Temporal Web Dynamics

Implications for Information Retrieval

Nattiya Kanhabua

1st ALEXANDRIA Workshop

L3S Research Center, Hannover, Germany

15 September 2014

Outline

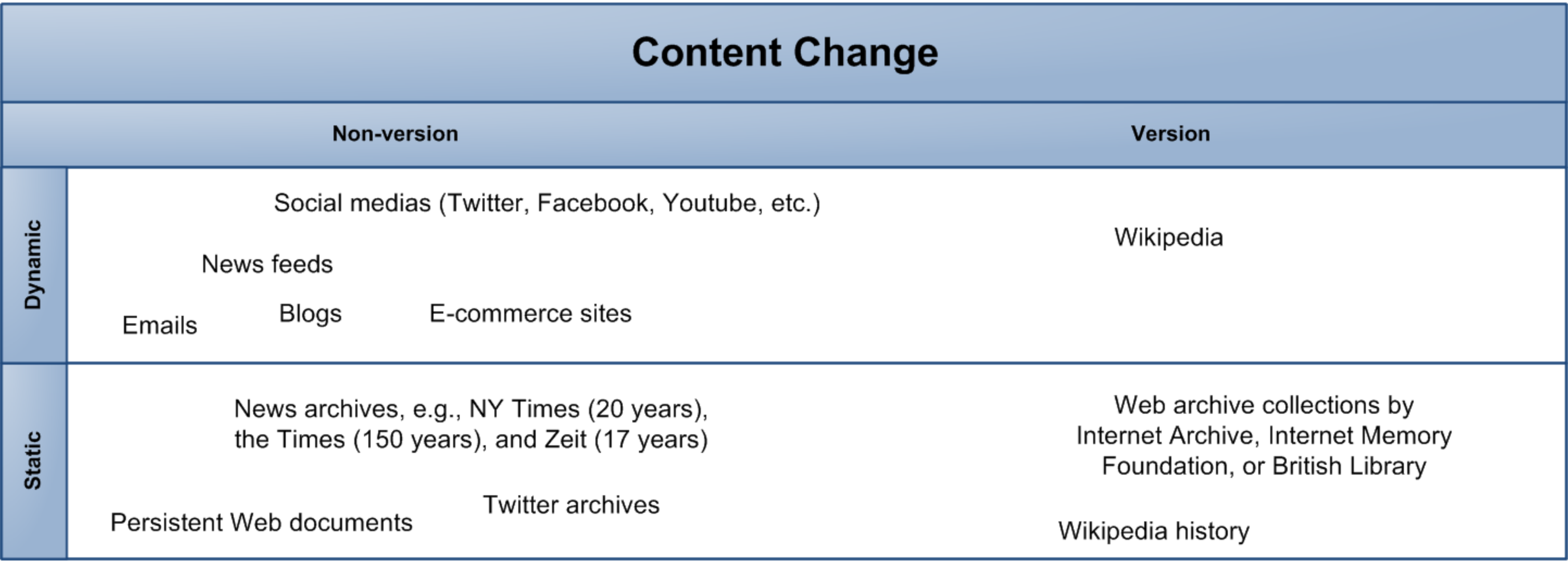
- *What* are temporal web dynamics?
- *Why* the dynamics impact search?
- Overview of time-aware approaches
  - Temporal Information Extraction
  - Temporal Query Analysis
  - Time-aware Retrieval and Ranking
- Conclusion and outlook

Temporal Web Dynamics

- Web is *changing* over time in many aspects, e.g., size, content, structure and how it is accessed by user interactions or queries.
  - – **Size**: web pages are added/deleted at all time
  - – **Content**: web pages are edited/modified
  - – **Query**: users' information needs changes

[Dumais, SIAM-SDM 2012; WebDyn 2010]
[Ke et al., CN 2006; Risvik et al., CN 2002]

Content/Structure Changes



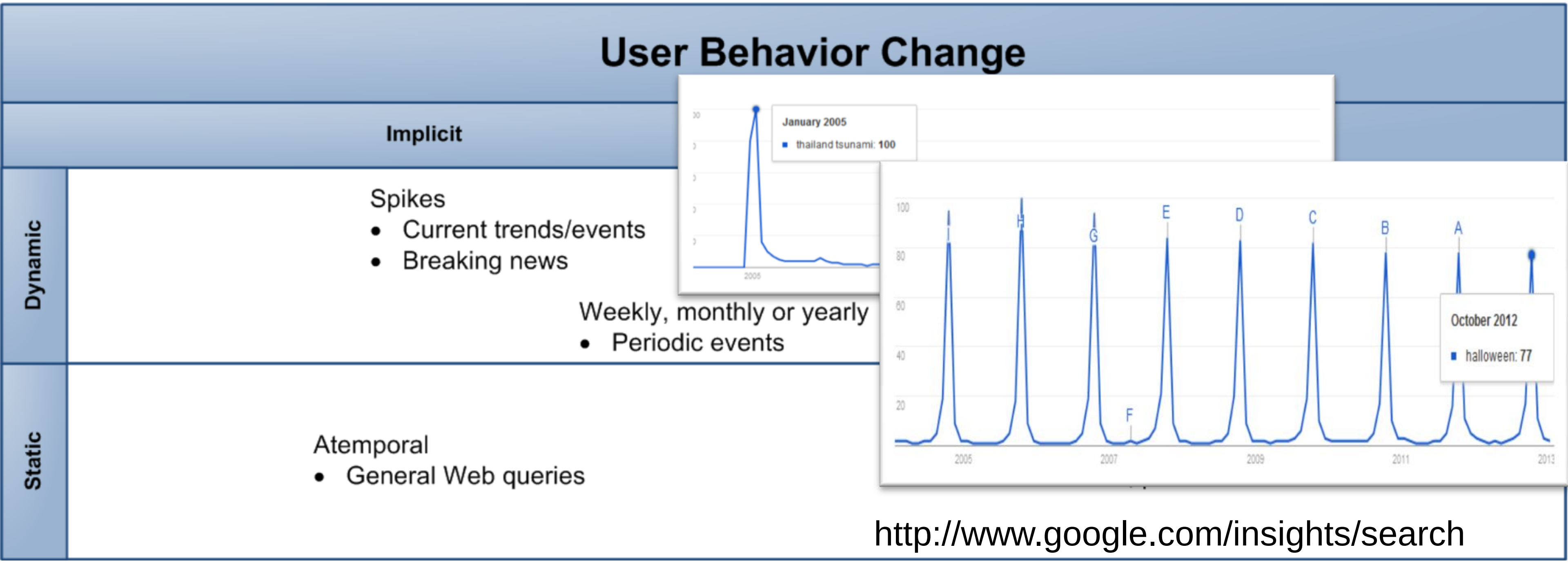
Content Change

	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) Twitter archives Persistent Web documents	Web archive collections by Internet Archive, Internet Memory Foundation, or British Library Wikipedia history

Fig. 1 Categorization of document collections with content changes over time.

Implications: Crawling, Indexing, Ranking

Changes in User Behavior



User Behavior Change

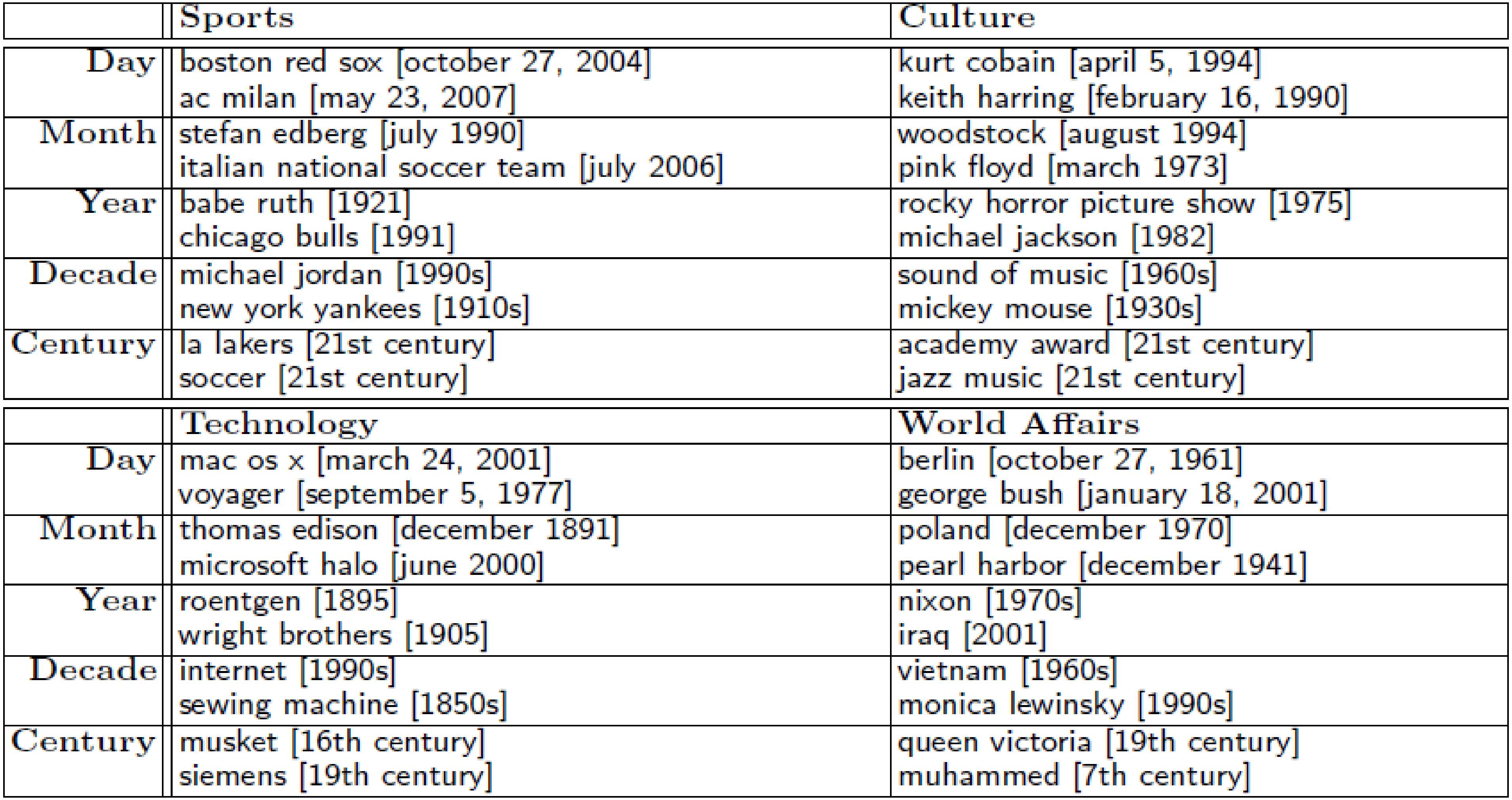
	Implicit
Dynamic	Spikes • Current trends/events • Breaking news Weekly, monthly or yearly • Periodic events
Static	Atemporal • General Web queries

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

Fig. 2 Categorization of queries with temporal information needs.

Implications: Query Analysis, Ranking

Temporal Query Examples

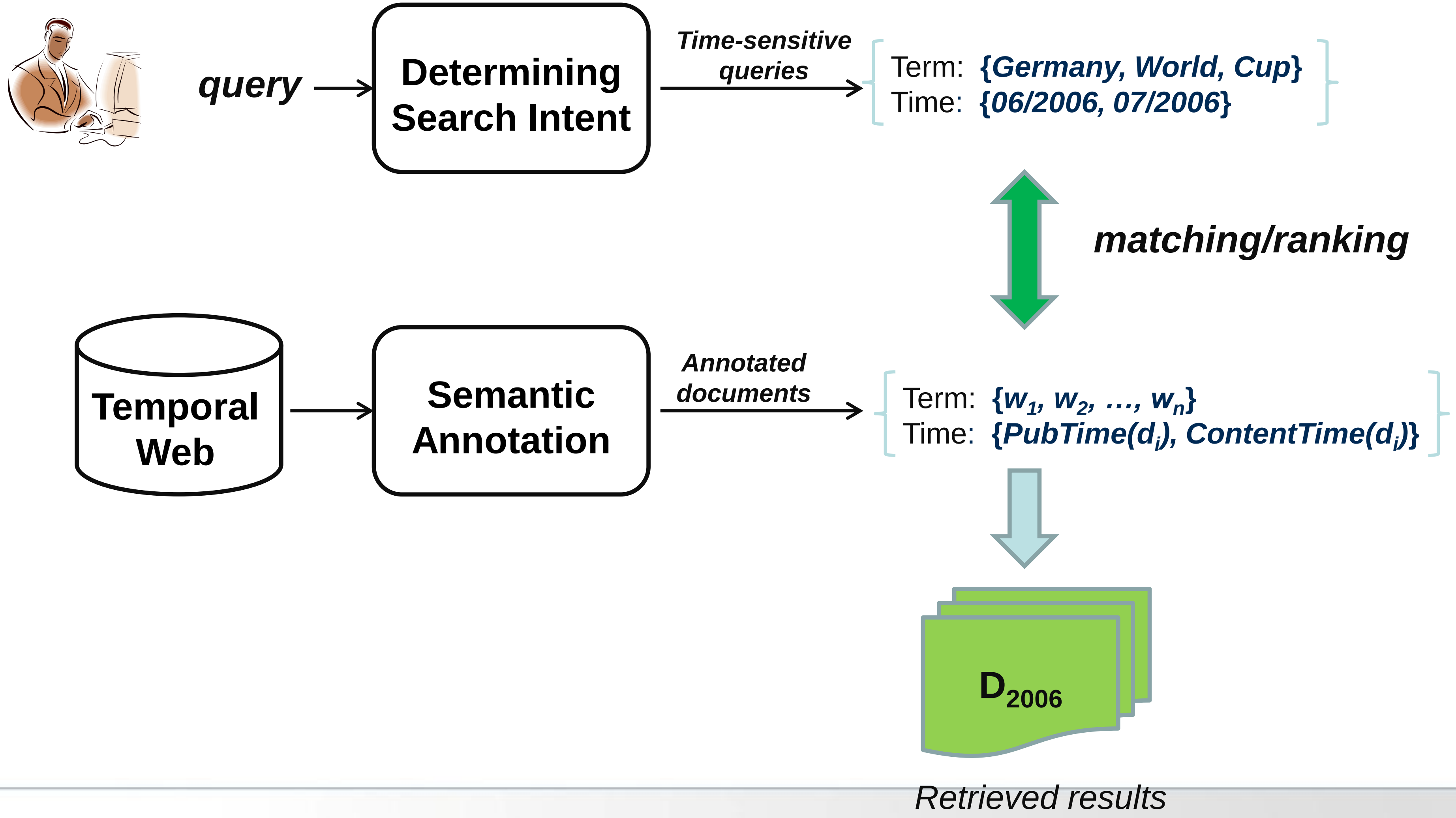


	Sports	Culture
Day	boston red sox [october 27, 2004] ac milan [may 23, 2007]	kurt cobain [april 5, 1994] keith harring [february 16, 1990]
Month	stefan edberg [july 1990] italian national soccer team [july 2006]	woodstock [august 1994] pink floyd [march 1973]
Year	babe ruth [1921] chicago bulls [1991]	rocky horror picture show [1975] michael jackson [1982]
Decade	michael jordan [1990s] new york yankees [1910s]	sound of music [1960s] mickey mouse [1930s]
Century	la lakers [21st century] soccer [21st century]	academy award [21st century] jazz music [21st century]

	Technology	World Affairs
Day	mac os x [march 24, 2001] voyager [september 5, 1977]	berlin [october 27, 1961] george bush [january 18, 2001]
Month	thomas edison [december 1891] microsoft halo [june 2000]	poland [december 1970] pearl harbor [december 1941]
Year	roentgen [1895] wright brothers [1905]	nixon [1970s] iraq [2001]
Decade	internet [1990s] sewing machine [1850s]	vietnam [1960s] monica lewinsky [1990s]
Century	musket [16th century] siemens [19th century]	queen victoria [19th century] muhammed [7th century]

[Berberich et al., ECIR 2010]

Implications for Search



query

Determining Search Intent

Time-sensitive queries

Term: {Germany, World, Cup}

Time: {06/2006, 07/2006}

matching/ranking

Temporal Web

Semantic Annotation

Annotated documents

Term: $\{w_1, w_2, \ldots, w_n\}$

Time: $\{PubTime(d_i), ContentTime(d_i)\}$

D_{2006}

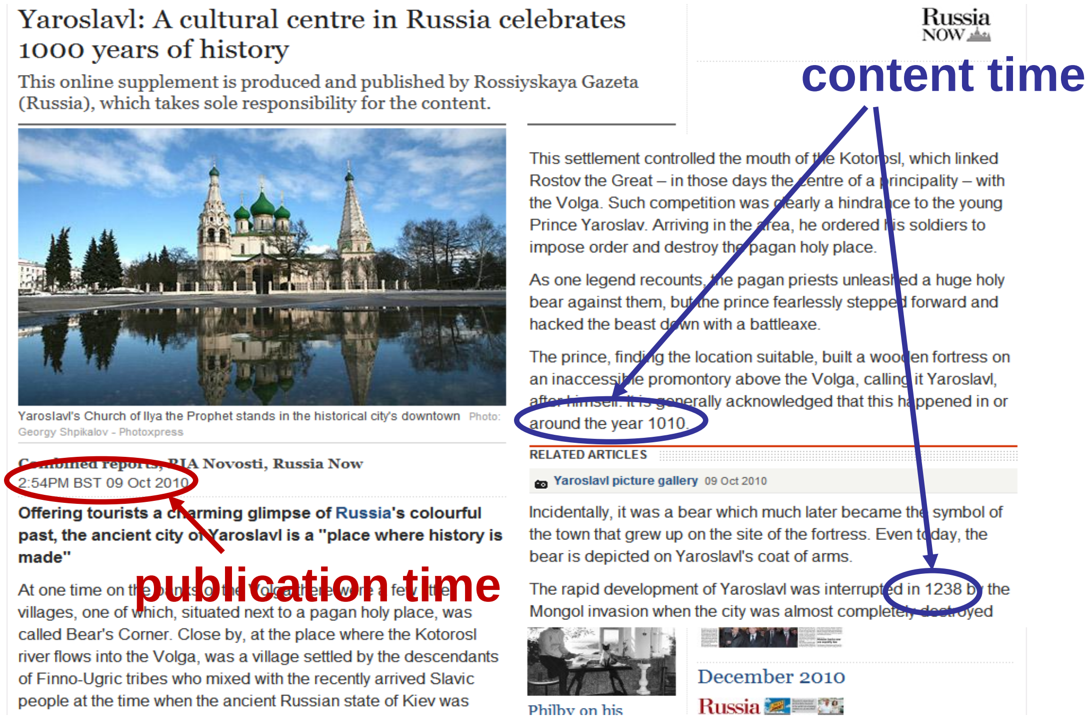
Retrieved results

Temporal Information Extraction

Two Time Aspects

Two time dimensions

1. Publication or modified time
2. Content or event time



Yaroslavl: A cultural centre in Russia celebrates 1000 years of history

This online supplement is produced and published by Rossiyskaya Gazeta (Russia), which takes sole responsibility for the content.

Russia NOW

Yaroslavl's Church of Ilya the Prophet stands in the historical city's downtown Photo: Georgy Shpikalov - Photoxpress

Combined reports, RIA Novosti, Russia Now
2:54PM BST 09 Oct 2010

publication time

Offering tourists a charming glimpse of Russia's colourful past, the ancient city of Yaroslavl is a "place where history is made"

At one time on the banks of the Volga there were a few little villages, one of which, situated next to a pagan holy place, was called Bear's Corner. Close by, at the place where the Kotorosl river flows into the Volga, was a village settled by the descendants of Finno-Ugric tribes who mixed with the recently arrived Slavic people at the time when the ancient Russian state of Kiev was

content time

This settlement controlled the mouth of the Kotorosl, which linked Rostov the Great – in those days the centre of a principality – with the Volga. Such competition was clearly a hindrance to the young Prince Yaroslav. Arriving in the area, he ordered his soldiers to impose order and destroy the pagan holy place.

As one legend recounts, the pagan priests unleashed a huge holy bear against them, but the prince fearlessly stepped forward and hacked the beast down with a battleaxe.

The prince, finding the location suitable, built a wooden fortress on an inaccessible promontory above the Volga, calling it Yaroslavl, after himself. It is generally acknowledged that this happened in or around the year 1010.

RELATED ARTICLES

Yaroslavl picture gallery 09 Oct 2010

Incidentally, it was a bear which much later became the symbol of the town that grew up on the site of the fortress. Even today, the bear is depicted on Yaroslavl's coat of arms.

The rapid development of Yaroslavl was interrupted in 1238 by the Mongol invasion when the city was almost completely destroyed

Philby on his

December 2010

Russia

Document Dating

Problem Statements

- Difficult to find the *trustworthy* time for web documents
  - Time gap between crawling and indexing
  - Decentralization and relocation of web documents
  - No standard metadata for time/date

“ For a given document with uncertain timestamp, can the contents be used to determine the timestamp with a sufficiently high confidence? ”

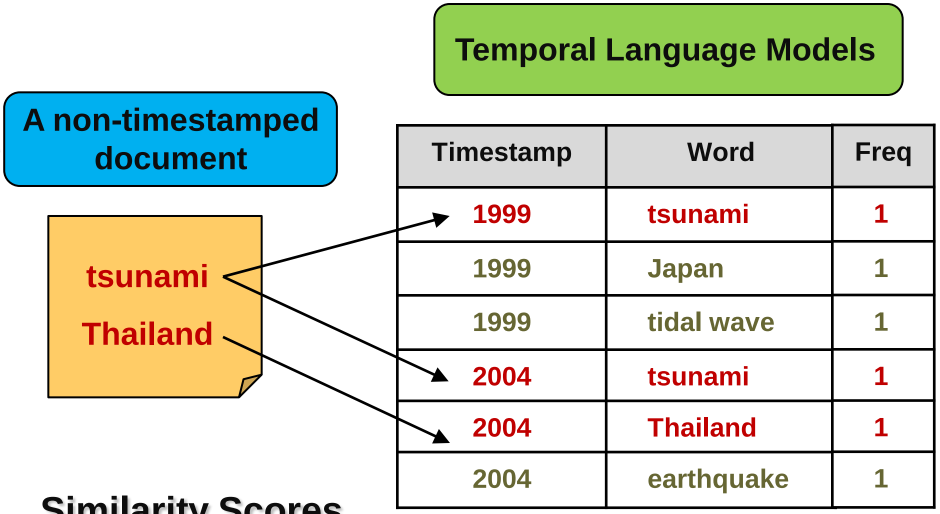
I found a bible-like document. But I have no idea when it was created?

Let’s me see…
This document is
probably
written in 850 A.C.
with 95% confidence.

Probabilistic Approach

Temporal Language Models

- Based on the statistic usage of words over time
- Compare each word of a non-timestamped document with a reference corpus
- Tentative timestamp -- a time partition mostly overlaps in word usage



Temporal Language Models

A non-timestamped document

tsunami

Thailand

Timestamp	Word	Freq
1999	tsunami	1
1999	Japan	1
1999	tidal wave	1
2004	tsunami	1
2004	Thailand	1
2004	earthquake	1

Similarity Scores

Score(1999) = 1

Score(2004) = 1 + 1 = 2

Most likely timestamp is 2004

$$Score(d_i, p_j) = \sum_{w \in d_i} P(w|d_i) \times \log \frac{P(w|p_j)}{P(w|C)}$$

[de Jong et al., AHC 2005; Kraaij, SIGIR Forum 2005; Kanhabua et al., ECDL 2008]

Extracting Content Time

- How to determine **relevant temporal expressions** tagged in a document?
  - *Not all* temporal expressions associated to an event *are equally relevant*

Reported by World Health Organization (WHO) on 29 July 2012 about an ongoing Ebola outbreak in Uganda since the beginning of July 2012

- Approaches: machine learning; rule-based

[Kanhabua et al., TAIA 2012; Strötgen et al., TempWeb 2012; Hoffart et al., AIJ 2012]

Temporal Query Analysis

Temporal Queries

- Temporal queries exist in the Web and archives
  - Relevancy is dependent on time
  - Documents are about events at particular time
  - Users: historians, librarians or journalists

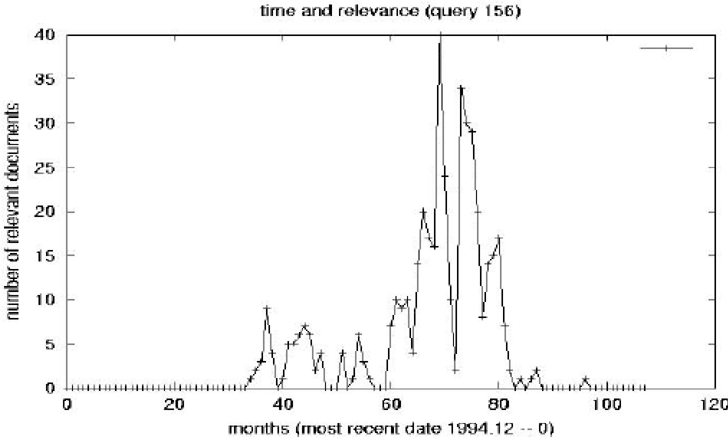


Figure 2.3: Query 156 “Efforts to Enact Gun Control Legislation”- Relevant documents mostly in the past.

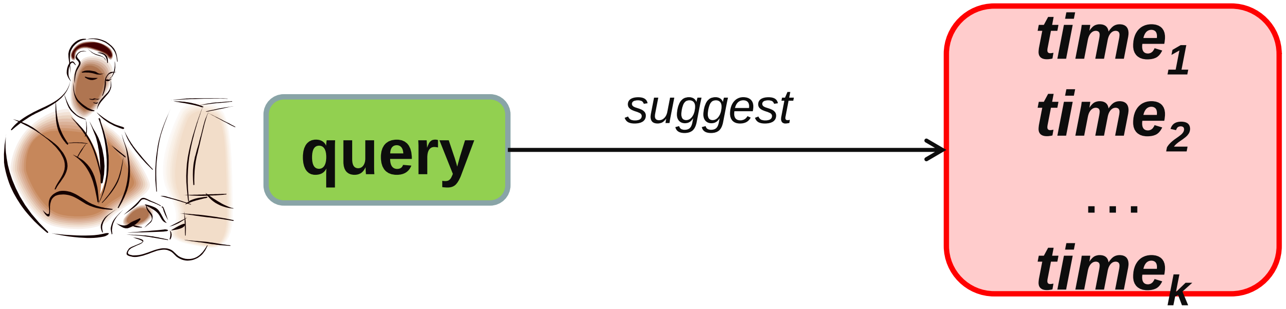
[Li et al., CIKM 2003; Jones and Diaz, ACM TOIS 2007; Berberich et al., ECIR 2010; Peetz et al., IR 2014]

Challenges

- Searching temporal document collections
  - E.g., digital libraries, web/news archives
- Problems: *semantic gaps* or lacking knowledge
  1. possibly **relevant time** of queries
  2. **terminology changes** over time

Challenges

- *Semantic gaps*: lacking knowledge about
  1. **possibly relevant time of queries**
  2. terminology changes over time



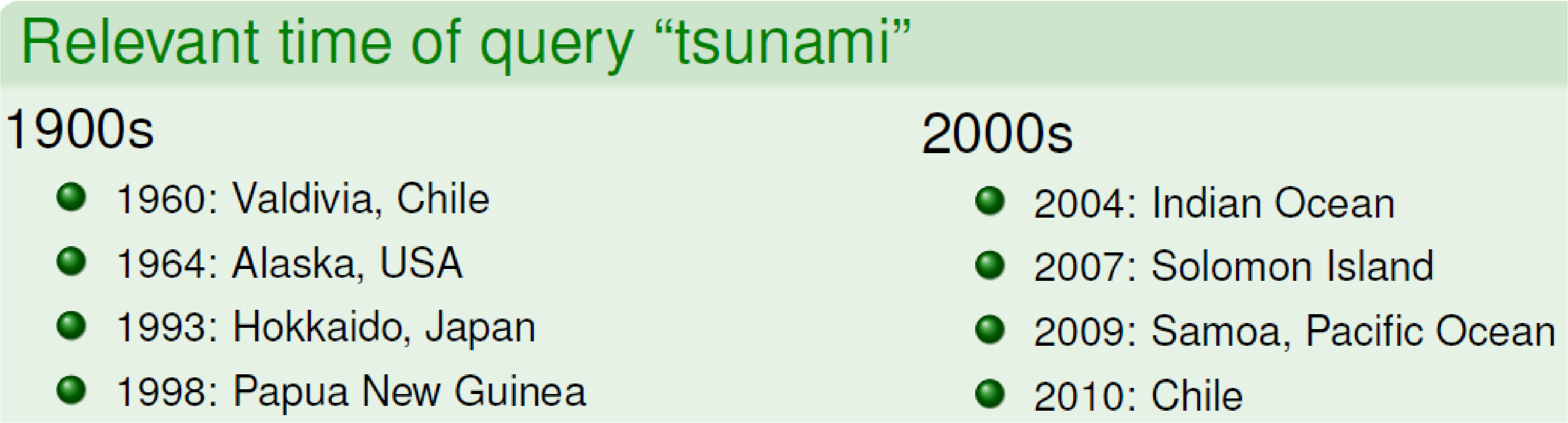
query

suggest

$time_1$
$time_2$
…
$time_k$

Challenges

- *Semantic gaps*: lacking knowledge about
  1. **possibly relevant time of queries**
  2. terminology changes over time



Relevant time of query “tsunami”

1900s

- 1960: Valdivia, Chile
- 1964: Alaska, USA
- 1993: Hokkaido, Japan
- 1998: Papua New Guinea

2000s

- 2004: Indian Ocean
- 2007: Solomon Island
- 2009: Samoa, Pacific Ocean
- 2010: Chile

How to determine the time of an implicit temporal query?

Current Approaches

1. Query log analysis
2. Search result analysis

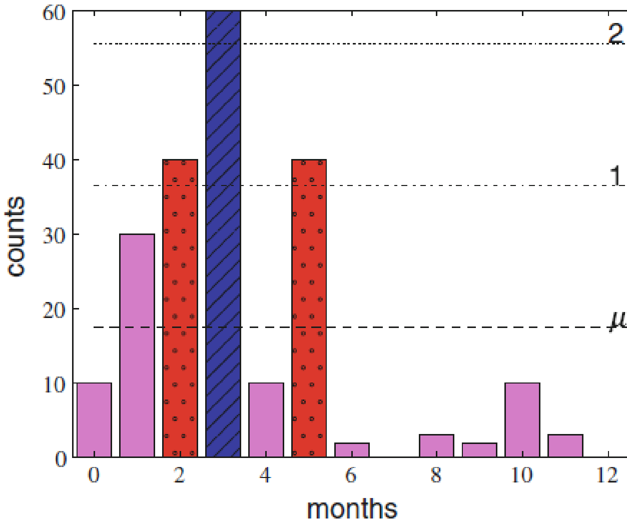
Query Log Analysis

- Mining query logs
  - Analyze query frequencies over time for identifying the *relevant time* of queries
  - Re-rank search results of implicit temporal queries using the determined time

[Metzler et al., SIGIR 2009; Zhang et al., EMNLP 2010]

Search Result Analysis

- Use temporal bursts for query modeling
  - Identify temporal bursts in the ranked lists of documents
  - Sample terms from the documents and update the query model



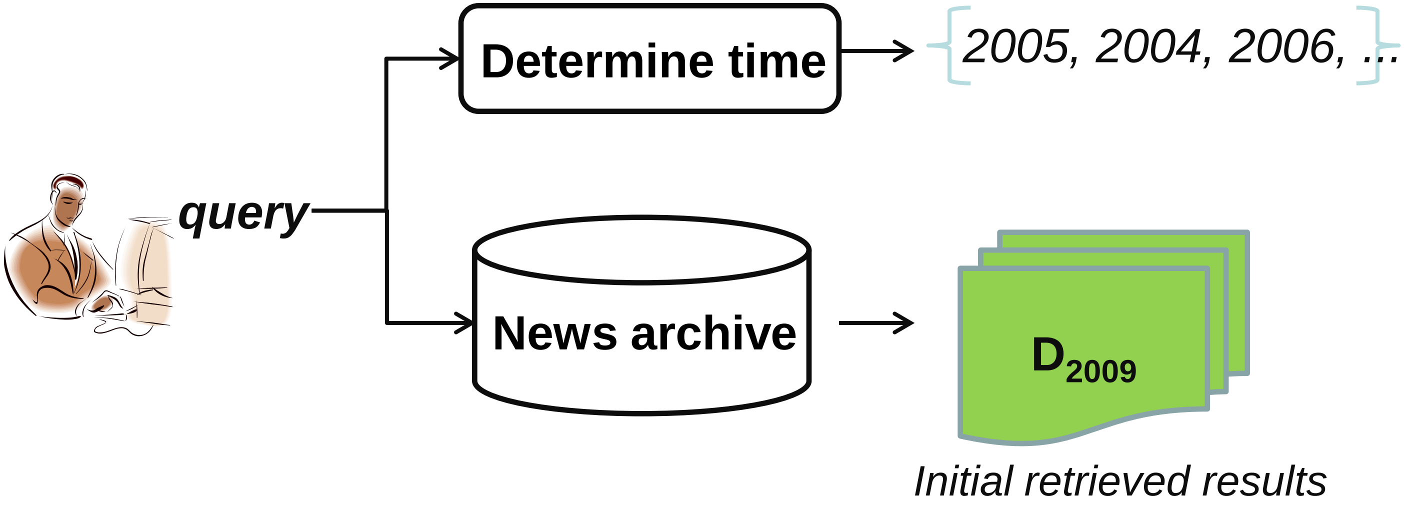
- Use temporal language models
  - Determine tentative time for a query
  - Re-rank search results using the determined time

[Kanhabua et al., ECDL 2010; Peetz et al., IR 2014]

Re-rank Search Results

- *Intuition:* documents published **closely to the time** of queries are more relevant
  - Assign ***document priors*** based on publication dates

$$S(q, d) = (1 - \alpha) \cdot S'(q_{\text{word}}, d_{\text{word}}) + \alpha \cdot S''(q_{\text{time}}, d_{\text{time}})$$

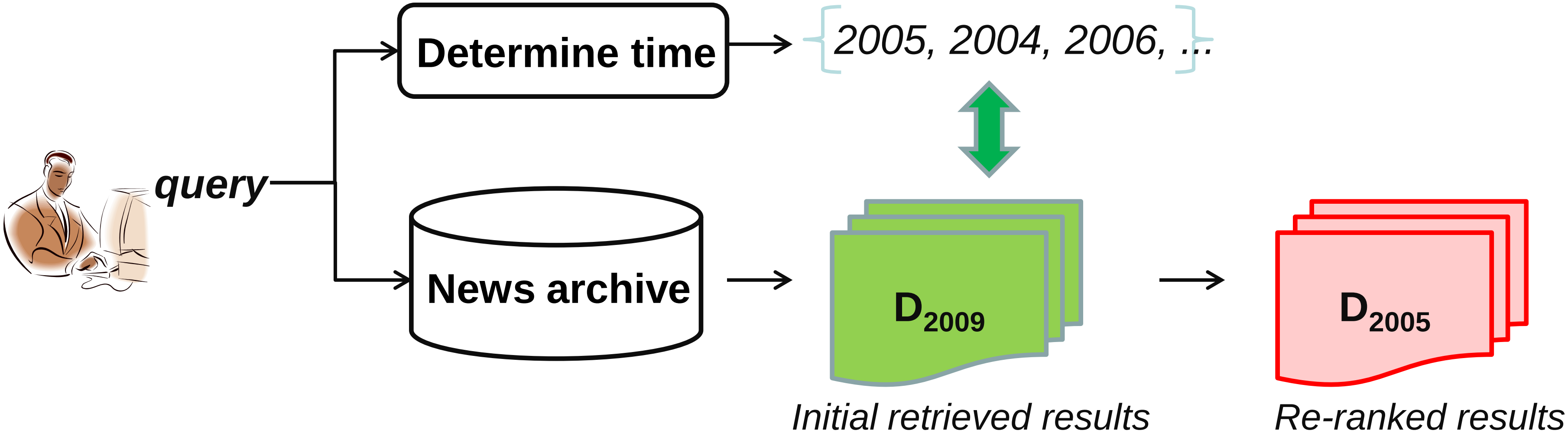


[Kanhabua et al., ECDL 2010]

Re-rank Search Results

- *Intuition:* documents published **closely to the time** of queries are more relevant
  - Assign ***document priors*** based on publication dates

$$S(q,d) = (1-\alpha) \cdot S'(q_{word}, d_{word}) + \alpha \cdot S''(q_{time}, d_{time})$$



query

Determine time

{2005, 2004, 2006, ...}

News archive

D_{2009}

Initial retrieved results

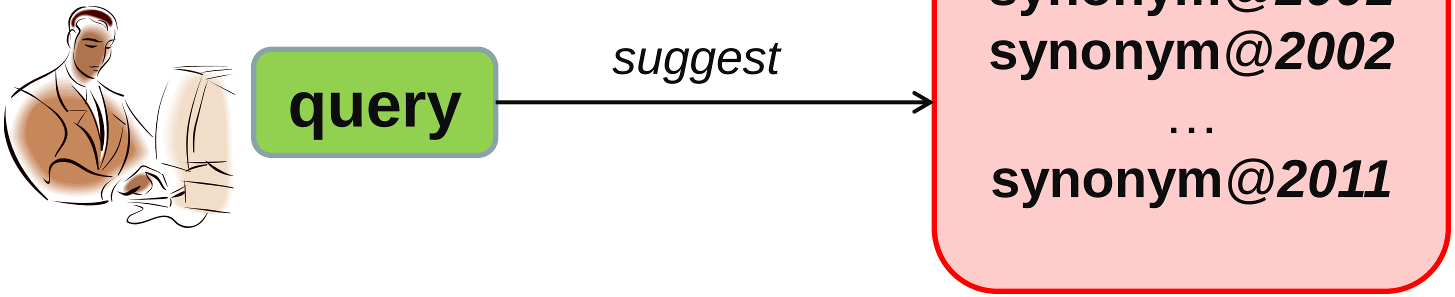
D_{2005}

Re-ranked results

[Kanhabua et al., ECDL 2010]

Challenges

- *Semantic gaps*: lacking knowledge about
  1. Possibly relevant time of queries
  2. **Named entity changes over time**



Named Entity Evolution

Problem Statements

- Queries of **named entities** (people, company, place)
  - Highly dynamic in appearance, i.e., relationships between terms changes over time
  - E.g. changes of roles, name alterations, or semantic shift

Named Entity Evolution

Problem Statements

- Queries of **named entities** (people, company, place)
  - Highly dynamic in appearance, i.e., relationships between terms changes over time
  - E.g. changes of roles, name alterations, or semantic shift

Scenario 1

Query: **“Pope Benedict XVI”** and written *before 2005*
Documents about **“Joseph Alois Ratzinger”** are relevant

Named Entity Evolution

Problem Statements

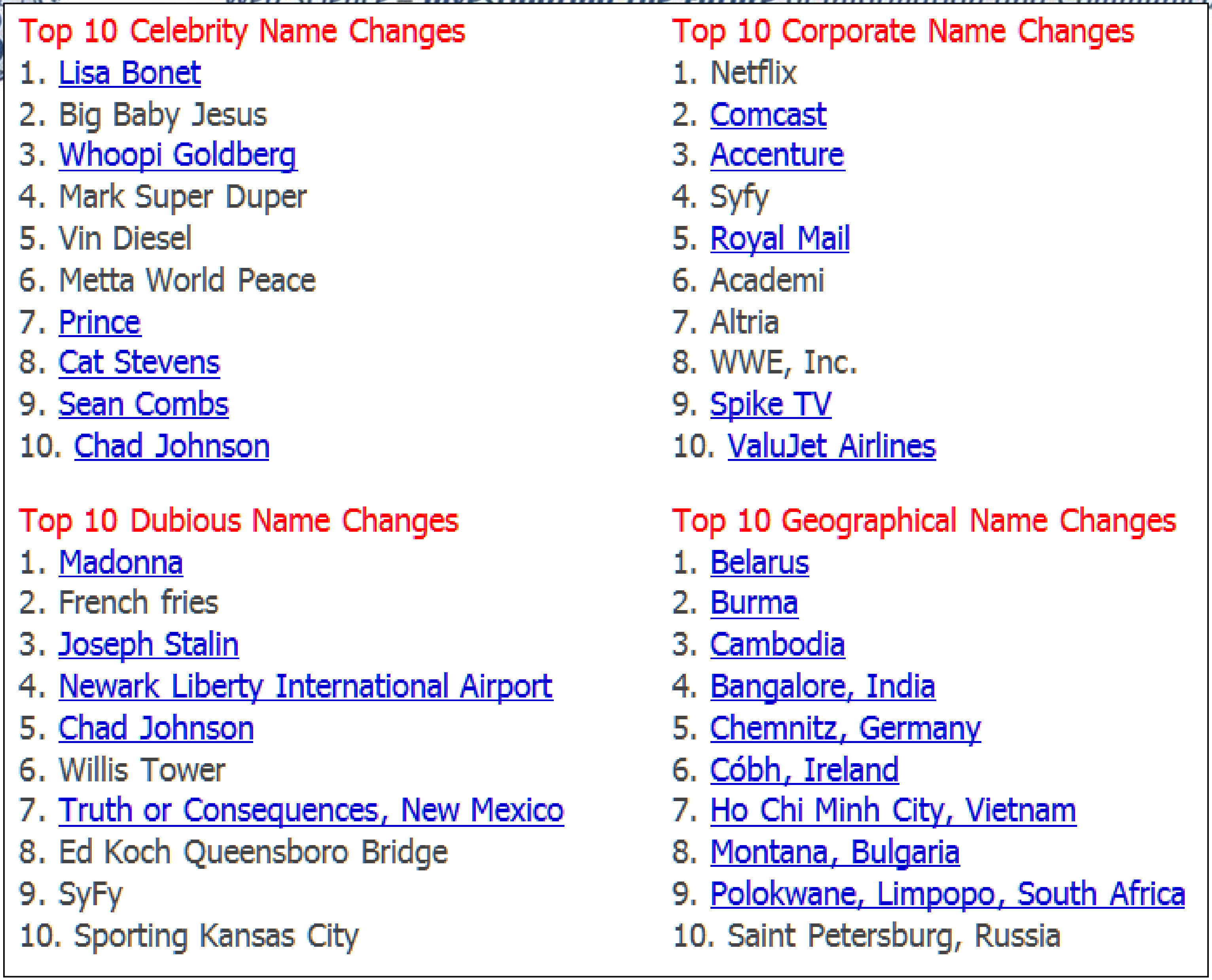
- Queries of **named entities** (people, company, place)
  - Highly dynamic in appearance, i.e., relationships between terms changes over time
  - E.g. changes of roles, name alterations, or semantic shift

Scenario 1

Query: **“Pope Benedict XVI”** and written *before 2005*
Documents about **“Joseph Alois Ratzinger”** are relevant

Scenario 2

Query: “**Hillary R. Clinton**” and written *from 1997 to 2002*
Documents about “**New York Senator**” and “**First Lady of the United States**” are relevant



Top 10 Celebrity Name Changes

1. Lisa Bonet
2. Big Baby Jesus
3. Whoopi Goldberg
4. Mark Super Duper
5. Vin Diesel
6. Metta World Peace
7. Prince
8. Cat Stevens
9. Sean Combs
10. Chad Johnson

Top 10 Corporate Name Changes

1. Netflix
2. Comcast
3. Accenture
4. Syfy
5. Royal Mail
6. Academi
7. Altria
8. WWE, Inc.
9. Spike TV
10. ValuJet Airlines

Top 10 Dubious Name Changes

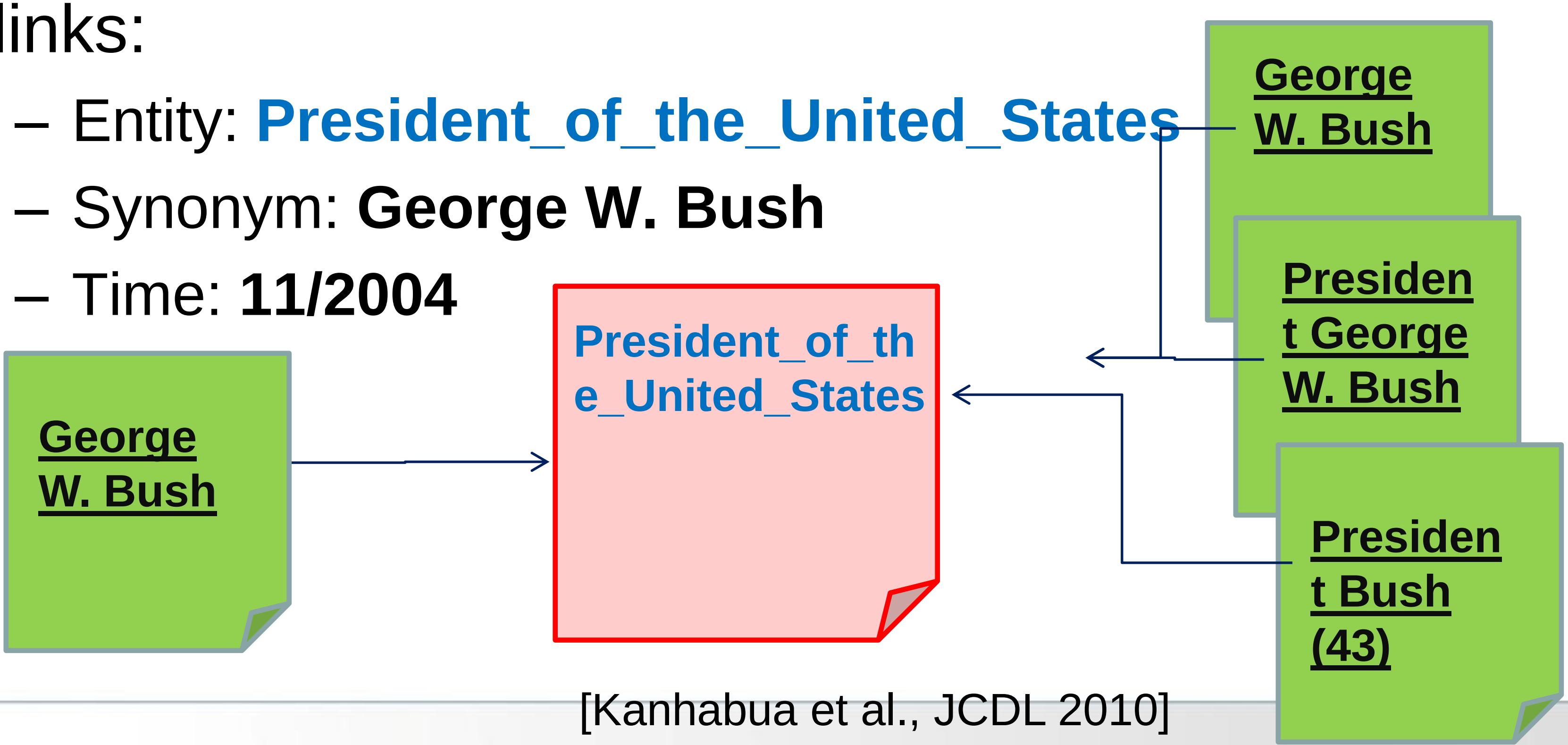
1. Madonna
2. French fries
3. Joseph Stalin
4. Newark Liberty International Airport
5. Chad Johnson
6. Willis Tower
7. Truth or Consequences, New Mexico
8. Ed Koch Queensboro Bridge
9. SyFy
10. Sporting Kansas City

Top 10 Geographical Name Changes

1. Belarus
2. Burma
3. Cambodia
4. Bangalore, India
5. Chemnitz, Germany
6. Cóbh, Ireland
7. Ho Chi Minh City, Vietnam
8. Montana, Bulgaria
9. Polokwane, Limpopo, South Africa
10. Saint Petersburg, Russia

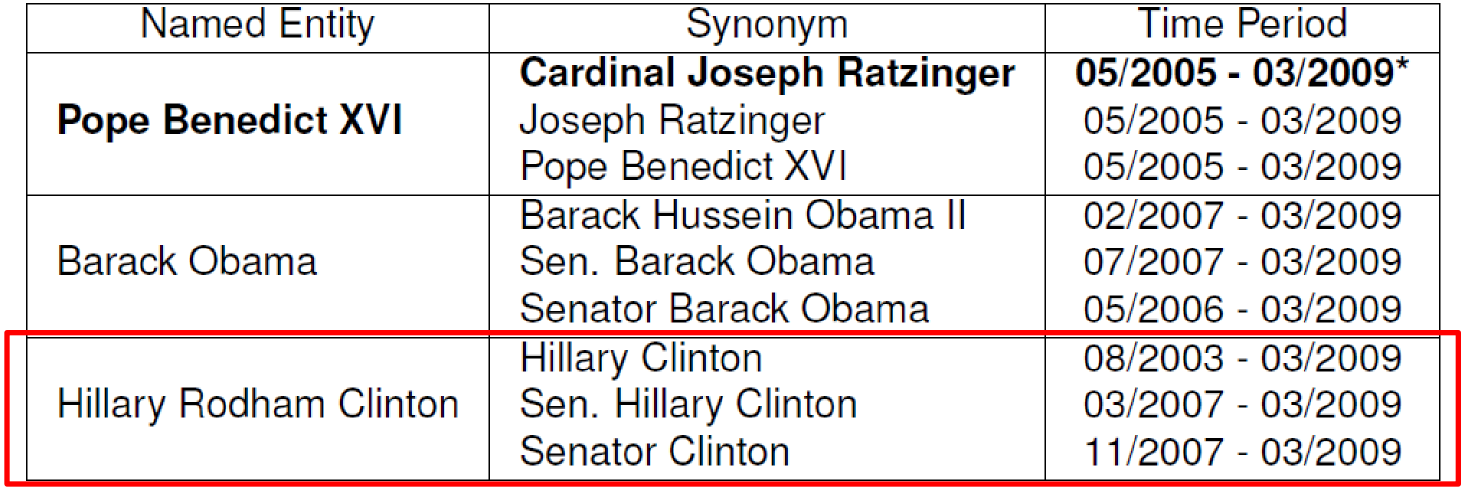
Find Temporal Synonyms

- Extract time-based synonyms from Wikipedia
- Find a set of **entity-synonym relationships** at *time* t_k
- For each $e_i \epsilon E_{tk}$, extract **anchor texts** from article links:
  - Entity: **President_of_the_United_States**
  - Synonym: **George W. Bush**
  - Time: **11/2004**



[Kanhabua et al., JCDL 2010]

Temporal Entity-Synonym



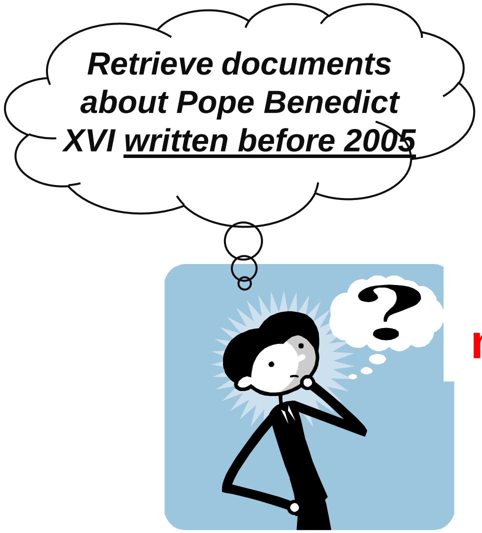
Named Entity	Synonym	Time Period
Pope Benedict XVI	**Cardinal Joseph Ratzinger**	**05/2005 - 03/2009***
	Joseph Ratzinger	05/2005 - 03/2009
	Pope Benedict XVI	05/2005 - 03/2009
Barack Obama	Barack Hussein Obama II	02/2007 - 03/2009
	Sen. Barack Obama	07/2007 - 03/2009
	Senator Barack Obama	05/2006 - 03/2009
Hillary Rodham Clinton	Hillary Clinton	08/2003 - 03/2009
	Sen. Hillary Clinton	03/2007 - 03/2009
	Senator Clinton	11/2007 - 03/2009

Note: the time of synonyms are timestamps of Wikipedia articles (8 years)

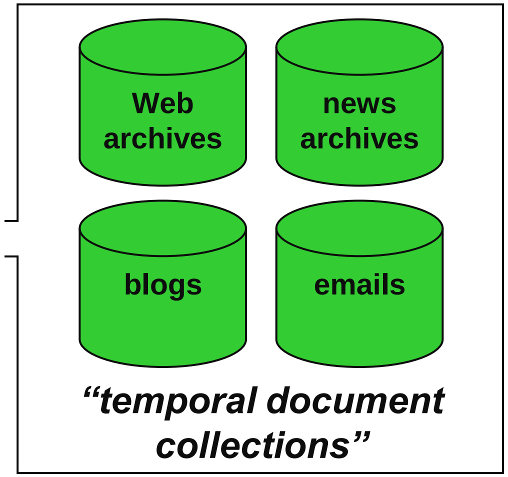
Time-aware Retrieval and Ranking

Searching the Past

- Time must be **explicitly modeled** in order to increase the effectiveness of *ranking*
  - To order search results so that the *most relevant* ones are ranked higher


Retrieve documents about Pope Benedict XVI written before 2005

Term-based IR approaches may give unsatisfied results



Query/Document Models

- A **temporal query** consists of:
  - Query keywords
  - Temporal expressions
- A **document** consists of:
  - Terms, i.e., bag-of-words
  - Publication time and temporal expressions

Time-aware Ranking Models

- Two main approaches
  1. Mixture model [Kanhabua et al., ECDL 2010]
     - Linearly combining *textual*- and *temporal* similarity
  2. Probabilistic model [Berberich et al., ECIR 2010]
     - Generating a query from the *textual part* and *temporal part* of a document independently

Mixture Model

- Linearly combine *textual*- and *temporal* similarity

$$S(q,d) = (1-\alpha)\cdot S'(q_{text}, d_{text}) + \alpha \cdot S''(q_{time}, d_{time})$$

  - *α* indicates the importance of similarity scores
    - Both scores are normalized before combining
  - Textual similarity can be determined using any term-based retrieval model
    - E.g., tf.idf or a unigram language model

Mixture Model

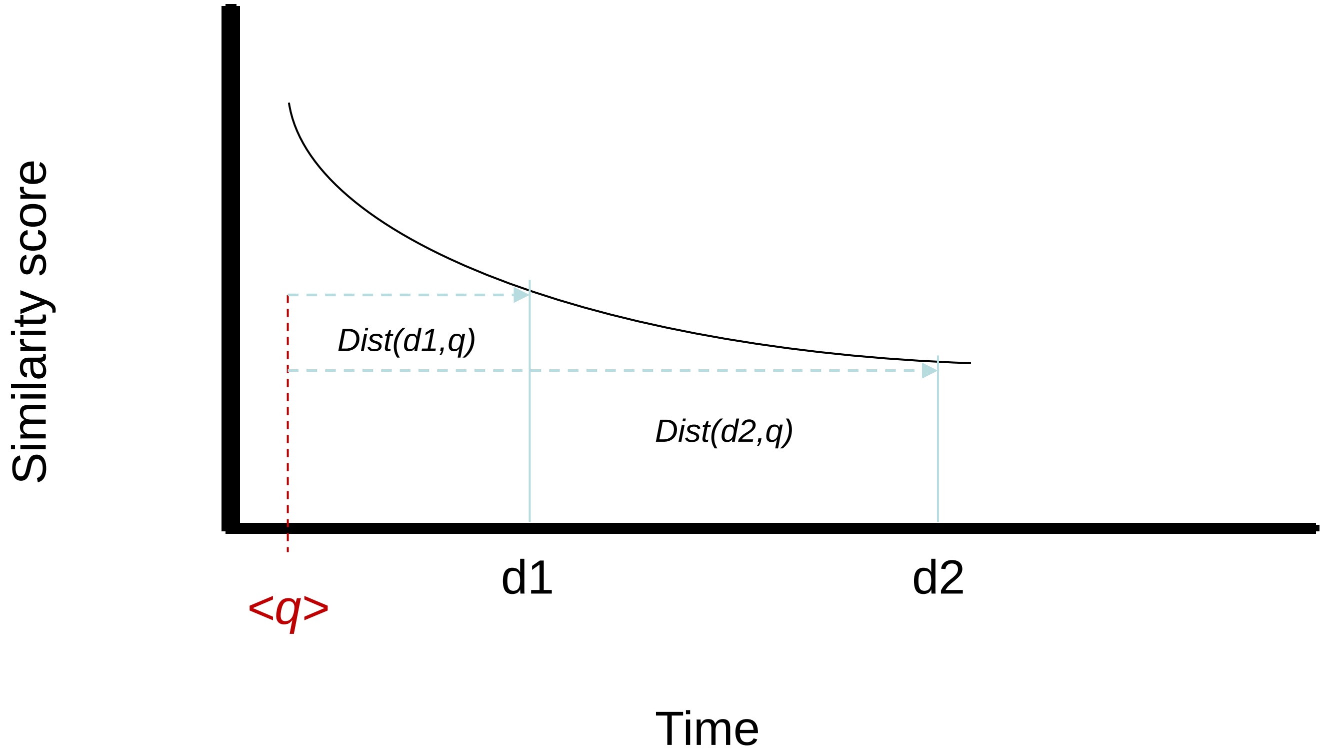
- Linearly combine *textual*- and *temporal* similarity

$$S(q,d) = (1-\alpha)\cdot S'(q_{text}, d_{text}) + \alpha \cdot S''(q_{time}, d_{time})$$

  - *α* indicates the importance of similarity scores
    - Both scores are normalized before combining
  - Textual similarity can be determined using any term-based retrieval model
    - E.g., tf.idf or a unigram language model

How to determine **temporal similarity**?

Temporal Similarity



[Kanhabua et al., ECDL 2010]

Conclusion and Outlook

- Temporal web dynamics and its impact
- State of the art temporal IR techniques
- Future work:
  - Search in versioned document collections
  - Efficient methods for document processing
  - Effective retrieval and ranking, e.g., return aggregated results or summaries
  - Support exploratory search in Web archives

References

- **[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
- **[Dumais, SIAM-SDM 2012]** Susan T. Dumais: Temporal Dynamics and Information Retrieval. SIAM-SDM 2012
- **[Hoffart et al., AIJ 2012]** Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Gerhard Weikum: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. 194: 28-61 (2013)
- **[de Jong et al., AHC 2005]** Franciska de Jong, Henning Rode, Djoerd Hiemstra: Temporal language models for the disclosure of historical text. AHC 2005: 161-168
- **[Kanhabua et al., JCDL 2010]** Nattiya Kanhabua, Kjetil Nørvåg: Exploiting time-based synonyms in searching document archives. JCDL 2010: 79-88
- **[Kanhabua et al., ECDL 2010]** Nattiya Kanhabua, Kjetil Nørvåg: Determining Time of Queries for Re-ranking Search Results. ECDL 2010: 261-272
- **[Kanhabua et al., TAIA 2012]** Nattiya Kanhabua, Sara Romano, Avaré Stewart: Identifying Relevant Temporal Expressions for Real-World Events. Time-aware Information Access Workshop 2012
- **[Ke et al., CN 2006]** Yiping Ke, Lin Deng, Wilfred Ng, Dik Lun Lee: Web dynamics and their ramifications for the development of Web search engines. Computer Networks 50(10): 1430-1447 (2006)
- **[Kraaij, SIGIR Forum 2005]** Wessel Kraaij: Variations on language modeling for information retrieval. SIGIR Forum 39(1): 61 (2005)
- **[Li et al., CIKM 2003]** Xiaoyan Li, W. Bruce Croft: Time-based language models. CIKM 2003: 469-475

References (cont')

- **[Metzler et al., SIGIR 2009]** Donald Metzler, Rosie Jones, Fuchun Peng, Ruiqiang Zhang: Improving search relevance for implicitly temporal queries. SIGIR 2009: 700-701
- **[Peetz et al., IR 2014]** Maria-Hendrike Peetz, Edgar Meij, Maarten de Rijke: Using temporal bursts for query modeling. Inf. Retr. 17(1): 74-108 (2014)
- **[Risvik et al., CN 2002]** Knut Magne Risvik, Rolf Michelsen: Search engines and Web dynamics. Computer Networks 39(3): 289-302 (2002)
- **[Strötgen et al., TempWeb 2012]** Jannik Strötgen, Omar Alonso, Michael Gertz: Identification of top relevant temporal expressions in documents. Temporal Web Workshop 2012.
- **[WebDyn 2010]** Web Dynamics course: http://www.mpi-inf.mpg.de/departments/d5/teaching/ss10/dyn/, Max-Planck Institute for Informatics, Saarbrücken, Germany, 2010
- **[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