Advanced Topics in Information Retrieval

Temporal Information Retrieval

Vinay Setty

vsetty@mpi-inf.mpg.de

Jannik Strötgen

jannik.stroetgen@mpi-inf.mpg.de

ATIR – June 23, 2016

Why **temporal information retrieval**?

Time in queries

temporal information needs are frequent

query log analyses

- 1.5% queries with **explicit temporal intent** [Nunes et al. 2008]
- 7% queries with **implicit temporal intent** [Metzler et al. 2009]
- 13.8% **explicit**, 17.1% **implicit** [Zhang et al. 2010]

different types of temporal information in IR

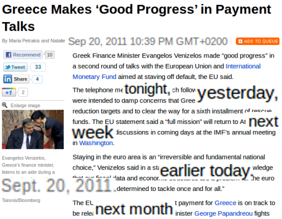
- time as **dimension of relevance**
- time as **query topic**

more in a few minutes

Outline

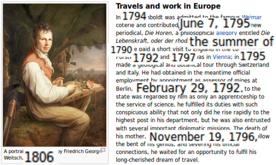
Temporal information extraction

temporal information is **frequent**



News articles.

Narrative documents.



Biographies.

temporal information can be **normalized**

same semantics → same value

- "heute", "aujourd'hui", "today", "June 8, 2016" → 2016-06-08

So far addressed: temporal tagging

Addressed types of temporal expressions

TimeML standard:
[Pustejovsky et al. 2005]

- **dates** ("May 1, 2015", "today")
- **times** ("9 pm", "last night")
- **durations** ("three years")
- **set expressions** ("twice a day")

dates and times may be:

- **explicit** ("May 1, 2015")
- **implicit** ("Christmas 2012")
- **relative** ("last night")
- **underspecified** ("Monday")

normalization difficultly varies between types, but:
all are obvious temporal expressions

So far ignored: free-text temporal expressions

Idea

standard text phrases may be associated with temporal scopes

So far ignored: free-text temporal expressions

Idea

standard text phrases may be associated with temporal scopes

- **temponyms** [Kuzey et al. 2016a, 2016b]

  refer to arbitrary kinds of **named events or facts with temporal scopes** that are merely given by a text phrase but have unique interpretations **given the context and background knowledge**.

- **temponym tagging**

  is the **detection** and **normalization** of temponyms.

Goal

further temporal enrichment of documents

Examples of temponyms

John F. Kennedy's death marked a watershed in the memories of almost every American.

President Obama awarded the nation's highest military honor to a Union soldier who was killed more than 150 years ago during the Battle of Gettysburg.

Examples of temponyms

John F. Kennedy's death marked a watershed in the memories of almost every American.

publication date: 2014-11-06

President Obama awarded the nation's highest military honor to a Union soldier who was killed **more than 150 years ago** during the Battle of Gettysburg.

normalized temporal information
(temporal tagging)

	1864

Examples of temponyms

John F. Kennedy's death marked a watershed in the memories of almost every American.

publication date: 2014-11-06

President Obama awarded the nation's highest military honor to a Union soldier who was killed **more than 150 years ago** during **the Battle of Gettysburg**.

normalized temporal information
(temporal tagging, temponym tagging)

1963-11-22

1864
[1863-07-01, 1863-07-03]

Examples of temponyms

phrases with temponyms

The Cuban Revolutionary War

The second inauguration of Bill Clinton

2008 Mexico City plane crash

2016 WWW Conference

Examples of temponyms

temporal tagging

–	–
The Cuban Revolutionary War	The second inauguration of Bill Clinton

2008	2016
2008 Mexico City plane crash	**2016** WWW Conference

Examples of temponyms

temporal tagging vs. temponym tagging

– vs.
[1953-07-26,1959-01-01]
The Cuban Revolutionary War

– vs.
1997-01-20
The second inauguration of Bill Clinton

2008 vs.
2008-11-04
2008 Mexico City plane crash

2016 vs.
[2016-04-11,2016-04-15]
2016 WWW Conference

temponyms add new or more precise temporal information

WWW'16 paper [Kuzey et al. 2016a]

- all temponyms, not only explicit ones, e.g., "during **his presidency**"
- often year-level temporal scopes approach with **Integer Linear Programming**

As Time Goes By: Comprehensive Tagging of Textual Phrases with Temporal Scopes

TempWeb'16 Approach [Kuzey et al. 2016b]

- explicit temponyms, day-level temporal scopes
- completely other approach than WWW'16 approach: **temponym tagging with HeidelTime**

Temponym Tagging: Temporal Scopes for Textual Phrases

Outline

Outline

How dynamic is the Web?

Ntoulas et al. (2004) study dynamics of the Web 2002–2003

Data

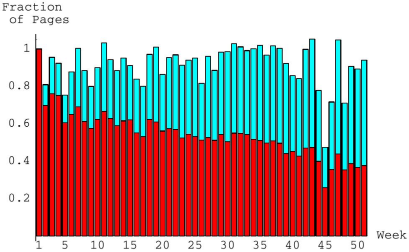
- weekly crawls of **154 web sites** over one year
- **top-ranked web sites** from topical categories in Google Directory (extension of DMOZ) from different top-level domains
- at most **200K web pages per web site** per weekly crawl

Domain	Fraction of pages in domain
.com	41%
.gov	18.7%
.edu	16.5%
.org	15.7%
.net	4.1%
.mil	2.9%
misc	1.1%

How dynamic is the Web?

Web pages

- on average **8% new web pages** per week
- **peek** in creation of new pages at the **end of each month**
- after **9 months** about **50% of web pages** have been deleted



week 1

- 4.8 M pages

week 45

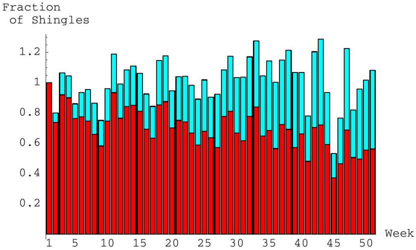
- one crawler crashed

work from 2004!

How dynamic is the Web?

content

- based on w-shingles (contiguous sequence of w words)
- **after one year** more than **50% of shingles** are still available
- each week about **5% of new shingles** are created



shingle size

- w = 50

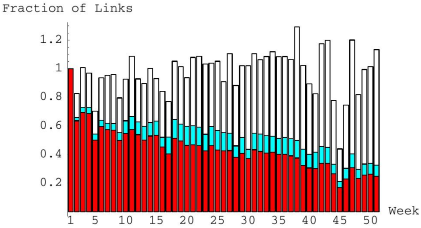
week 1

- 4.3 B unique shingles

How dynamic is the Web?

links

- after one year **only 24% of links** are still available
- on average **25% of new links** are created **every week**



red:

- first-week links

blue

- new links from 1st week pages

white

- new links from “new” pages

Dynamics and age

the Web is **highly dynamic**

- new content is continuously added
- old content is deleted and potentially lost forever

Web archives

- e.g., archive.org, internetmemory.org
- have been preserving **old snapshots** of web pages since 1996

improved digitalization

- e.g., using OCR (optical character recognition)
- have allowed (newspaper) archives to make old documents (e.g., from 1700s) searchable

Dynamics and age

several challenges

- How to index highly redundant document collections like web archives?
- How to make use of temporal information such as publication dates?
- How to search documents written in archaic language?

Outline

Indexing Redundant Content

Zhang & Suel (2007):
approach to index **highly-redundant document collections** (e.g., web archives)

main ideas:

- **break up documents** into shorter segments
- **segments should be shared** between overlapping documents
- use a **two-level index structure** to index associations between **words-and-segments** and **segments-and-documents**

$$d_1 \boxed{\begin{array}{c}\text{aac}\\ \text{bab}\\ \text{ccb}\end{array}} \rightarrow s_1 \boxed{\text{aac}},\ s_2 \boxed{\text{bab}},\ s_3 \boxed{\text{ccb}}$$

$$\boxed{\text{a}} \rightarrow \boxed{s_1, s_2, s_4, s_7, \ldots} \qquad \boxed{s_1} \rightarrow \boxed{d_1, d_3, d_9, \ldots}$$

Indexing Redundant Content

problem

how to break up documents into smaller segments so that segments are **shared between overlapping documents**

	naïve approach	better approach
d_1: aac bab ccb	aac / bab / ccb	a / acb / abc / cb
d_2: acb abc cba	acb / abc / cba	acb / abc / cb / a

Indexing Redundant Content

hash breaking (naïve approach)

- compute **hash code** h[i] for each term d[i] in document
- break document **at all indices i** such that h[i] % w = 0

Winnowing (as a better approach with guarantees)

- compute **hash code** h[i] for all subsequences d[i ... i+b-1] of length b
- **slide window** of size **w** over the array of hash codes h[0 .. |d|-b]
  - if h[i] is **strictly smaller** than all other values h[j] in current window, **cut the document** between i and i -1
  - if **multiple positions** i in the current window with minimal value h[i]
    - if we have previously cut directly before one of them, don't perform a cut
    - otherwise, cut before the rightmost position i having minimal value h[i]

Indexing Redundant Content

query processing

- **needs to be adapted** to reflect that the same segment can occur in many documents
- when **seeing a segment** in a posting list of the first index, **look up documents containing it** in the second index
- **effectiveness of skipping** for conjunctive queries is **reduced**
  - terms could be spread over different segments in a document
  - segments can be contained in documents with arbitrary document identifiers

Time travel text search

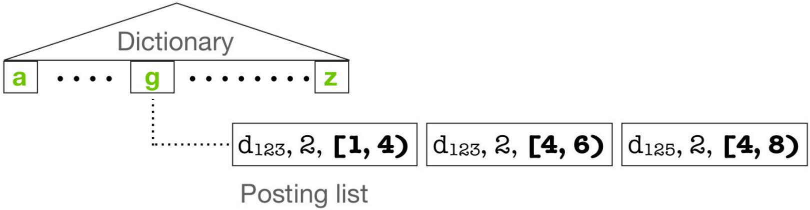
text search on **version document collections**

time-travel keyword query q@t

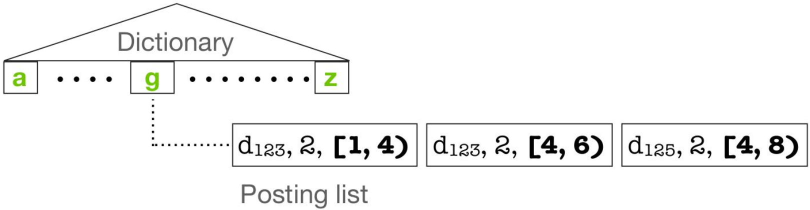
- combines keywords q with a time of interest t to **search "as of" the indicated time** in the past

time-travel inverted index

- adds a **valid-time interval [t_b, t_e)** to postings indicating when the information therein was current



Time travel text search



time-travel keyword query q@t

- **read posting lists** for keywords in q
- **filter out postings** whose valid-time interval does not contain t, i.e.: $t \notin [t_b, t_e)$

Outline

Time in queries

temporal information needs are frequent

query log analyses

- 1.5% queries with **explicit temporal intent** [Nunes et al. 2008]
- 7% queries with **implicit temporal intent** [Metzler et al. 2009]
- 13.8% **explicit**, 17.1% **implicit** [Zhang et al. 2010]

different types of temporal information in IR

- time as **dimension of relevance**
- time as **query topic**

Time in documents

Documents come with different kinds of **temporal information**

- **publication dates** (DCT): when document was published
- **temporal expressions**: time periods the document talks about

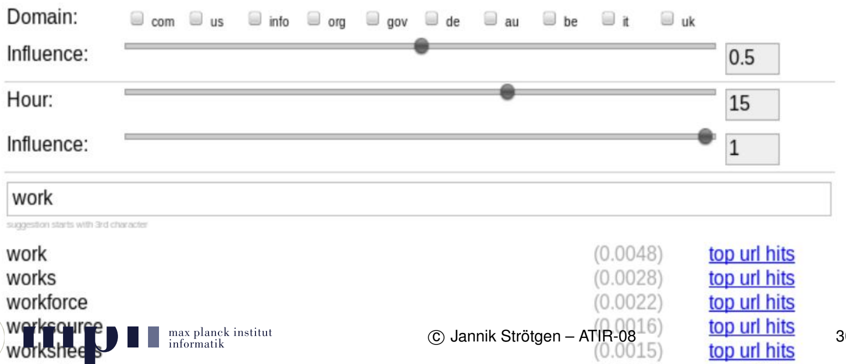
what is helpful depends on how time is used

Time as dimension of relevance

- temporal tagging is not needed
- **document creation time** and **query time** are utilized
- examples: news-related queries, freshness of search results
- besides improving search results:
  - query time for time-sensitive query auto-completion
    [Sengstock & Gertz 2011]

Time as dimension of relevance

- temporal tagging is not needed
- **document creation time** and **query time** are utilized
- examples: news-related queries, freshness of search results
- besides improving search results:
  - query time for time-sensitive query auto-completion [Sengstock & Gertz 2011]



Time as dimension of relevance

- temporal tagging is not needed
- **document creation time** and **query time** are utilized
- examples: news-related queries, freshness of search results
- besides improving search results:
  - query time for time-sensitive query auto-completion [Sengstock & Gertz 2011]

Time as dimension of relevance

- temporal tagging is not needed
- **document creation time** and **query time** are utilized
- examples: news-related queries, freshness of search results
- besides improving search results:
  - query time for time-sensitive query auto-completion [Sengstock & Gertz 2011]

suggestions for query "work..." at **different times**:

- 6am: workwear
- 3pm: workforce
- 9pm: workout

Time as query topic

- **temporal tagging is required**
- temporal information in the **content**
- document creation time is not meaningful
- example: queries with explicit time expressions

$$q_{text} = \langle \text{Germanwings} \rangle,\ q_{time} = [2015\text{-}03\text{-}01,\ 2015\text{-}04\text{-}30]$$

March 25, 2015

Germanwings plane crash: Leaders visit Alps site

The German, French and Spanish leaders have arrived together in the French Alps to visit the scene where a Germanwings plane crashed on **Tuesday**, killing all 150 on board.
...

November 10, 2015

Lufthansa tries to force striking staff back to work

The carrier confirmed **Tuesday** it had applied for a German court order ...
Lufthansa is still recovering from the blow that it suffered when disaster struck its subsidiary Germanwings in **March**. ...

left: http://www.bbc.com/news/world-europe-32046250 [last accessed: Nov 21, 2015]
right: http://money.cnn.com/2015/11/10/news/companies/lufthansa-strike-court-injunction [last accessed: Nov 21, 2015]

Source: Strötgen & Gertz 2016

Time as query topic

$$q_{text} = \langle \text{Germanwings} \rangle, \; q_{time} = [2015\text{-}03\text{-}01, 2015\text{-}04\text{-}30]$$

March 25, 2015	November 10, 2015
Germanwings plane crash: Leaders visit Alps site	**Lufthansa tries to force striking staff back to work**
The German, French and Spanish leaders have arrived together in the French Alps to visit the scene where a Germanwings plane crashed on **Tuesday**, killing all 150 on board. ...	The carrier confirmed **Tuesday** it had applied for a German court order ... Lufthansa is still recovering from the blow that it suffered when disaster struck its subsidiary Germanwings in **March**. ...

left: http://www.bbc.com/news/world-europe-32046250 [last accessed: Nov 21, 2015]
right: http://money.cnn.com/2015/11/10/news/companies/lufthansa-strike-court-injunction [last accessed: Nov 21, 2015]

Source: Strötgen & Gertz 2016

- relevant news document
- DCT in q_{time}
- Tuesday in q_{time}

- relevant news document
- DCT not in q_{time}
- Tuesday not in q_{time}
- March in q_{time}

Time in queries

queries can be **temporally classified** along several dimensions

- query can refer to a single or multiple time periods
  - **temporally unambiguous**
    (e.g., fifa world cup 2014, battle of waterloo)
  - **temporally ambiguous**
    (e.g., summer olympics, world war)
- time period is explicitly mentioned or implicitly assumed
  - **explicitly temporal**
    (e.g., fifa world cup 2014, presidential election 2016)
  - **implicitly temporal**
    (e.g., superbowl, london bombing)

Time in queries

queries can be **temporally classified** along several dimensions

- query aims for information about the past, present, or future
  - **past**
    (e.g., historic map of rome, news reports about moon landing)
  - **recent**
    (e.g., orlando shooting, tesla stock price)
  - **future**
    (e.g., uefa euro final, academy awards 2016)
- query can refer to any time period at all
  - **atemporal**
    (e.g., muffin recipe, side effects of paracetamol, muscle cramps)

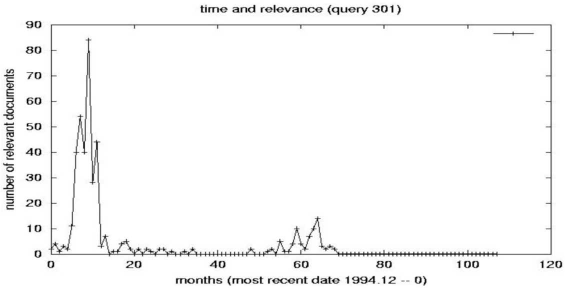
Outline

Temporal document priors

freshness of documents

Li and Croft (2003): approach based on language models for **queries favoring more recent documents**

- analysis of publication dates of relevant documents



recency query

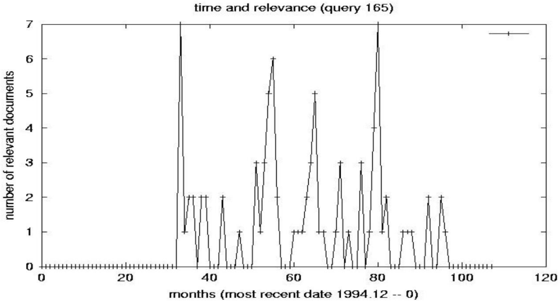
query 301: international organized crime

Temporal document priors

freshness of documents

Li and Croft (2003): approach based on language models for **queries favoring more recent documents**

- analysis of publication dates of relevant documents



rather uniform

query 165: Tobacco Company Advertising and the Young

Temporal document priors

freshness of documents – recency queries

- Query likelihood approach with **temporal document prior P[d]** depending on **DCT t** of a document and **current time c**

$$P[d|q] \propto P[d] \cdot \prod_v P[v|d] \qquad P[d] = \lambda e^{-\lambda(c-t)}$$

- typically: uniform prior probability P[d], i.e., P[d] can be ignored
- now: exponential distribution for prior probabilities, i.e., recent documents have higher prior probability P[d]

experiments show

ranking improvements – if applied on recency queries

Query classification

not all queries are equal

- treating every query as recency query decreases ranking quality
- it is important to distinguish queries

query logs can be analyzed to detect

implicitly temporal and atemporal queries;
temporally ambiguous and unambiguous queries

how?

question in assignment 5

Outline

Time as query topic

Berberich et al. (2010):

language modeling approach for **temporal information needs**

approach addresses main **shortcoming** of standard IR

- **temporal expressions** are treated as **terms**
- their **semantics is lost**

approach handles

- **explicitly temporal queries**

  i.e., queries with temporal expression

  e.g., “Michael Jordan 1990s”

Problems of standard IR approaches

temporal and geographic expressions

- (seem to be) treated as regular terms
- semantics is lost

→ **should be extracted and normalized**

query functionality

- how to search for time intervals?
- how to search for geographic regions?

→ **should be defined and provided**

results

- same ranking as for standard text search
- no time-/geo-centric exploration features

→ **special ranking is required**

→ **time-/geo-centric exploration should be possible**

max planck institut informatik
© Jannik Strötgen – ATIR-08 42 / 62

Time as query topic

temporal expressions are (often) **vague**

- precise time interval they refer to is uncertain
- this uncertainty needs to be reflected
- e.g., *in the 1990s* can refer to
  [1992, 1995], [1990, 1999], [1992, 1993], etc.

approach models temporal expressions as **sets of time intervals**

- temporal expressions as four-tuples (tb_l, tb_u, te_l, te_u)
- temporal expression T = (tb_l, tb_u, te_l, te_u) can **refer to any time interval** [tb, te] such that the following holds

$$tb_l \leq tb \leq tb_u \quad \wedge \quad tb \leq te \quad \wedge \quad te_l \leq te \leq te_u$$

Time as query topic

documents

- modeled as **set of textual terms** d_{text} and **set of temporal expressions** d_{time}

queries

- modeled as **set of textual terms** q_{text} and **set of temporal expressions** q_{time}

query-likelihood approach

- assumes **independence** between textual terms and temporal expressions

$$P[q|d] = P[q_{text}|d_{text}] \cdot P[q_{time}|d_{time}]$$

Time as query topic

query likelihood of **textual part** $P[q_{text}|d_{text}]$ estimated with

- **unigram language model** with Jelinek-Mercer smoothing

query likelihood of **temporal part** $P[q_{time}|d_{time}]$ estimated

- assuming independence between temporal expressions
- assuming uniform probability for temporal expressions from document d
- assuming uniform probability for time intervals Q can refer to
- assuming uniform probability for time intervals T can refer to

Berberich et al. (2010)'s evaluation shows

importance of treating time in a special way

Time as query topic

- independence between temporal expressions

$$P[q_{time}|d_{time}] = \prod_{Q \in q_{time}} P[Q|d_{time}]$$

- uniform probability for temporal expressions from d

$$P[Q|d_{time}] = \frac{1}{|d_{time}|} \sum_{T \in D_{time}} P[Q|T]$$

- uniform probability for time intervals Q can refer to

$$P[Q|T] = \frac{1}{|Q|} \sum_{[q_b, q_e] \in Q} P[[q_b, q_e]|T]$$

- uniform probability for time intervals T can refer to

$$P[[q_b, q_e]|T] = 1([q_b, q_e] \in T)$$

Time as query topic

Stroetgen & Gertz (2013): **proximity-aware ranking**

- no independence between terms and temporal expressions
- three dimensions: text, time, geo
  i.e., no independence between all three dimensions

Multi-word textual query:

- query: search engine

Document 1
....
....
....**search engine** ....
....
....

Document 2
......**search**............
....
....
.........**engine**........
....

Information need from last week

What did Alexander von Humboldt do

between **late 18th century**

and **early 19th century**

in Latein America?

Time as query topic

multidimensional query model with **query dimensions**

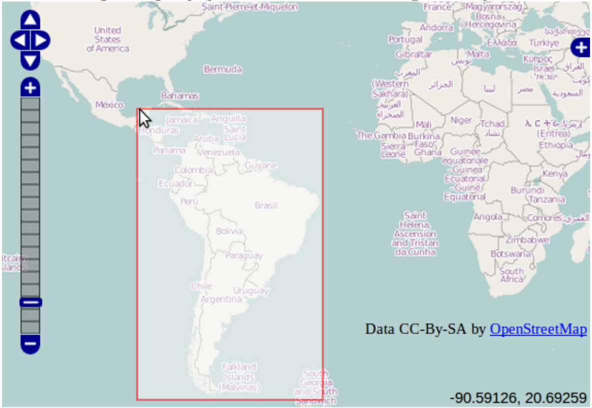
- **textual** query (q_{text})
- **temporal** query (q_{temp}): time intervals of interest
- **geographic** query (q_{geo}): regions of interest

Time as query topic

Textual search query: Alexander von Humboldt

Temporal query: [1794,1802]

Spatial query (press SHIFT to draw rectangle on map):

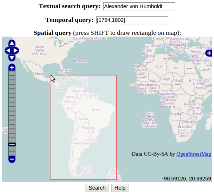


Search Help

Time as query topic

multidimensional query

- q_{text}: Alexander
- q_{temp}: late 18th – early 19th century
- q_{geo}: box(Latin America)

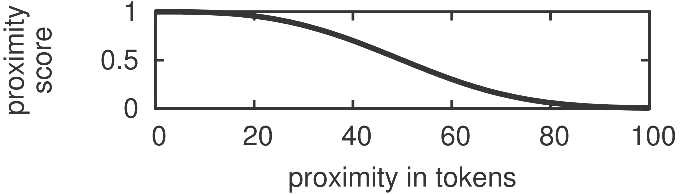


Document 1	Document 2
.... **Alexander** visited **Cuba** in **1800**.............	.... Paul visited **Cuba** in 2001 ....
....	....
...Until 2001	...Until **1800** ....
....	....
...brother of Paul	..brother of **Alexander**...

Time as query topic

term proximity score

- proximity of **terms satisfying all query dimensions**



final score

- textual, temporal, geographic scores
- **term proximity score**

Time as query topic

evaluation – data set

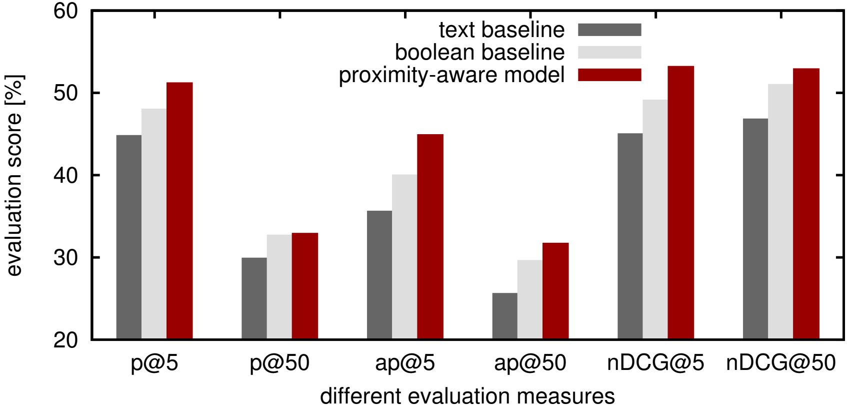
- NTCIR GeoTime [Gey et al. (2010)]

e.g., When and where did a volcano erupt in Africa during 2002?
$\rightarrow$ q_{text} = volcano erupt; q_{temp} = 2002; q_{geo} = box(Africa)

comparison

- **proximity-aware ranking approach**
- **text baseline**: text score (q_{text} = volcano erupt 2002 Africa)
- **boolean baseline**: text score & boolean filtering

Time as query topic



- boolean baseline outperforms text baseline
- proximity-aware model outperforms both baselines

Outline

Historical Document Collections

- improved **digitization methods** (e.g., OCR)
- (very) **old documents** now being **digitally available**

examples

- The New York Times Archive (1851 – today)
- The Times Archive (1785 – now)
- Google Books (1500 – now)
- HathiTrust (1500 – now)

Historical Document Collections

challenges and opportunities

- **unknown publication dates** of documents

  can be estimated based on similar documents with known publication dates

- **vocabulary gap** between today's queries and old documents

  needs to be bridged for effective information retrieval

- **longitudinal document collections** allow analyses that give insights into, e.g., the evolution of language and historic events

Historical Document Collections

IR on historical document collection suffers from **vocabulary gap** between today's queries and old documents

- **language evolution** (e.g., "and if he hear thee, thou wilt anger him")
- **terminology evolution** (e.g., Leningrad/Saint Petersburg)

Koolen et al. (2006) treat the problem as

- **cross-language information retrieval problem**
- translate documents using **rewriting rules** mined from the document collection

Historical Document Collections

phonetic sequence similarity

- **transcribe** historical and modern words **into phonemes**

  veeghen (historical) → v e g @ n

  vegen (modern) → v e g @ n
- find pairs of historical and modern word with same pronunciation
- split words into sequences of consonants and vowels

  historical: v ee gh e n

  modern: v e g e n
- **align sequences** and spot rewritings (e.g., ee → e, gh → g)
- rewritings that are **often observed** become **rewriting rules**

Summary

- **Temponyms**: phrases with temporal scopes
- **different aspects of time** can be distinguished in IR
- Web is **highly dynamic**
- **Temporal information** (e.g., publication dates and temporal expressions) can be leveraged for more effective IR
- **Web archives** keep often highly-similar old snapshots of web pages, allowing for efficient indexing and time-travel text search
- **Historical document collections** contain documents published long time ago,

  are challenging to search, but insightful to analyze

Thank you for your attention!

References

- Nunes et al. 2008: Use of Temporal Expressions in Web Search, ECIR.
- Metzler et al. 2009: Improving Search Relevance for Implicitly Temporal Queries, SIGIR.
- Zhang et al. 2010: Learning Recurrent Event Queries for Web Search, EMNLP.
- Pustejovsky et al. 2005: Temporal and Event Information in Natural Language Text, LRE journal.
- Kuzey et al. 2016a: As Time Goes By: Comprehensive Tagging of Textual Phrases with Temporal Scopes, WWW.
- Kuzey et al. 2016b: Temponym Tagging: Temporal Scopes for Textual Phrases, TempWeb.
- Ntoulas et al. 2004: What's New on the Web? The Evolution of the Web from a Search Engine Perspective, WWW.

References

- Zhang & Suel 2007: Efficient Search in Large Textual Collections with Redundancy, WWW.
- Sengstock & Gertz 2011: CONQUER: A System for Efficient Context-aware Query Suggestions, WWW.
- Strötgen & Gertz 2016: Domain-sensitive Temporal Tagging (M&CP, to appear).
- Li & Croft 2003: Time-Based Language Models, CIKM.
- Berberich et al. 2010: A Language Modeling Approach for Temporal Information Needs, ECIR.
- Strötgen & Gertz 2013: Proximity^2-aware Ranking for Textual, Temporal, and Geographic Queries, CIKM.
- Gey et al. 2010: NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search, NTCIR.
- Koolen et al. 2006: A Cross-Language Approach to Historic Document Retrieval, ECIR.

Thanks

some slides / examples are taken from / similar to those of:

- Klaus Berberich, Saarland University, previous ATIR lecture