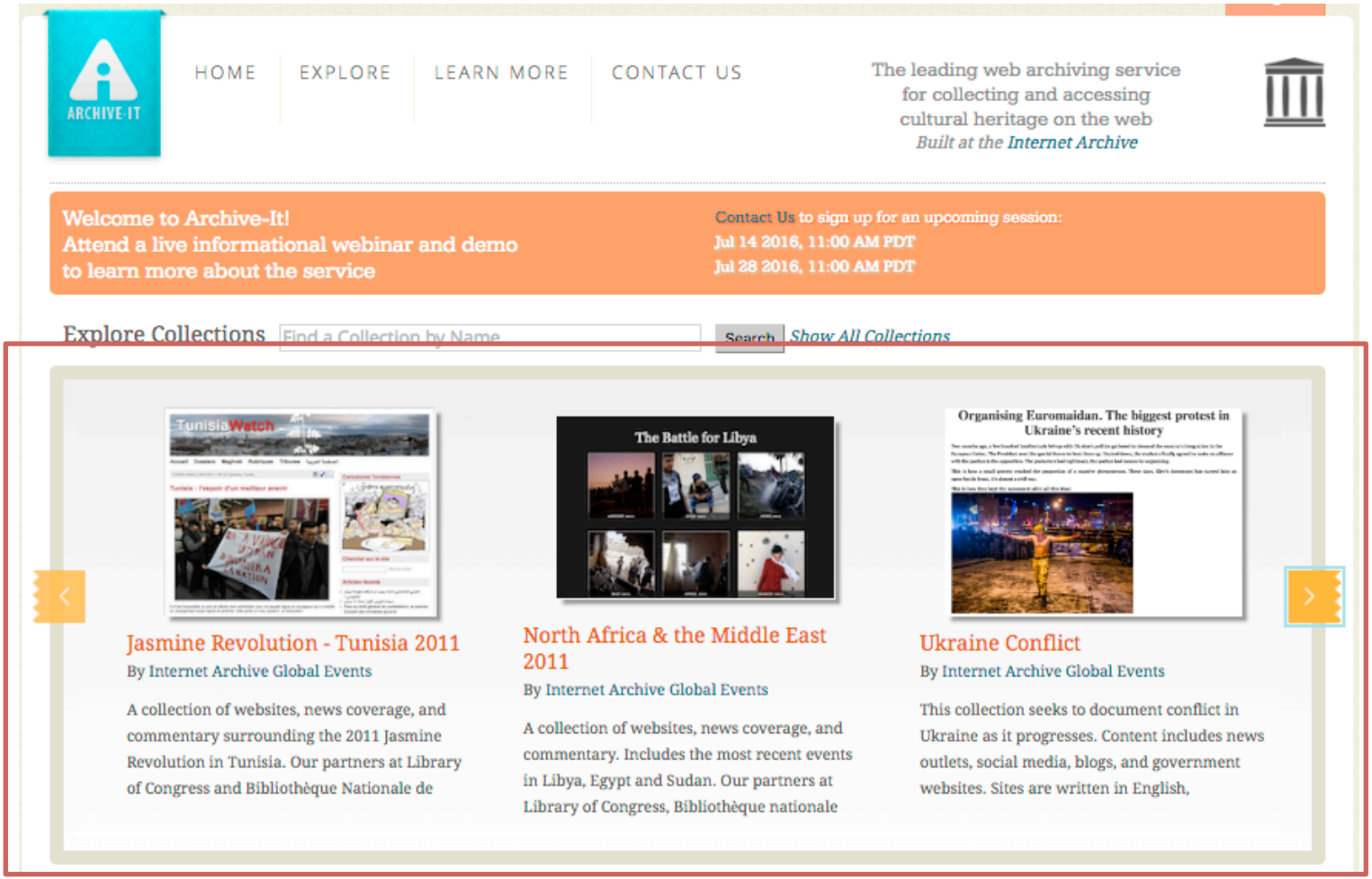
WayBack Machine



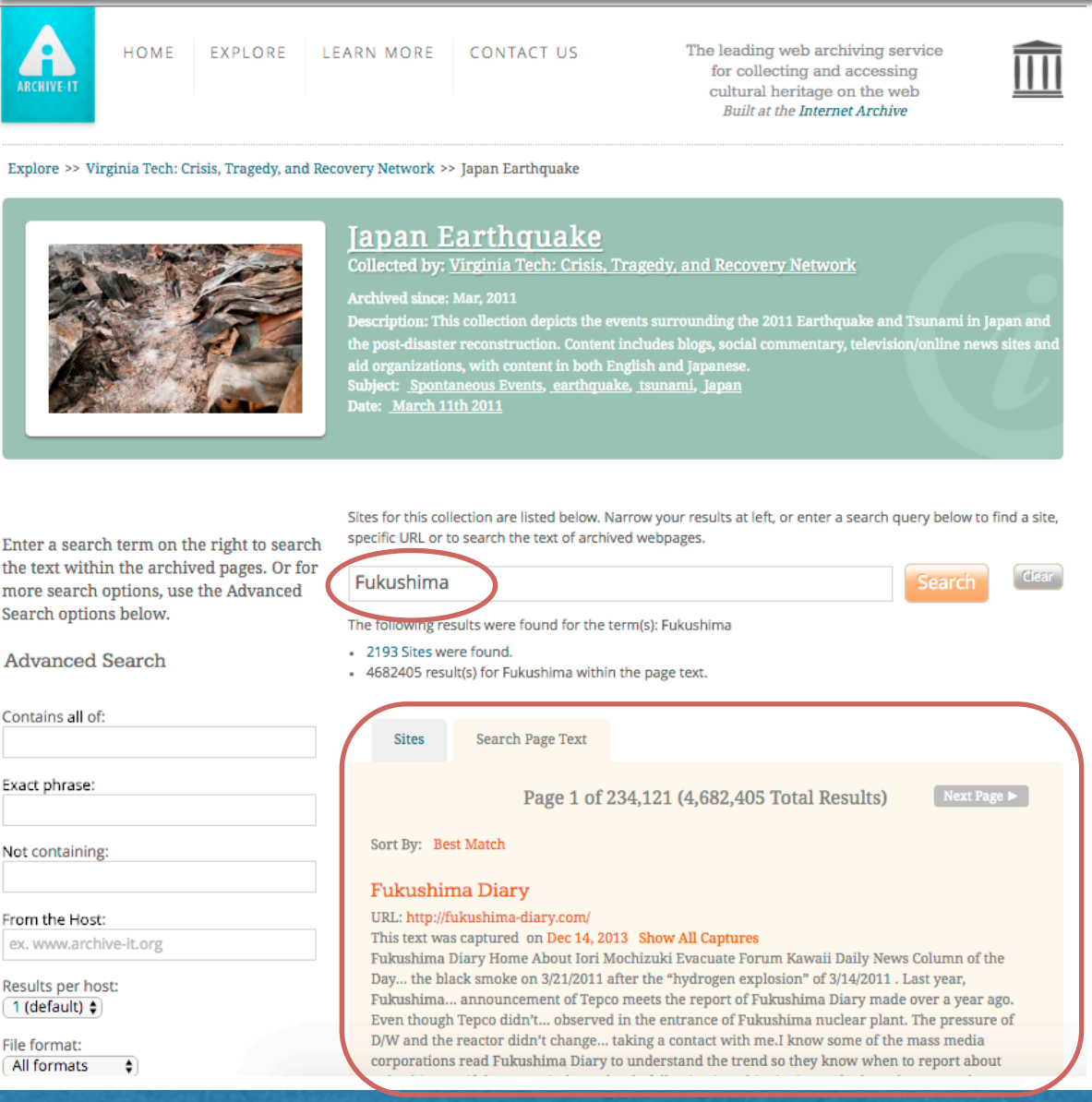
Archive-IT


ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US
The leading web archiving service for collecting and accessing cultural heritage on the web
Built at the Internet Archive
Welcome to Archive-It!
Attend a live informational webinar and demo to learn more about the service
Contact Us to sign up for an upcoming session:
Jul 14 2016, 11:00 AM PDT
Jul 28 2016, 11:00 AM PDT
the human rights documentation initiative
University of Texas Libraries
Human Rights Documentation Initiative
By University of Texas at Austin Libraries, Human Rights Documentation Initiative
The University of Texas Libraries' Human Rights Documentation Initiative Collection features fragile websites containing human rights documentation and related content
Maryland State Document Collection
By University of Maryland
This collection contains material created by the State of Maryland related to state planning.
IT History Society
Ensuring the Future by Preserving the Past
IT History Society
By IT History Society
The IT History Society has created this comprehensive archive of IT websites which is a valuable resource for historians, archivists and the general public.

Archive-IT


ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US
The leading web archiving service for collecting and accessing cultural heritage on the web
Built at the Internet Archive
Welcome to Archive-It!
Attend a live informational webinar and demo to learn more about the service
Contact Us to sign up for an upcoming session:
Jul 14 2016, 11:00 AM PDT
Jul 28 2016, 11:00 AM PDT
Explore Collections
Find a Collection by Name
Search
Show All Collections
Jasmine Revolution - Tunisia 2011
By Internet Archive Global Events
A collection of websites, news coverage, and commentary surrounding the 2011 Jasmine Revolution in Tunisia. Our partners at Library of Congress and Bibliothèque Nationale de
The Battle for Libya
North Africa & the Middle East 2011
By Internet Archive Global Events
A collection of websites, news coverage, and commentary. Includes the most recent events in Libya, Egypt and Sudan. Our partners at Library of Congress, Bibliothèque nationale
Organising Euromaidan. The biggest protest in Ukraine's recent history
Ukraine Conflict
By Internet Archive Global Events
This collection seeks to document conflict in Ukraine as it progresses. Content includes news outlets, social media, blogs, and government websites. Sites are written in English,

Archive-IT: Keyword-based Search



Archive-IT: Keyword-based Search

https://wayback.archive-it.org/2438/20131214060311/http://fukushima-dia...

You are viewing an archived web page, collected at the request of Virginia Tech: Crisis, Tragedy, and Recovery Network using Archive-It. This page was captured on 6:03:11 Dec 14, 2013, and is part of the Japan Earthquake collection. The information on this web page may be out of date. See All versions of this archived page. hide

HOME ABOUT IORI MOCHIZUKI EVACUATE FORUM KAWAII

FUKUSHIMA DIARY

WE ARE AGAINST MEDIA BLACK OUT - PLEASE SUPPORT US SO WE MAY INFORM THE WORLD

日本の窮状を世界に。

DAILY NEWS COLUMN OF THE DAY BREAKING NEWS BACKGROUND EVACUATE FORUM KAWAII

Tepco "Reactor3 had black smoke on 3/21/2011 again. Radiation level significantly spiked up."

*Posted by **Mochizuki** on December 13th, 2013 · No Comments*

Note : If you are from the international mass media, Don't read this site before taking a contact with me.

Related to this article.. **Tepco "Major release of fission product observed on 3/20~21/2011" / "Spiked the radiation level in Kanto region"** [URL]

In the report, Tepco also admitted reactor3 had the black smoke on 3/21/2011 after the "hydrogen explosion" of 3/14/2011.

Last year, Fukushima Diary reported the possibility of the second meltdown of reactor3, which may have occurred from 3/21 to 3/22/2013.

(cf, [Reactor3] Possible second meltdown happened 3/21 ~ 3/22/2011 [URL])

This announcement of Tepco meets the report of Fukushima Diary made over a year ago.

Even though Tepco didn't mention the cause of the black smoke, they admitted the "significant" spike of radiation level was observed in the entrance of Fukushima nuclear plant.

The pressure of D/W and the reactor didn't change.

Officially reactor1, 3 and 4 had one explosion for each. However, there is the possibility that

FRIEND US

SUBSCRIBE

FOLLOW US

Search

北極スヴァールバル調査サポート（1オレオ10$、オレオ数はQuantityで変更可能）はこちらです」
目標オレオ数300／現在オレオ数300
Facebook特設グループよりもTumblrをご希望の方はPaypalのメッセージ欄よりお知らせください。

You can edit Quantity

1OREO $10.00 USD

Facebookのリンクか名前をお願いします。

Archive-IT: Keyword-based Search

Japan Earthquake Web Archive (Virginia Tech: Crisis, Tragedy, and Recovery Network)

Enter Web Address: http:// All Take Me Back

Searched for http://fukushima-diary.com/ 80 Results RSS

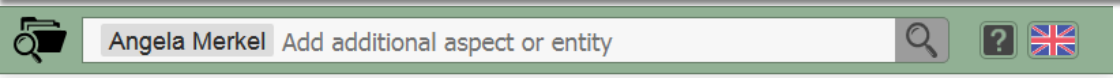
Look up URL in general Internet Archive web collection Proxy Mode Help

* denotes when page was updated

Found 80 Captures between Oct 29, 2011 - Mar 18, 2016

2011	2012	2013	2014	2015	2016
26 pages	15 pages	16 pages	16 pages	6 pages	1 page
Oct 29, 2011 *	Jan 4, 2012 *	Feb 13, 2013 *	Jan 12, 2014 *	Mar 5, 2015 *	Mar 18, 2016 *
Oct 30, 2011 *	Jan 11, 2012 *	Mar 10, 2013 *	Feb 12, 2014 *	Mar 5, 2015	
Oct 31, 2011 *	Jan 18, 2012 *	Mar 11, 2013 *	Feb 13, 2014 *	Mar 5, 2015	
Nov 1, 2011 *	Jan 25, 2012 *	Mar 12, 2013 *	Mar 12, 2014 *	Apr 5, 2015 *	
Nov 2, 2011 *	Feb 1, 2012 *	Mar 13, 2013 *	Apr 12, 2014 *	Apr 5, 2015	
Nov 2, 2011 *	Feb 8, 2012 *	Mar 14, 2013 *	Apr 13, 2014 *	Apr 5, 2015 *	
Nov 3, 2011 *	Feb 15, 2012 *	Mar 15, 2013 *	Apr 18, 2014 *		
Nov 4, 2011 *	Feb 22, 2012 *	Mar 16, 2013 *	May 12, 2014 *		
Nov 5, 2011 *	Mar 7, 2012 *	Apr 13, 2013 *	Jun 12, 2014 *		
Nov 6, 2011 *	Mar 10, 2012 *	Jun 13, 2013 *	Jun 13, 2014 *		
Nov 8, 2011 *	Mar 28, 2012 *	Aug 13, 2013 *	Jul 13, 2014 *		
Nov 9, 2011 *	May 23, 2012 *	Oct 12, 2013 *	Aug 12, 2014 *		
Nov 9, 2011 *	Oct 12, 2012 *	Oct 13, 2013 *	Aug 13, 2014 *		
Nov 10, 2011 *	Oct 20, 2012 *	Nov 12, 2013 *	Sep 12, 2014 *		
Nov 11, 2011 *	Nov 1, 2012 *	Dec 12, 2013 *	Oct 12, 2014 *		
Nov 12, 2011 *		Dec 14, 2013 *	Oct 13, 2014		
Nov 13, 2011 *					
Nov 14, 2011 *					
Nov 15, 2011 *					
Nov 16, 2011 *					
Nov 16, 2011 *					
Nov 23, 2011 *					
Nov 30, 2011 *					
Dec 7, 2011 *					
Dec 15, 2011 *					

Entity-oriented Web Archive Search



Angela Merkel - Wikipedia, the free encyclopedia
Wayback » https://en.wikipedia.org/wiki/Angela_Merkel
captured between 1/23/04 and 7/11/12
Angela Dorothea **Merkel** (née Kasner; born 17 July 1954) is a German politician and former research scientist. **Merkel** has been the Chancellor of Germany since 2005 ...

Angela Merkel - Forbes
Wayback » http://www.forbes.com/profile/angela-merkel/
captured between 10/9/10 and 6/7/16
German Chancellor **Angela Merkel** continues her reign as the most powerful woman on the planet for 10 years running. Why? She clinched a third four-year term of Europe ...

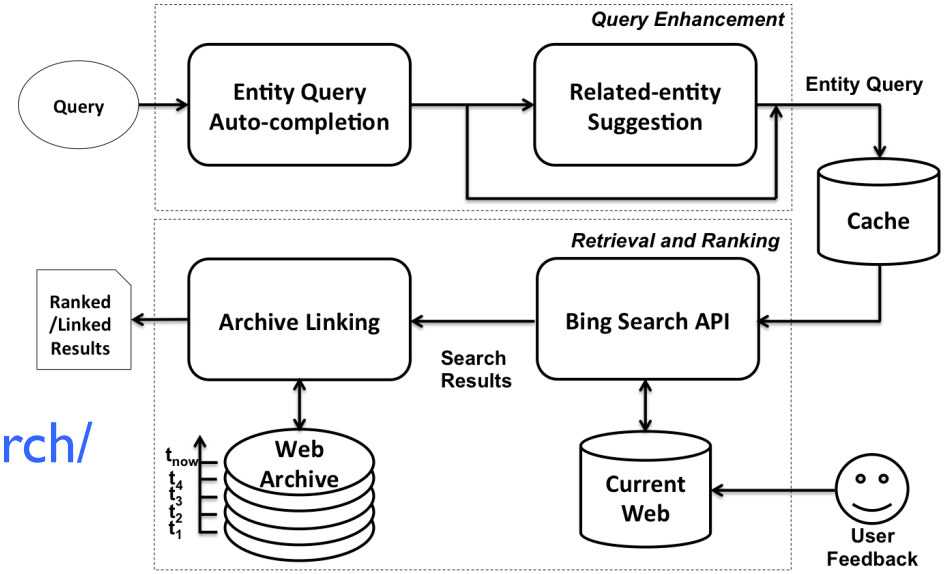
Angela Merkel - Chancellor - Biography.com
Wayback » http://www.biography.com/people/angela-merkel-9406424#!
captured between 10/30/11 and 12/4/15
Angela Merkel is a German politician best known as the first female chancellor of Germany and one of the architects of the European Union.

Angela Merkel | World news | The Guardian
Wayback » http://www.theguardian.com/world/angela-merkel
captured between 7/31/13 and 6/3/16
EU referendum live with Andrew Sparrow EU referendum: **Merkel** says UK will lose out if it leaves EU - as it happened

Angela Merkel - facebook.com
Wayback » https://www.facebook.com/AngelaMerkel
captured between 9/2/09 and 5/30/16
Angela Merkel, Berlin, Germany. 2,038,800 likes · 31,232 talking about this. Facebook-Seite der CDU-Vorsitzenden, Bundeskanzlerin **Angela Merkel**.

Related entities:

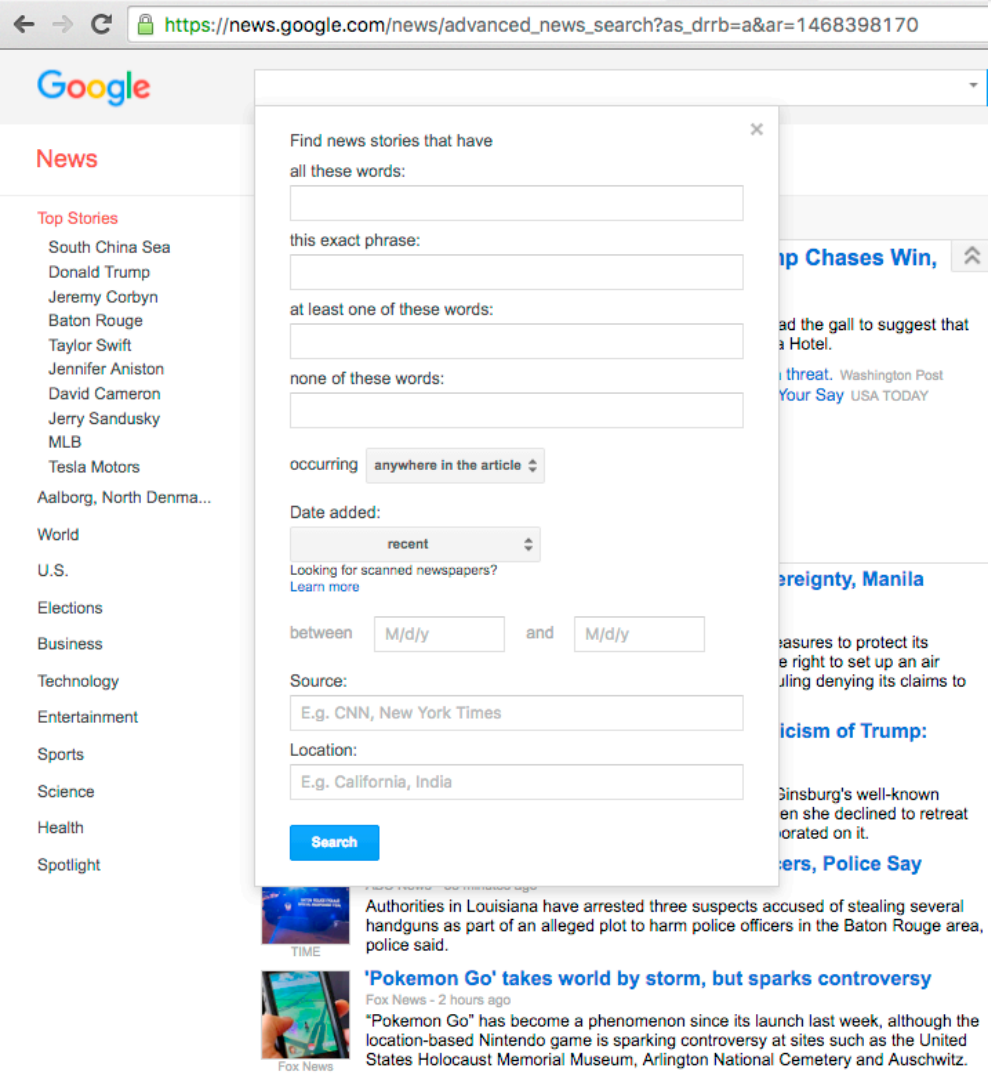
Thomas de Maizière
Hans-Peter Friedrich
José Sócrates
Jens Stoltenberg
Edmund Stoiber
Ilse Aigner
European debt crisis
Annegret Kramp-Karrenbauer
Bundeswehr
Barbara Hendricks (politician)
Richard von Weizsäcker



http://alexandria-project.eu/archivesearch/

[Kanhabua et al., TPDL 2016]

Google News Archive



Search news archive

Searching for news shows you the most recent articles relevant to your query. Sometimes you may be interested in searching news over historical periods of time to find stories that are most significant over those periods, rather than the most recent. You can do this by searching the news archives.

To search news archives:

1. Go to news.google.com.
2. Type in your query in the search box at the top of the page.
3. Select **Enter**.
4. From the search results page that appears, go to Search tools below the search box.
5. From the menu that appears, click the **Recent** drop-down list.
6. Click **Archive**.
7. Click **Enter**.
8. You will see search results ranked by significance

The Archive option lets you search for web news content until the year 2003. You can restrict your search to a more specific data range within this period. To do this, in step 6 above, select Custom range and type your specified dates.

Searching within Google News won't show results older than the year 2003. If you're looking for older content, you have two options:

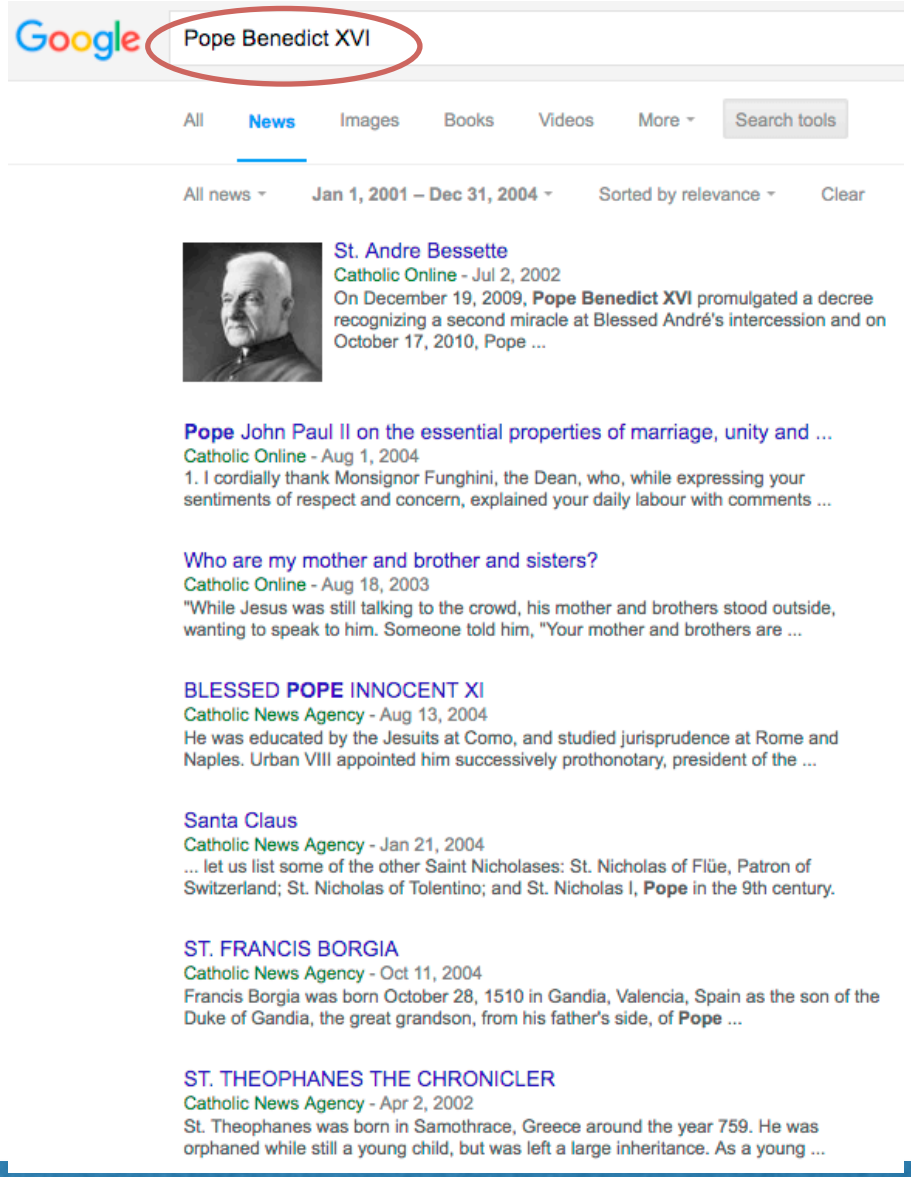
- Use Google Web Search at www.google.com. Note that Google Web Search doesn't support custom date ranges earlier than 1970 or link to content behind a paywall.
- Search for scanned newspapers (see section below).

Search newspaper archives

To locate an article from a scanned newspaper, go to www.google.com and type in site:google.com/newspapers, followed by the search terms you'd like to use. For example, if you're searching for a scanned article on the Berlin wall, you would type in:

site:google.com/newspapers "the Berlin wall"

Google News Archive



Do not handle *name changes* over time

Application Areas

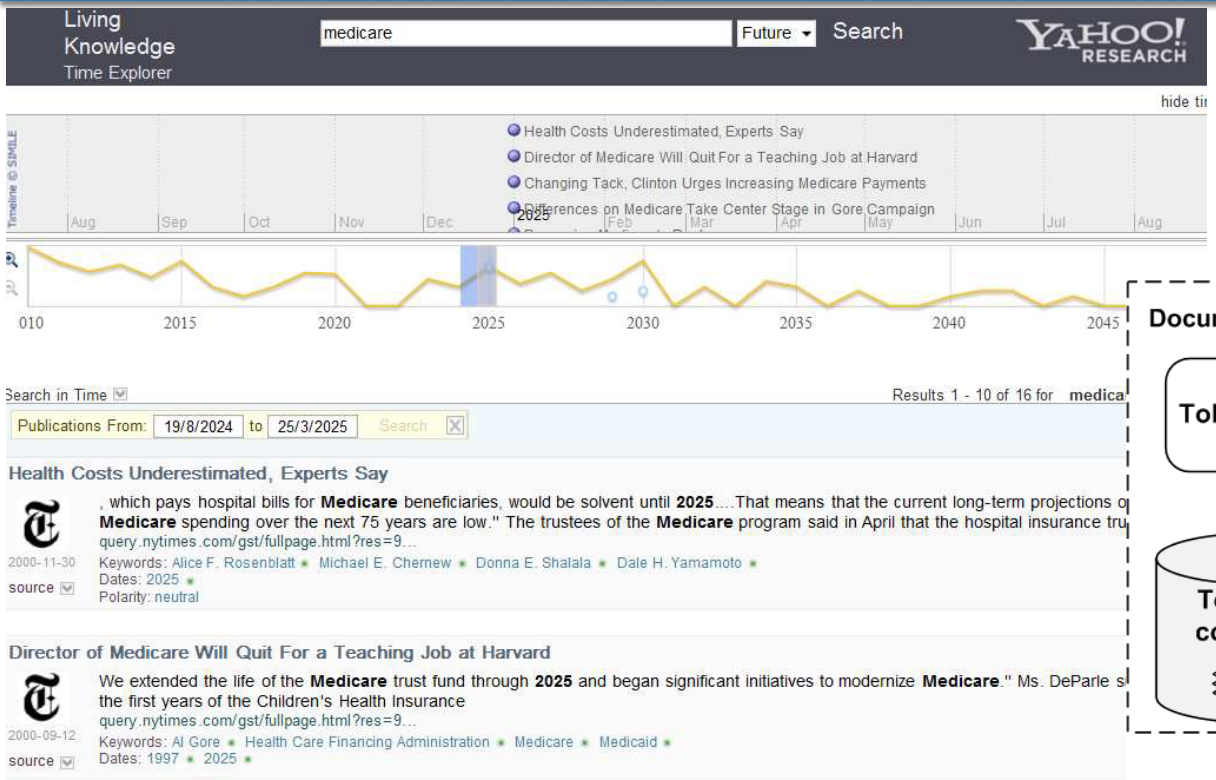
(1)
Web Archive Search
News Archive Search

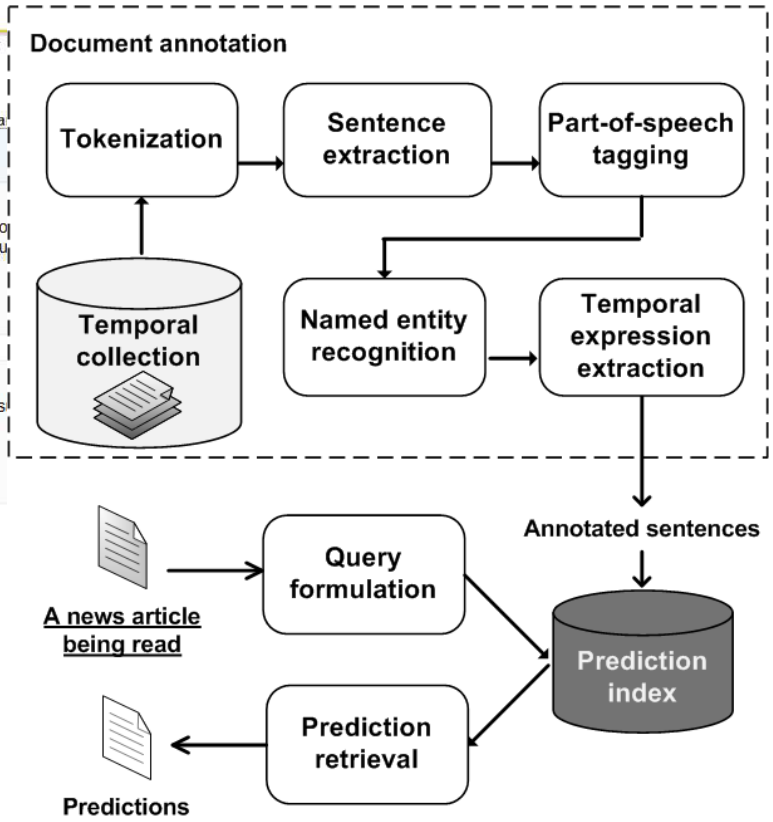
(2)
Temporal Analytics and Exploration

(3)
Temporal Clustering and Summarization

(4)
Search the Past
Future Event Retrieval

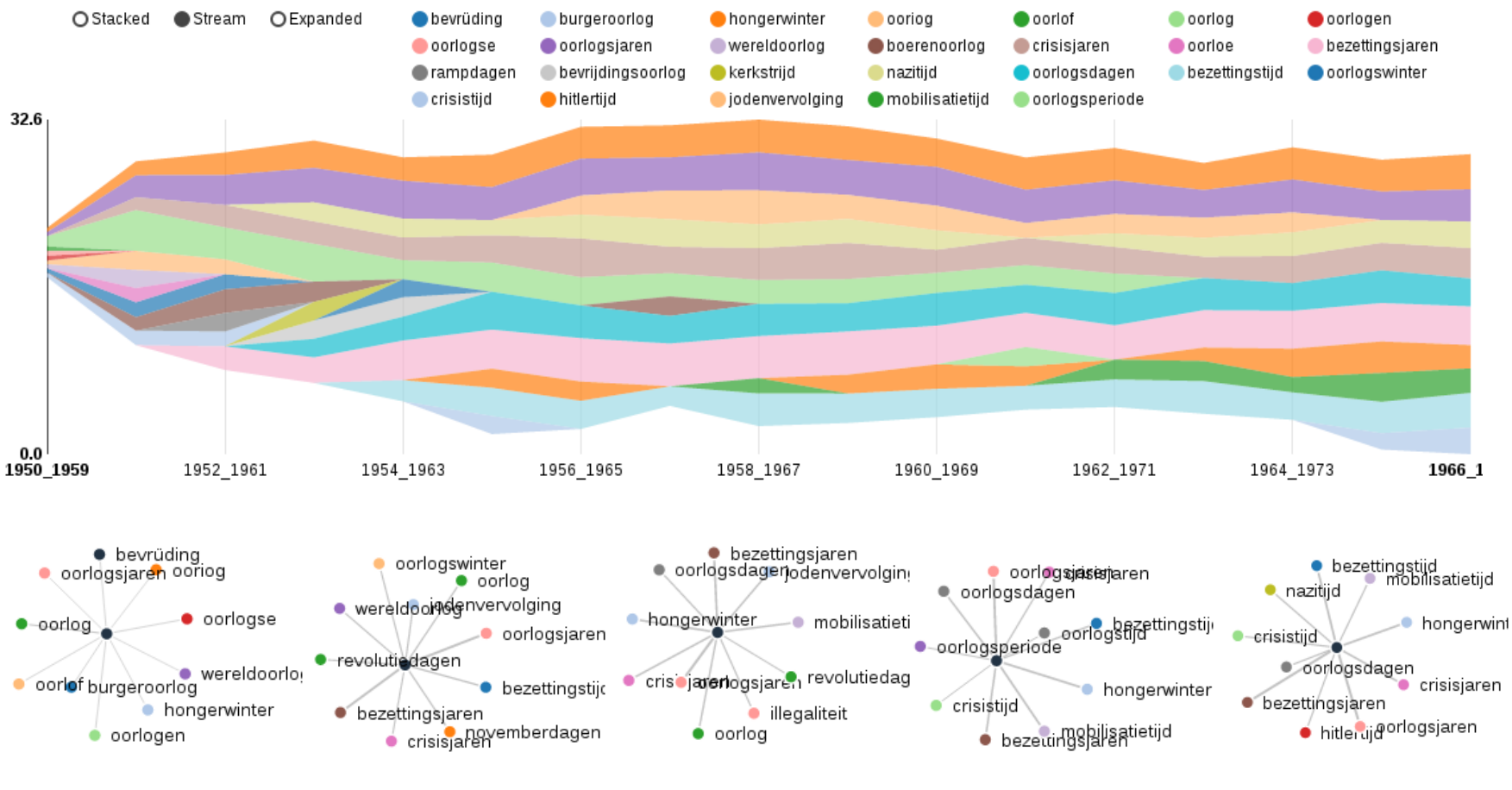
Yahoo! Time Explorer





[Matthews et al., HCIR Workshop 2010]

ShiCo: Visualizing Shifting Concepts over Time



[Martinez-Ortiz et al., Histoinfomatics 2016]

Ad Hoc Monitoring of Vocabulary Shifts over Time

- Data: >600.000 digitized newspaper issues from the Dutch National Library
- Multi-dimensional word-vector space using Google's word2vec (word embeddings)
- **Tracing concepts:** identify words remain and disappear from network

1 model = 10 years

40 models for period between 1950-1990

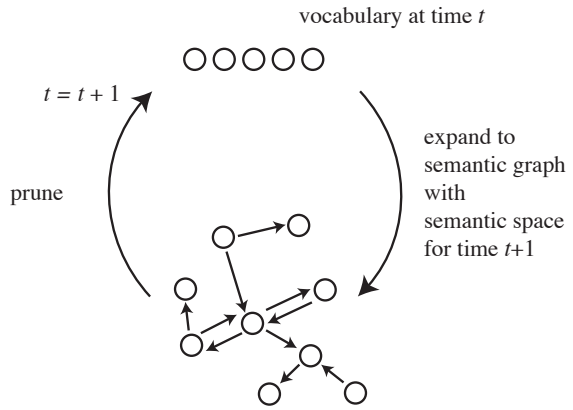


Figure 1: Schematic representation of the generation algorithm for generating vocabularies over time.

[Kenter et al., CIKM 2015]

Application Areas

(1)
Web Archive Search
News Archive Search

(2)
Temporal Analytics and Exploration

(3)
Temporal Clustering and Summarization

(4)
Search the Past
Future Event Retrieval

Clustering and Exploring Search Results using Timeline

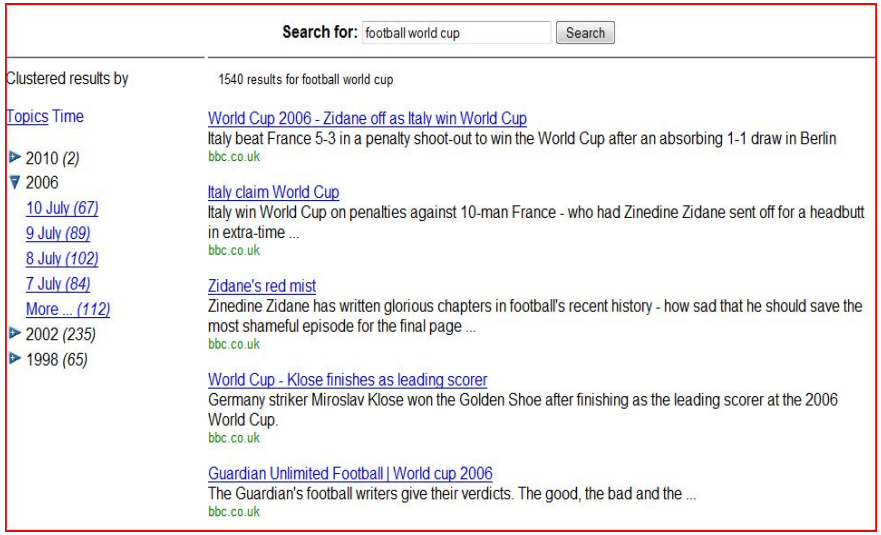


Figure 1: Timeline cluster for the query [football world cup]

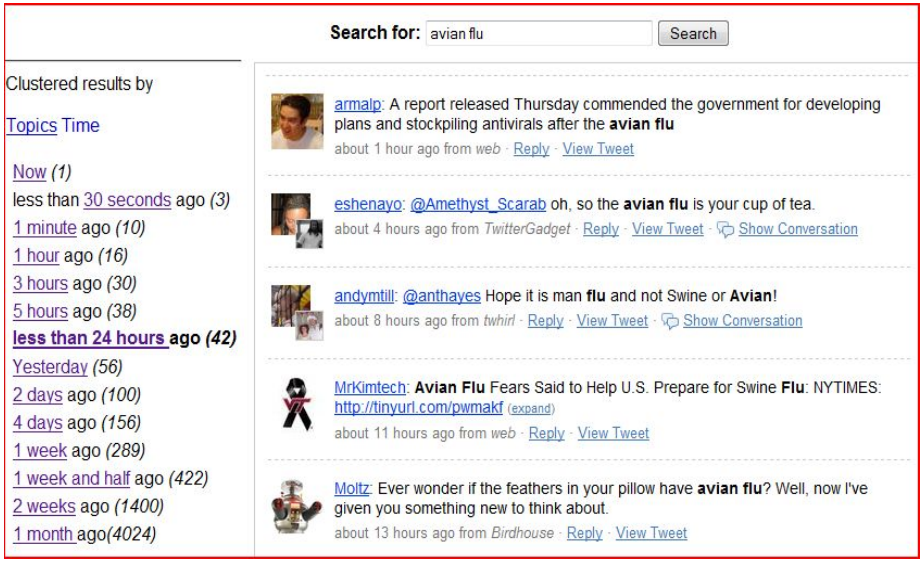


Figure 2: Timeline cluster for [avian flu] twitts.

[Alonso et al., CIKM 2009]

Timeline Summarization

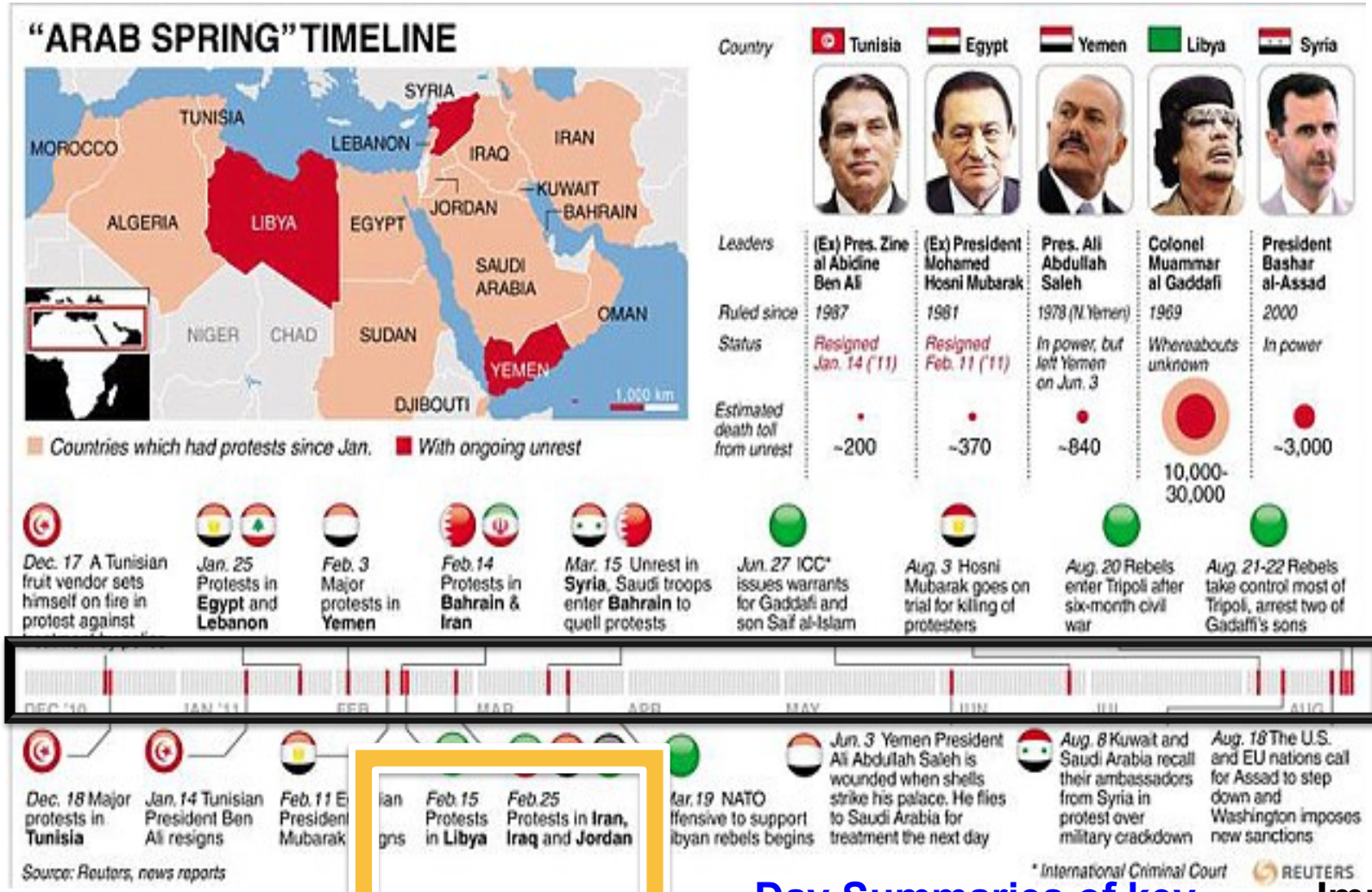
- Query topic: ***Arab Spring***
- What and how did it happen?

- A summarization with the temporal structure consisting of the list of ***daily key events***

- Example:
  - 11 Feb 2011: Egypt President Hosni Mubarak resigned
  - 15 Feb 2011: protests broke out against Muammar Gaddafi's regime
  - 03 Mar 2011: Egypt Prime Minister Ahmed Shafik resigned

[Tran et al., TAIA 2013]

Timeline Example



Day Summaries of key events

Important dates

Existing Methods

- Multi-document summarization [Erkan and Radev, *J. Artif. Int. Res.* 2004]
- Use of burstiness + interest score (sum TFxIDF similarity to neighbor sentences) [Chieu et al., SIGIR 2004]
- Temporal summary of landmark documents, authors, and topics [Sipos et al., CIKM 2012]
- Topic relevancy + coverage + coherence + diversity based on word distribution [Yan et al., SIGIR 2011]

Entity-centric Timeline Summarization of News Events

- Learn from expert-created timeline summaries, and optimize with different criteria
- Use named entities as cues to summarize long-running news events

Germanwings Flight 9525

crashed into the Alp

Digne-les-Bain

"accident site spread across 5 acres ..", "is horrible"

Germanwings

several Germanwings flight cancelled after crew refused to fly

Germanwings Flight 9525

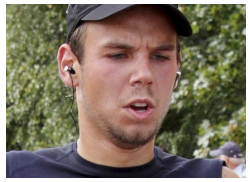
"detailed of victims in the crash of GF 4U9525 …

Joseph-König-Gymnasium

"classes cancelled at JKG after 16 students confirmed to have died

Francois Hollande

President FH: "a tragedy on our soil" ….. report from President Francois Hollande conflicts with….



Andreas Lubitz

"blackbox data analysis confirmed co-pilot Andreas Lubitz deliberately…

Carsten Spohr

Lufthansa CEO stunned that co-pilot crashed …. gives a speech about ….

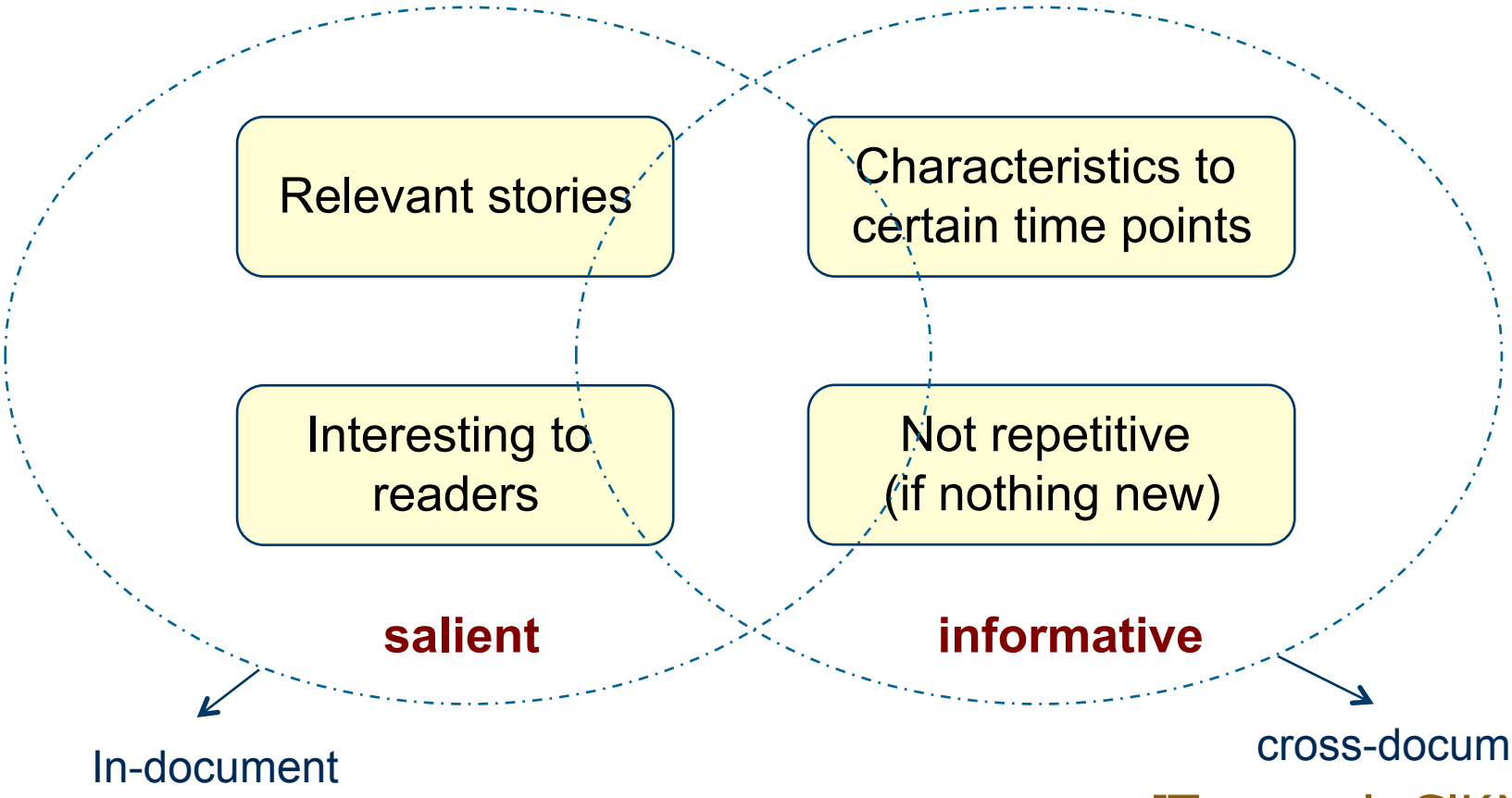
Joseph-König-Gymnasium

A moment of silence is held Thursday at JKG

24 March | 25 March | 26 March

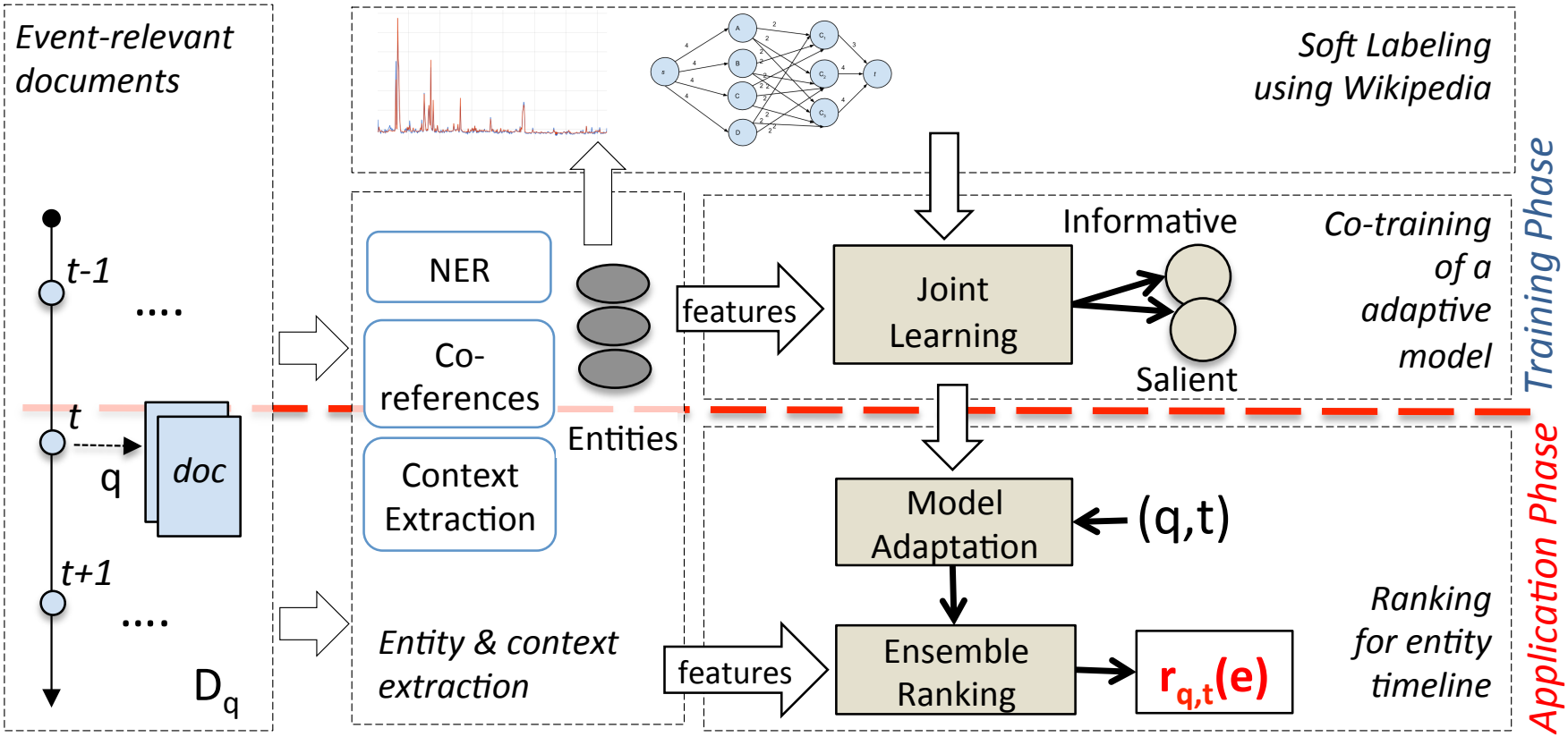
[Tran et al., CIKM 2015]

Key Entities for a News Event Timeline



[Tran et al., CIKM 2015]

System Framework



[Tran et al., CIKM 2015]

Remarks

- Entities as units summarize better complex events
  - Need a dynamic and adaptive ranking for entities
  - High-impact events can be learnt using social evidences at scale
- But still the first step:
  - Better extraction of entities and contexts
  - More flexible adaptive learning-to-rank
  - Larger set of events

[Tran et al., CIKM 2015]

Application Areas

(1)
Web Archive Search
News Archive Search

(2)
Temporal Analytics and Exploration

(3)
Temporal Clustering and Summarization

(4)
Search the Past
Future Event Retrieval

Time-aware Contextualization

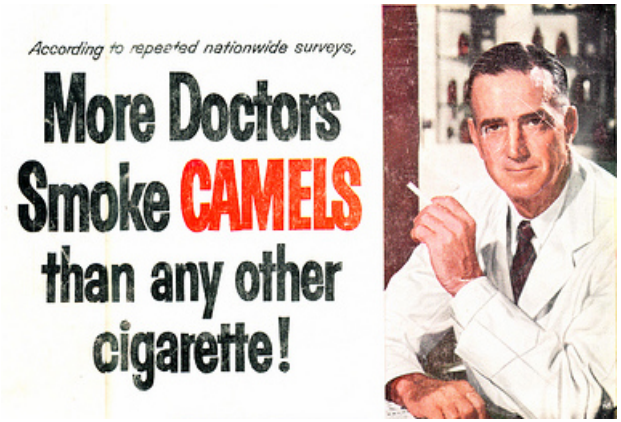
Advertisement poster from the 1950s

Time-aware contextualization

Prior to 1964, many of the cigarette companies advertised their brand by falsely claiming that their product did not have serious health risks. A couple of examples would be "Play safe with Philip Morris" and "More doctors smoke Camels". Such claims were made both to increase the sales of their product and to combat the increasing public knowledge of smoking's negative health effects.

[Tran et al., WSDM 2015]

Entity Linking



Physician
http://en.wikipedia.org/wiki/Physician

Camel (cigarette)
http://en.wikipedia.org/wiki/Camel_(cigarette)

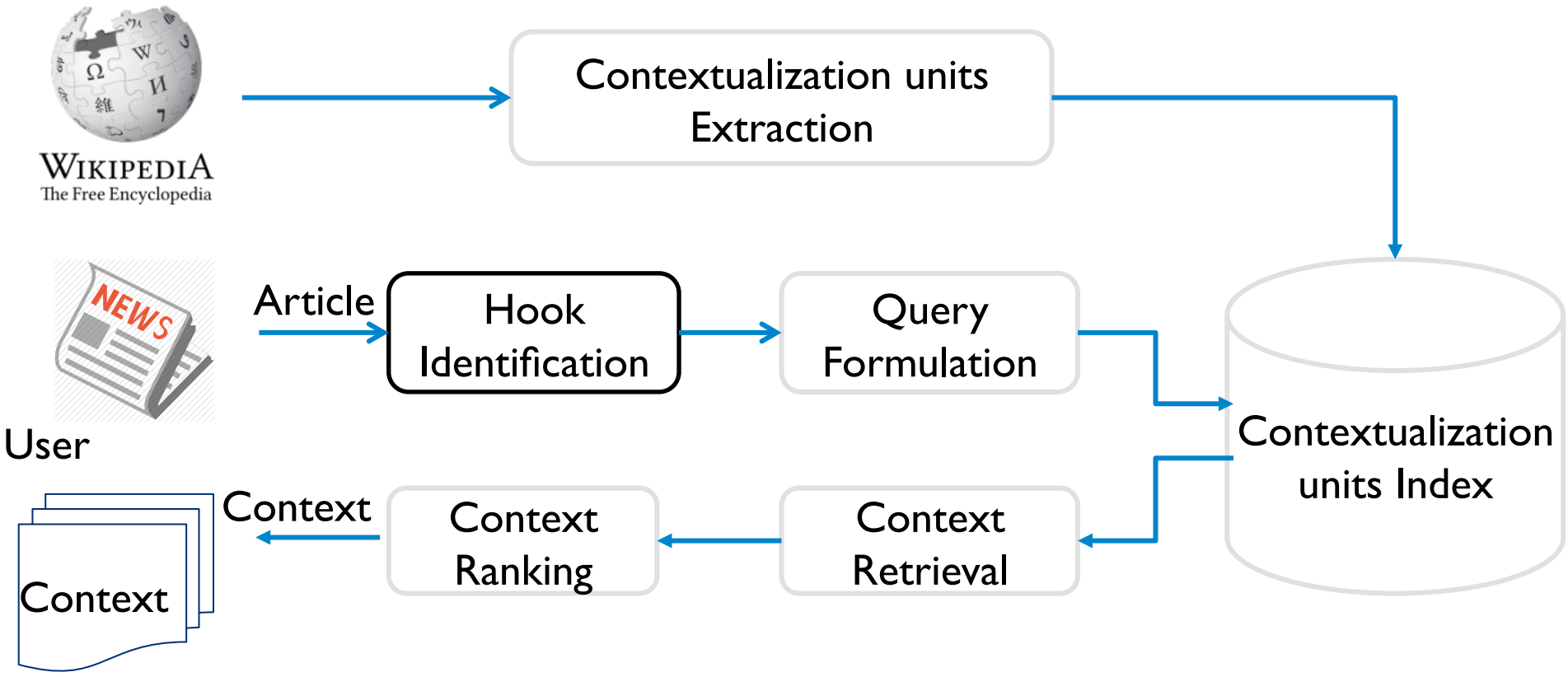
Cigarette
http://en.wikipedia.org/wiki/Cigarette

Entity linking is not sufficient

- Wikipedia pages tend to contain large amounts of content
- Relevant information might be distributed over various articles
- The crucial temporal aspect is missing in pure linking approaches

[Tran et al., WSDM 2015]

Approach Overview



[Tran et al., WSDM 2015]

Searching the Future

- People are naturally curious about the future
  - What will happen to EU economies in next 5 years?
  - What will be potential effects of climate changes?

Current Works

- Searching the future [Baeza-Yates SIGIR Forum 2005]
  - Extract temporal expressions from news articles
  - Retrieve future information using a probabilistic model, i.e., multiplying textual similarity and a time confidence

- Supporting analysis of future-related information in news and the Web [Jatowt et al., JCDL 2009]
  - Extract future mentions from news snippets obtained from search engines
  - Summarize and aggregate results using clustering methods, but no ranking

Recorded Future

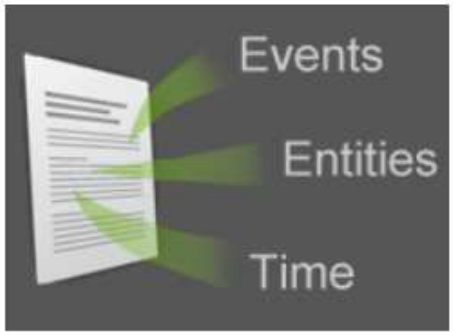
1. Scour the web

We continually scan thousands of high-quality news publications, blogs, public niche sources, trade publications, government web sites, financial databases and more.

2. Extract, analyze & rank

We extract information from text including entities, events, and the time that these events occur.

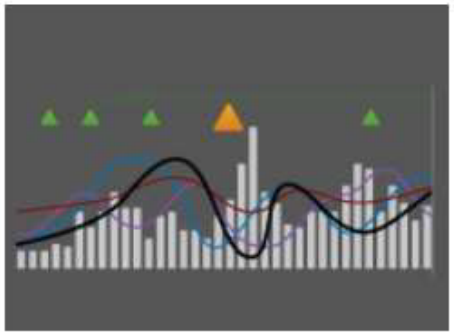
We also measure momentum for each item in our index, as well as sentiment.



3. Make it useful

You can explore the past, present and predicted future of almost anything.

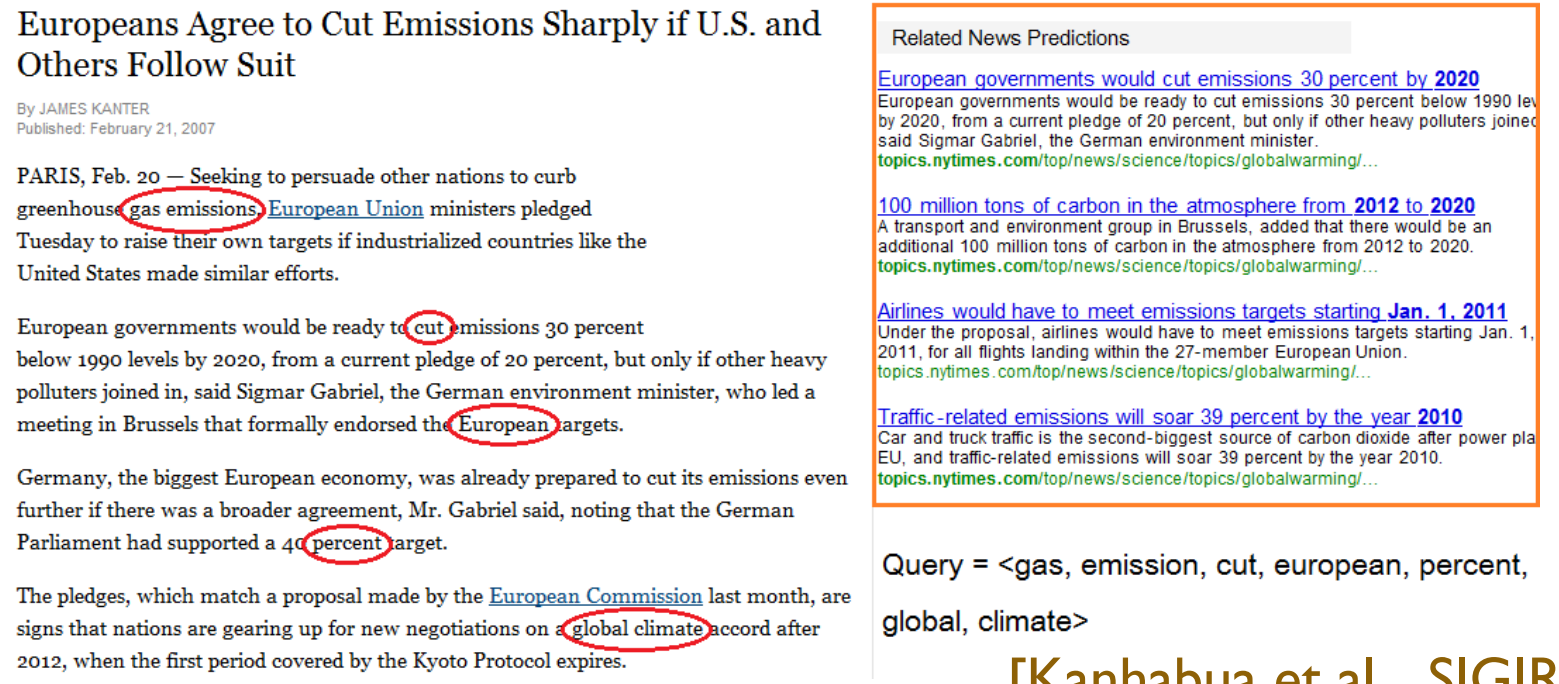
Powerful visualization tools allow you to quickly see temporal patterns, or link networks of related information.



http://www.recordedfuture.com/

Ranking News Predictions

- Over 32% of 2.5M documents from Yahoo! News (July'09 – July'10) contain at least one prediction
- Retrieve *predictions* related to a news story in news archives and rank by relevance



Europeans Agree to Cut Emissions Sharply if U.S. and Others Follow Suit

By JAMES KANTER
Published: February 21, 2007

PARIS, Feb. 20 — Seeking to persuade other nations to curb greenhouse gas emissions, European Union ministers pledged Tuesday to raise their own targets if industrialized countries like the United States made similar efforts.

European governments would be ready to cut emissions 30 percent below 1990 levels by 2020, from a current pledge of 20 percent, but only if other heavy polluters joined in, said Sigmar Gabriel, the German environment minister, who led a meeting in Brussels that formally endorsed the European targets.

Germany, the biggest European economy, was already prepared to cut its emissions even further if there was a broader agreement, Mr. Gabriel said, noting that the German Parliament had supported a 40 percent target.

The pledges, which match a proposal made by the European Commission last month, are signs that nations are gearing up for new negotiations on a global climate accord after 2012, when the first period covered by the Kyoto Protocol expires.

Related News Predictions

European governments would cut emissions 30 percent by **2020**
European governments would be ready to cut emissions 30 percent below 1990 lev by 2020, from a current pledge of 20 percent, but only if other heavy polluters joined said Sigmar Gabriel, the German environment minister.
topics.nytimes.com/top/news/science/topics/globalwarming/...

100 million tons of carbon in the atmosphere from **2012** to **2020**
A transport and environment group in Brussels, added that there would be an additional 100 million tons of carbon in the atmosphere from 2012 to 2020.
topics.nytimes.com/top/news/science/topics/globalwarming/...

Airlines would have to meet emissions targets starting **Jan. 1, 2011**
Under the proposal, airlines would have to meet emissions targets starting Jan. 1, 2011, for all flights landing within the 27-member European Union.
topics.nytimes.com/top/news/science/topics/globalwarming/...

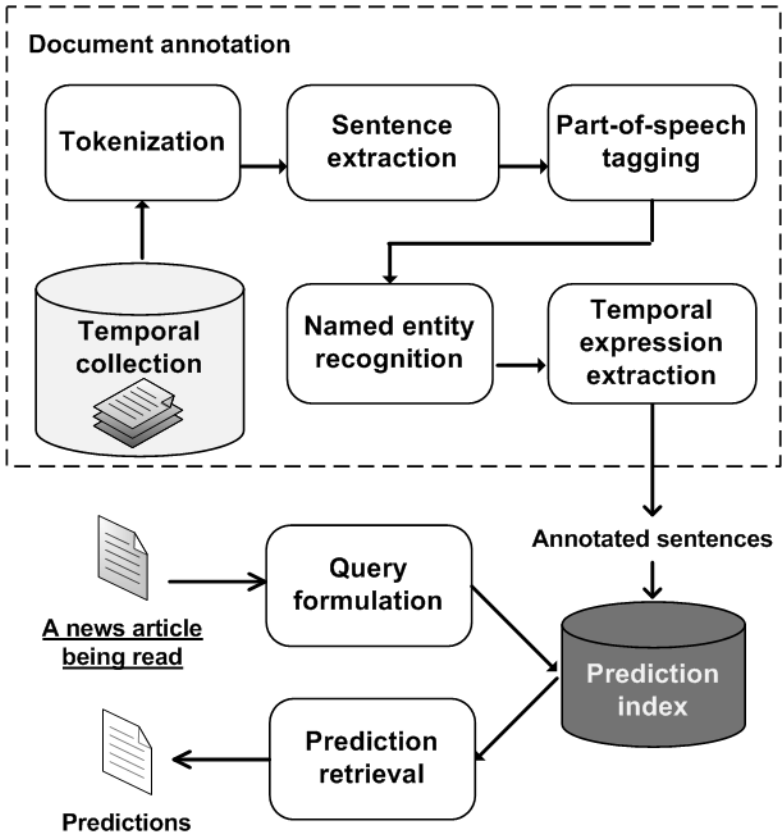
Traffic-related emissions will soar 39 percent by the year **2010**
Car and truck traffic is the second-biggest source of carbon dioxide after power pla EU, and traffic-related emissions will soar 39 percent by the year 2010.
topics.nytimes.com/top/news/science/topics/globalwarming/...

Query = <gas, emission, cut, european, percent, global, climate>

[Kanhabua et al., SIGIR 2011]

System Architecture

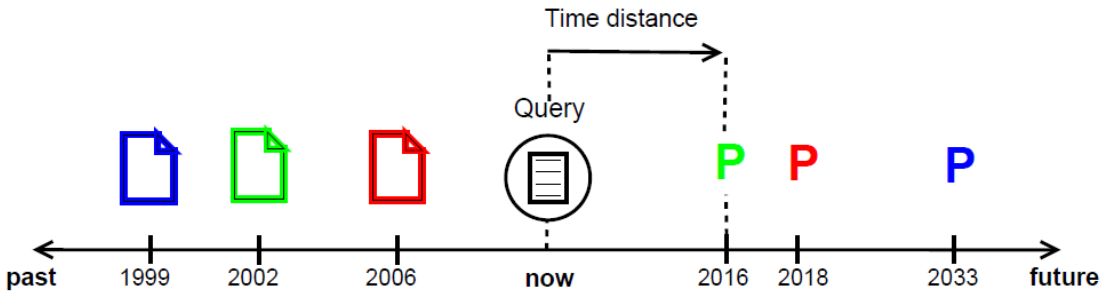
- **Step 1: Document annotation**
  - Extract temporal expressions using time and event recognition
  - Normalize them to dates so they can be anchored on a timeline
  - Output: sentences annotated with named entities and dates, i.e., *predictions*
- **Step 2: Retrieving predictions**
  - Automatically generate a query from a news article *being read*
  - Retrieve predictions that match the query
  - Rank predictions by relevance (i.e., a prediction is "relevant" if it is about the *topics of the article*)



[Kanhabua et al., SIGIR 2011]

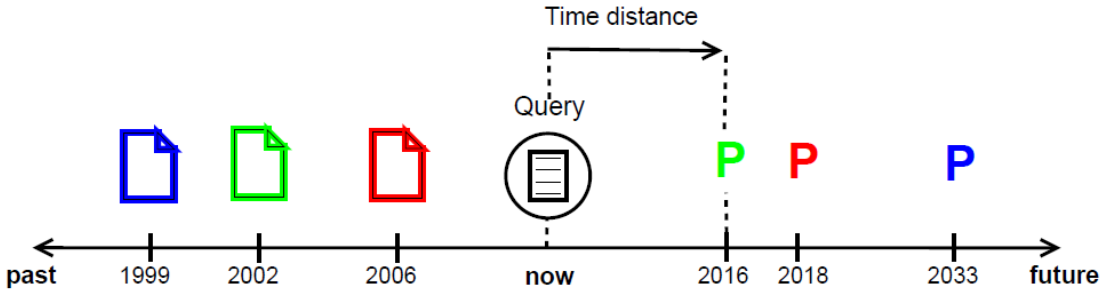
Temporal Similarity

- **Hypothesis I.** Predictions that are more recent to the query are more relevant

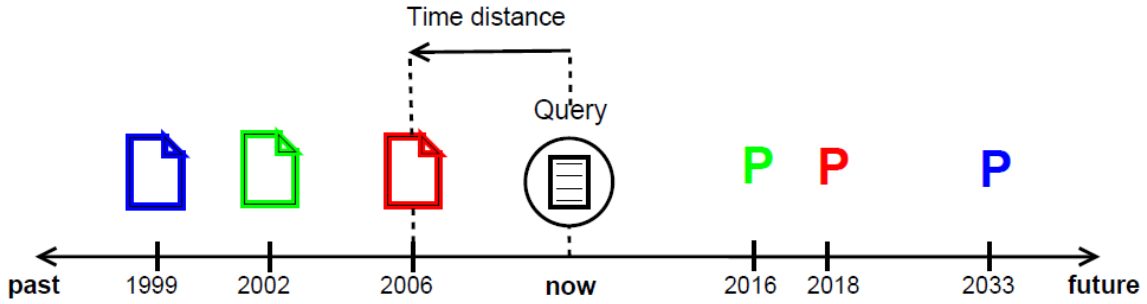


Temporal Similarity

- **Hypothesis I.** Predictions that are more recent to the query are more relevant



- **Hypothesis II.** Predictions extracted from more recent documents are more relevant



Ranking Method

- **Learning-to-rank**: Given an unseen (q, p), p is ranked using a model trained over a set of labeled query/prediction

$$score(q, p) = \sum_{i=1}^{N} w_i \times f_i$$

  - SVM-*MAP* [Yue et al., SIGIR 2007]
  - RankSVM [Joachims, KDD 2002]
  - SGD-SVM [Zhang, ICML 2004]
  - PegasosSVM [Shalev-Shwartz et al., ICML 2007]
  - PA-Perceptron [Crammer et al., J. Mach. Learn. 2006]

Results

- New York Times Annotated Corpus
  - 1.8 million articles, over 20 years
  - More than 25% contain *at least one prediction* (44,335,519 sentences, 548,491 predictions and 939,455 future dates)

- Results:
  - Topic features play an important role in ranking
  - Features in top-5 features with *lowest weights* are entity-based features

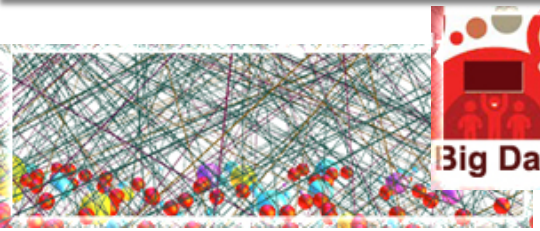
- Open issues:
  - Extract predictions from other sources, e.g., Wikipedia, blogs, comments, etc.
  - Sentiment analysis for future-related information

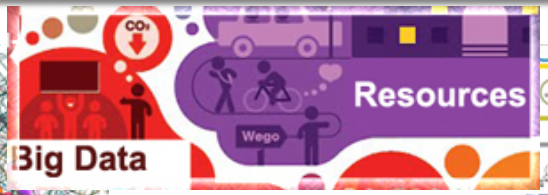
Conclusions and Outlooks

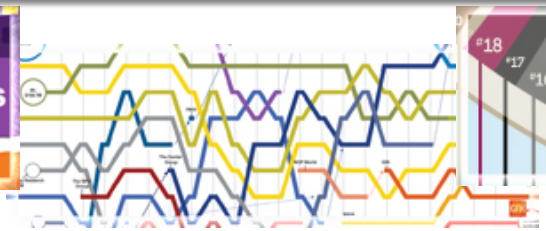
Summary

- Part I
  - Introduction to Temporal IR
  - Temporal Indexing and Query Processing
  - Time-aware Retrieval and Ranking
- Part II
  - Temporal Query Analysis
  - Applications of Temporal IR

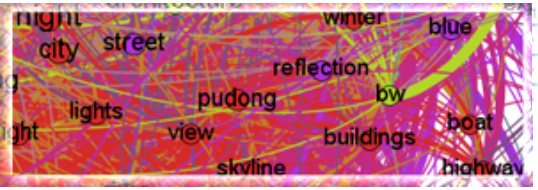
SoBigData: Research Infrastructure

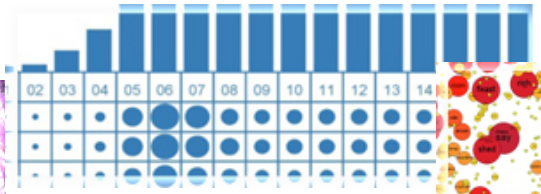


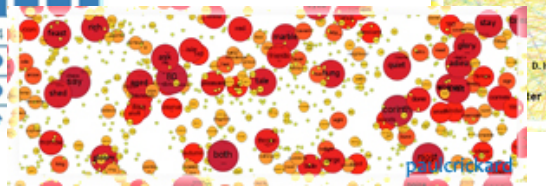




A **Multidisciplinary** European **Infrastructure** for **Big Data** and **Social Data Mining** providing an integrated ecosystem for **ethically sensitive scientific discoveries** and advanced applications of social data mining on various dimensions of social life, as recorded by "big data"







http://www.sobigdata.eu

Next Research Topics

Visual Analytics

Semantics visualization and real-time solutions for the simulation, visualization and rendering of data

Web Analytics

New models and tools for understanding user behaviour for improving the users' web experience.

Human Mobility Analytics

Methods for mobility analytics, pre-processing tools, integration of diverse mobility observations, advanced mobility data mining

Social Network Analysis

Methods and algorithms covering the major research topics in social networks, community discovery, evolutionary analysis

Social Data

Intersection of natural, social and engineering sciences to address the challenges that social data generate

Text and Social Media Mining

Natural processing language, mining social media, information extraction, ontology-based semantic annotation

http://www.sobigdata.eu


thank you
danke
謝謝
teşekkür ederim
gracias
спасибо
merci
obrigado
dziękuje
bedankt
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
grazie
arigatô
dakujem
terima kasih
감사합니다
ありがとう
мерси
tapadh leat
dank je
hvala
takk
sagolun
Баярлалаа
рахмат
хвала

References

- [Adar et al., WSDM 2009] Eytan Adar, Jaime Teevan, Susan T. Dumais, Jonathan L. Elsas: The web changes everything: understanding the dynamics of web content. WSDM 2009: 282-291
- [Alonso et al., CIKM 2009] Omar Alonso, Michael Gertz, Ricardo A. Baeza-Yates: Clustering and exploring search results using timeline constructions. CIKM 2009: 97-106
- [Anand et al., SIGIR 2011] Avishek Anand, Srikanta J. Bedathur, Klaus Berberich, Ralf Schenkel: Temporal index sharding for space-time efficiency in archive search. SIGIR 2011: 545-554
- [Beitzel et al., SIGIR 2004] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David A. Grossman, Ophir Frieder: Hourly analysis of a very large topically categorized web query log. SIGIR 2004: 321-328
- [Berberich and Bedathur, TAIA 2013] Klaus Berberich, Srikanta Bedathur: Temporal Diversification of Search Results. TAIA 2013

References

- [Berberich et al., SIGIR 2007] Klaus Berberich, Srikanta J. Bedathur, Thomas Neumann, Gerhard Weikum: A time machine for text search. SIGIR 2007: 519-526
- [Berberich et al., ECIR 2010] Klaus Berberich, Srikanta J. Bedathur, Omar Alonso, Gerhard Weikum: A Language Modeling Approach for Temporal Information Needs. ECIR 2010: 13-25
- [Bian et al., WWW 2010] Jiang Bian, Xin Li, Fan Li, Zhaohui Zheng, Hongyuan Zha: Ranking specialization for web search: a divide-and-conquer approach by using topical RankSVM. WWW 2010: 131-140
- [Chieu et al., SIGIR 2004] Hai Leong Chieu, Yoong Keok Lee: Query based event extraction along a timeline. SIGIR 2004: 425-432
- [Crammer et al., J. Mach. Learn. 2006] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer: Online Passive-Aggressive Algorithms. J. Mach. Learn. 2006: 551-585

References

- [Dai et al., SIGIR 2011] Na Dai, Milad Shokouhi, Brian D. Davison: Learning to rank for freshness and relevance. SIGIR 2011: 95-104
- [Diaz and Jones, SIGIR 2004] Fernando Diaz, Rosie Jones: Using temporal profiles of queries for precision prediction. SIGIR 2004: 18-24
- [Dong et al., WSDM 2010] Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, Fernando Diaz: Towards recency ranking in web search. WSDM 2010: 11-20
- [Efron and Golovchinsky, SIGIR 2011] Miles Efron, Gene Golovchinsky: Estimation methods for ranking recent information. SIGIR 2011: 495-504
- [Erkan and Radev, *J. Artif. Int. Res.* 2004] Günes Erkan, Dragomir R Radev: LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research: 457-479

References

- [He et al., CIKM 2009] Jinru He, Hao Yan, Torsten Suel: Compact full-text indexing of versioned document collections. CIKM 2009: 415-424
- [He et al., CIKM 2010] Jinru He, Junyuan Zeng, Torsten Suel: Improved index compression techniques for versioned document collections. CIKM 2010: 1239-1248
- [Holzmann et al., JCDL 2016] Helge Holzmann, Wolfgang Nejdl, Avishek Anand: The Dawn of Today's Popular Domains: A Study of the Archived German Web over 18 Years. JCDL 2016: 73-82
- [Joachims, KDD 2002] Thorsten Joachims: Optimizing search engines using clickthrough data. KDD 2002: 133-142
- [Joho et al., TempWeb 2014] Hideo Joho, Adam Jatowt, Roi Blanco: NTCIR temporalia: a test collection for temporal information access research. WWW 2015 Companion: 845-850

References

- [Kanhabua et al., SIGIR 2011] Nattiya Kanhabua, Roi Blanco, Michael Matthews: Ranking related news predictions. SIGIR 2011: 755-764
- [Kanhabua et al., FnTIR 2015] Nattiya Kanhabua, Roi Blanco, Kjetil Nørvåg: Temporal Information Retrieval. Foundations and Trends in Information Retrieval 9(2): 91-208 (2015)
- [Kanhabua et al., Neu-IR 2016] Nattiya Kanhabua, Huamin Ren, Thomas B. Moeslund: Learning Dynamic Classes of Events using Stacked Multilayer Perceptron Networks. CoRR abs/1606.07219 (2016)
- [Kanhabua et al., TPDL 2016] Nattiya Kanhabua, Philipp Kemkes, Wolfgang Nejdl, Tu Ngoc Nguyen, Felipe Reis, Nam Khanh Tran: How to Search the Internet Archive Without Indexing It. TPDL 2016

References

- [Kenter et al., CIKM 2015] Tom Kenter, Melvin Wevers, Pim Huijnen, Maarten de Rijke: Ad Hoc Monitoring of Vocabulary Shifts over Time. CIKM 2015: 1191-1200
- [Kulkarni et al., WSDM 2011] Anagha Kulkarni, Jaime Teevan, Krysta Marie Svore, Susan T. Dumais: Understanding temporal query dynamics. WSDM 2011: 167-176
- [Lavrenko and Croft, SIGIR 2001] Victor Lavrenko, W. Bruce Croft: Relevance-Based Language Models. SIGIR 2001: 120-127
- [Li and Croft, CIKM 2003] Xiaoyan Li, W. Bruce Croft: Time-based language models. CIKM 2003: 469-475
- [Martinez-Ortiz et al., Histoinfomatics 2016] Carlos Martinez-Ortiz, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul, Joris van Eijnatten: Design and implementation of ShiCo: Visualising shifting concepts over time. DH Histoinfomatics Workshop 2016

References

- [Matthews et al., HCIR Workshop 2010] Michael Matthews, Pancho Tolchinsky, Roi Blanco, Jordi Atserias, Peter Mika, Hugo Zaragoza: Searching through time in the new york times. HCIR Workshop 2010
- [Metzler et al., SIGIR 2009] Donald Metzler, Rosie Jones, Fuchun Peng, Ruiqiang Zhang: Improving search relevance for implicitly temporal queries. SIGIR 2009
- [Nguyen et al., ECIR 2014] Tu Ngoc Nguyen, Nattiya Kanhabua: Leveraging Dynamic Query Subtopics for Time-Aware Search Result Diversification. ECIR 2014: 222-234
- [Ntoulas et al., 2004] Alexandros Ntoulas, Junghoo Cho, Christopher Olston: What's new on the web?: the evolution of the web from a search engine perspective. WWW 2004: 1-12

References

- [Nunes et al., ECIR 2008] Sérgio Nunes, Cristina Ribeiro, Gabriel David: Using neighbors to date web documents. WIDM 2007: 129-136
- [Radinsky et al., WSDM 2013] Kira Radinsky, Paul N. Bennett: Predicting content change on the web. WSDM 2013: 415-424
- [Radinsky et al., WWW 2012] Kira Radinsky, Krysta Marie Svore, Susan T. Dumais, Jaime Teevan, Alex Bocharov, Eric Horvitz: Modeling and predicting behavioral dynamics on the web. WWW 2012: 599-608
- [Shalev-Shwartz et al., ICML 2007] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, Andrew Cotter: Pegasos: primal estimated sub-gradient solver for SVM. Math. Program. 127(1): 3-30

References

- [Shokouhi, SIGIR 2011] Milad Shokouhi: Detecting seasonal queries by time-series analysis. SIGIR 2011: 1171-1172
- [Singh et al., CHIIR 2016] Jaspreet Singh, Wolfgang Nejdl, Avishek Anand: History by Diversity: Helping Historians search News Archives. CHIIR 2016: 183-192
- [Sipos et al., CIKM 2012] Ruben Sipos, Adith Swaminathan, Pannaga Shivaswamy, Thorsten Joachims: Temporal corpus summarization using submodular word coverage. CIKM 2012: 754-763
- [Tran et al., CIKM 2015] Tuan A. Tran, Claudia Niederée, Nattiya Kanhabua, Ujwal Gadiraju, Avishek Anand: Balancing Novelty and Salience: Adaptive Learning to Rank Entities for Timeline Summarization of High-impact Events. CIKM 2015: 1201-1210

References

- [Tran et al., TAIA 2013] Giang Binh Tran, Tuan A. Tran, Nam-Khanh Tran, Mohammad Alrifai, Nattiya Kanhabua: Leveraging Learning To Rank in an Optimization Framework for Timeline Summarization. SIGIR TAIA Workshop 2013
- [Tran et al., WSDM 2015] Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua, Claudia Niederée: Back to the Past: Supporting Interpretations of Forgotten Stories by Time-aware Re-Contextualization. WSDM 2015: 339-348
- [Vlachos et al., SIGMOD 2004] Michail Vlachos, Christopher Meek, Zografoula Vagena, Dimitrios Gunopulos: Identifying Similarities, Periodicities and Bursts for Online Search Queries. SIGMOD 2004: 131-142
- [Whiting et al., SIGIR 2013] Stewart Whiting, Joemon M. Jose: Recent and robust query auto-completion. WWW 2014: 971-982

References

- [Yan et al., SIGIR 2011] Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, Yan Zhang: Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. SIGIR 2011: 745-754
- [Yue et al., SIGIR 2007] Yisong Yue, Thomas Finley, Filip Radlinski, Thorsten Joachims: A support vector method for optimizing average precision. SIGIR 2007: 271-278
- [Zhang et al., EMNLP 2010] Ruiqiang Zhang, Yuki Konda, Anlei Dong, Pranam Kolari, Yi Chang, Zhaohui Zheng: Learning Recurrent Event Queries for Web Search. EMNLP 2010: 1129-1139
- [Zhang, ICML 2004] Tong Zhang: Solving large scale linear prediction problems using stochastic gradient descent algorithms. ICML 2004