Temporal Web Dynamics

Implications from Search Perspective

Speaker: Nattiya Kanhabua
Web Science Course
KBS, University of Hannover
6 May 2014
Outline

• Temporal Web Dynamics

• Research Problems
  – Temporal Information Extraction
  – Temporal Query Analysis
  – Time-aware Retrieval and Ranking

• Application to Temporal Search
Temporal Web Dynamics

• Web is *changing* over time in many aspects:
  – **Size**: web pages are added/deleted all the 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]
Web and Index Sizes

http://www.worldwidewebsize.com/

2000
First billion-URL index
The world’s largest!
≈5000 PCs in clusters!

2004
Index grows to 4.2 billion pages

2008
Google counts 1 trillion unique URLs

2009
TBs or PBs of data/index
Tens of thousands of PCs

2012

Impacts: crawling, indexing, and caching
Content Dynamics

• WayBack Machine
  – Web archive search by the Internet Archive
Content Dynamics

The Information Retrieval Group

The Information Retrieval group led by Keith van Rijsbergen has a vigorous programme of research, aimed at developing novel, effective, and efficient retrieval models of information. The group plays a leading role in the International Information Retrieval Community and has set trends in many aspects of this research. The IR group at Glasgow is one of the oldest and most active research centres in the world.

The group, part of the Department of Computing Science at the University of Glasgow, has a long and strong research history in a wide range of topics in information retrieval, stemming from seminal work on the IR system SMART. The group's interests include multimedia information retrieval, and the development of novel interactive techniques. Our research preserves a strong emphasis on practical solutions for large-scale document collections, and maintains strong links with researchers in Machine Learning and Information Retrieval, as well as with industry through technology transfer. Members of the group have also been in organizing major conferences, workshops and summer schools in the area of information retrieval.

Research Projects:
- Terrier
- PEER
- Resource Finder
- Content-based Retrieval
- Navigation

For more information, visit: http://www.ics.glu/ir

1998

2006

Impacts: document representation and retrieval

2012
Query Dynamics

• Search queries exhibit temporal patterns
  – Spikes or seasonality

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

Impacts: search intent and query representation
### Temporal Query Examples

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

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

[Berberich et al., ECIR 2010]
Query/Document Matching

**Determining Search Intent**
- **Term:** \{Germany, World, Cup\}
- **Time:** \{06/2006, 07/2006\}

**Semantic Annotation**
- **Term:** \{w_1, w_2, ..., w_n\}
- **Time:** \{PubTime(d_i), ContentTime(d_i)\}

**Temporal Web**

**Matching**

Retrieved results

\[D_{2006}\]
Temporal Information Extraction
Two Time Aspects

Two time dimensions

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

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

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

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Word</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>tsunami</td>
<td>1</td>
</tr>
<tr>
<td>1999</td>
<td>Japan</td>
<td>1</td>
</tr>
<tr>
<td>1999</td>
<td>tidal wave</td>
<td>1</td>
</tr>
<tr>
<td>2004</td>
<td>tsunami</td>
<td>1</td>
</tr>
<tr>
<td>2004</td>
<td>Thailand</td>
<td>1</td>
</tr>
<tr>
<td>2004</td>
<td>earthquake</td>
<td>1</td>
</tr>
</tbody>
</table>

Similarity Scores
Score(1999) = 1
Score(2004) = 1 + 1 = 2

Most likely timestamp is 2004

[de Jong et al., AHC 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
Approach

• The task of identifying relevant time is regarded as a classification problem
  – Two classes: (1) relevant and (2) irrelevant

• Definition: relevant if overlaps the starting, ending or ongoing time of the event

• Machine learning: three classes of features
  – Sentence-based features
  – Document-based features
  – Corpus-specific features

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

• Sentence-based
  – $sen\text{Len}$, $sen\text{Pos}$, $is\text{Context}$, $cnt\text{EntityInS}$, $cnt\text{TExpInS}$, $cnt\text{TPointInS}$, $cnt\text{TPeriodInS}$, $entity\text{Pos}$, $entity\text{PosDist}$, $TExp\text{Pos}$, $TExp\text{PosDist}$, $time\text{Dist}$, $entityTExp\text{PosDist}$

• Document-based
  – $cnt\text{EntityInD}$, $cnt\text{EntitySen}$, $cnt\text{TExpInD}$, $cnt\text{TPointInD}$, $cnt\text{TPeriodInD}$

• Domain-specific
  – $is\text{Neg}$, $is\text{History}$

[Kanhabua et al., TAIA 2012]
Temporal Query Analysis
Determining Search Intent

• Two types of temporal queries:
  1. **Explicit**: time is provided, “US President 2012“
  2. **Implicit**: time is *not* provided, "Germany FIFA World Cup"
      • Temporal intent can be implicitly inferred

• Previous studies on temporal queries:
  – 1.5% of web queries are *explicit*
  – ~7% of web queries are *implicit*

[Nunes et al., ECIR 2008; Metzler et al., SIGIR 2009]
Query Log Analysis

- Leverage real-world query logs
  - Search query frequencies over time

- Apply time-series analysis
  - Time-series decomposition for detecting seasonal queries

[Metzler et al., SIGIR 2009; Shokouhi, SIGIR 2011]
Time-series Decomposition

Query: Easter
Matching: Re-visited

**Temporal Web**

- **Query**
  - Determining Search Intent
  - Time-sensitive Queries
    - Term: \{Germany, World, Cup\}
    - Time: \{06/2006, 07/2006\}

- **Annotated Documents**
  - Semantic Annotation
    - Term: \{w_1, w_2, ..., w_n\}
    - Time: \{PubTime(d_i), ContentTime(d_i)\}

- **Matching**

  - Retrieved results
  - Ranked results

\(D_{2006}\)
Time-aware Retrieval and Ranking
Searching the Past

- Searching documents created/edited over time
  - E.g., web archives, news archives, blogs, or emails
  - A journalist wants to write a timeline of a news article
  - A Wikipedia contributor searches for historical information about an entity of interests

Term-based IR approaches may give unsatisfied results
Challenges

• 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

• **Time uncertainty** should be taken into account
  – Two temporal expressions can refer to *the same* time period even though they are *not equally written*
  
  – E.g. the query **“Independence Day 2011”**
    • A retrieval model relying on *term-matching only will fail* to retrieve documents mentioning “July 4, 2011”
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

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

- \(\alpha\) 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?

[Kanhabua et al., ECDL 2010]
Assume that temporal expressions in the query are generated independently from a two-step generative model:

- $P(t_q | t_d)$ can be estimated based on publication time using an exponential decay function [Kanhabua et al., ECDL 2010]

- Linear interpolation smoothing is applied to eliminate zero probabilities

I.e., an unseen temporal expression $t_q$ in $d$

Similarity score

<table>
<thead>
<tr>
<th>Time</th>
<th>Dist(d1, q)</th>
<th>Dist(d2, q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Kanhabua et al., ECDL 2010]
Temporal Similarity

• Assume that temporal expressions in the query are generated independently from a two-step generative model:

\[
S''(q_{time}, d_{time}) = \prod_{t_q \in q_{time}} P(t_q | d_{time}) = \prod_{t_q \in q_{time}} \left( \frac{1}{|d_{time}|} \sum_{t_d \in d_{time}} P(t_q | t_d) \right)
\]

– \( P(t_q | t_d) \) can be estimated based on publication time using an exponential decay function

– Linear interpolation smoothing is applied to eliminates zero probabilities
  • I.e., an unseen temporal expression \( t_q \) in \( d \)

[Kanhabua et al., ECDL 2010]
Application to Temporal IR
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
# Examples of Name Changes

### 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

*QUEST Demo: [http://research.idi.ntnu.no/wislab/quest/](http://research.idi.ntnu.no/wislab/quest/)*
Current Approaches

- Temporal co-occurrence
- Temporal association rule mining
- Temporal knowledge extraction
  - Ontology
  - Wikipedia history

[Berberich et al., WebDB 2009; Kanhabua et al., JCDL 2010]
[Kaluarachchi et al., CIKM 2010; Tahmasebi et al., COLING 2012]
Temporal Co-occurrence

• Temporal co-occurrence
  – Measure the degree of relatedness of two entities at different times by comparing term contexts
  – Require a recurrent computation at querying time, which reduce efficiency and scalability

[Berberich et al., WebDB 2009]
Association Rule Mining

• Temporal association rule mining
  – Discover semantically identical concepts (or named entities) that are used in different time
  – Two entities are semantically related if their associated events occur multiple times in a collection
  – Events are represented as sentences containing a subject, a verb, objects, and nouns

[Kaluarachchi et al., CIKM 2010]
Temporal Knowledge Bases

- **YAGO ontology**
  - Extract named entities from the YAGO ontology
  - Track named entity evolution using the New York Times Annotated Corpus

- **Wikipedia history**
  - Define a *time-based synonym* as a term semantically related to a named entity at a particular time period
  - Extract synonyms of named entities from *anchor texts* in article links using the whole history of Wikipedia

[Mazeika et al., CIKM 2011; Kanhabua et al., JCDL 2010]
Search with Name Changes

• Extract time-based synonyms from Wikipedia
  – Synonyms are words with similar meanings
  – In this context, synonyms refer name variants (name changes, titles, or roles) of a named entity
    • E.g., "Cardinal Joseph Ratzinger" is a synonym of "Pope Benedict XVI" before 2005

• Two types of time-based synonyms
  1. Time-independent
  2. Time-dependent

[Kanhabua et al., JCDL 2010]
Recognize Named Entities

Step 1: Partition Wikipedia regarding to the time granularity \( g = \text{month} \) to obtain its snapshots \( W = \{W_{t_1}, \ldots, W_{t_z}\} \)

**Figure**: A snapshot of Wikipedia and current revisions at time \( t_k \)

[Kanhabua et al., JCDL 2010]
Recognize Named Entities

Step 1: Partition Wikipedia regarding to the time granularity $g = \text{month}$ to obtain its snapshots $\mathbb{W} = \{W_{t_1}, \ldots, W_{t_z}\}$

Step 2: For each snapshot $W_{t_k} \in \mathbb{W}$, identify named entity pages to obtain a set of named entities $E_{t_k} = \{e_1, \ldots, e_j\}$

Example

[Bunescu and Pašca EACL’2006]

1) Multi-word titles and all words are capitalized
   - President_of_the_United_States $\Rightarrow$ named entity

2) Single-word titles with multiple capital letters
   - UNICEF and WHO are named entities

3) 75% of occurrences in the article text itself are capitalized
Recognize Named Entities

**Step 1:** Partition Wikipedia regarding to the time granularity \( g = \text{month} \) to obtain its snapshots \( \mathcal{W} = \{W_{t_1}, \ldots, W_{t_z}\} \)

**Step 2:** For each snapshot \( W_{t_k} \in \mathcal{W} \), identify named entity pages to obtain a set of named entities \( E_{t_k} = \{e_1, \ldots, e_j\} \)

---

 Risk (game)

From Wikipedia, the free encyclopedia

Risk is a strategic board game, produced by Parker Brothers (now a division of Hasbro). It was invented by French film director Albert Lamorisse and originally released in 1957, as La Conquête du Monde (“The Conquest of the World”), in France.

Risk is a turn-based game for two to six players. The standard version is played on a board depicting a stylised Napoleonic-era political map of the Earth, divided into forty-two territories, which are grouped into six continents. Players control armies with which they attempt to capture territories from other players. The primary object of the game is “world domination,” or “to occupy every territory on the board and in so doing, eliminate all other players.”[1] Using area movement, Risk ignores limitations such as the vast size of the world and the logistics of long campaigns.

In the 40th Anniversary Collector’s Edition the movement route between the territories of East Africa and Middle East was removed; this was later confirmed to be a manufacturing error, an error repeated in Risk II. Subsequent editions restored the missing route.[2] While the European versions of Risk had included the variation "Secret Mission Risk" for some time, the U.S. version did not have this added until 1993.[3]
Find Synonyms

Find a set of entity-synonym relationships at time $t_k$

For each $e_i \in E_{tk}$, extract anchor texts from article links:

- Entity: President_of_the_United_States
- Synonym: George W. Bush
- Time: 11/2004
Initial Results

- Time periods are not accurate

<table>
<thead>
<tr>
<th>Named Entity</th>
<th>Synonym</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pope Benedict XVI</td>
<td>Cardinal Joseph Ratzinger</td>
<td>05/2005 - 03/2009*</td>
</tr>
<tr>
<td></td>
<td>Joseph Ratzinger</td>
<td>05/2005 - 03/2009</td>
</tr>
<tr>
<td></td>
<td>Pope Benedict XVI</td>
<td>05/2005 - 03/2009</td>
</tr>
<tr>
<td>Barack Obama</td>
<td>Barack Hussein Obama II</td>
<td>02/2007 - 03/2009</td>
</tr>
<tr>
<td></td>
<td>Senator Barack Obama</td>
<td>05/2006 - 03/2009</td>
</tr>
<tr>
<td>Hillary Rodham Clinton</td>
<td>Hillary Clinton</td>
<td>08/2003 - 03/2009</td>
</tr>
</tbody>
</table>

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

[Kanhabua et al., JCDL 2010]
Enhancement using NYT

- Analyze **NYT Corpus** to discover accurate time
  - 20-year time span (1987-2007)
- Use the **burst detection** algorithm
  - Time periods of synonyms = burst intervals

[Kanhabua et al., JCDL 2010]
Enhancement using NYT

• Analyze **NYT Corpus** to discover accurate time
• Use the **burst detection** algorithm
  – Time periods of synonyms = burst intervals

Results from burst-detection algorithm

<table>
<thead>
<tr>
<th>Synonym</th>
<th>Entity</th>
<th>Burst Weight</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>President Reagan</td>
<td>Ronald Reagan</td>
<td>5506.858</td>
<td>01/1987 - 02/1989</td>
</tr>
<tr>
<td>President Ronald</td>
<td>Ronald Reagan</td>
<td>100.401</td>
<td>01/1989 - 03/1990</td>
</tr>
<tr>
<td>President Ronald</td>
<td>Ronald Reagan</td>
<td>67.208</td>
<td>07/1990 - 02/1993</td>
</tr>
<tr>
<td>Senator Clinton</td>
<td>Hillary Rodham Clinton</td>
<td>18.214</td>
<td>01/2001 - 10/2001</td>
</tr>
<tr>
<td>Senator Clinton</td>
<td>Hillary Rodham Clinton</td>
<td>17.732</td>
<td>05/2002 - 01/2003</td>
</tr>
</tbody>
</table>

Initial results
Query Expansion

1. A user enters an entity as a query

[QUEST: Query Expansion using Synonyms over Time]

Search Box

Enter query:

Pope Benedict XVI

Published from:
01 1987

To:
01 2007

[Search Archive]

[Kanhabua et al., ECML PKDD 2010]
Query Expansion

1. A user **enters an entity** as a query
2. The system **retrieves synonyms** wrt. the query

[Synonyms]

- Pope Benedict XVI
- Benedictus XVI
- Benedicto XVI
- Benedict XVI
- Joseph Alois Ratzinger
- Benedict xvi
- Benedictus 16

[Show time periods]
[Show scores]

[Kanhabua et al., ECML PKDD 2010]
Query Expansion

1. A user **enters an entity** as a query
2. The system **retrieves synonyms** wrt. the query
3. The user select synonyms to **expand the query**

**QUEST: Query Expansion using Synonyms over Time**

- **Search Box**
  - Enter query: *Pope Benedict XVI OR Cardinal Joseph Ratzinger*
  - e.g. *Pope Benedict XVI, Barack Obama, Hillary Rodham Clinton, or Marilyn Monroe*

- **Published from:** 01 1987
- **To:** 01 2007

- **Search Archive**

- Results: 1 - 10 of 97532 for *Pope Benedict XVI OR Cardinal Joseph Ratzinger* AND pubDate:[198701 TO 200701]

  **Apr 21, 2005 - Few See Taint in Service By Pope in Hitler Youth**
  Few See Taint in Service By Pope in Hitler Youth The day after Cardinal Joseph Ratzinger became Pope Benedict XVI  
  http://query.nytimes.com/gst/fullpage.html?res=9D00E7DE1731F932A15757C0A9639C8B63

  **Apr 21, 2005 - A Swift Surge That Defied Expectations**
  A Swift Surge That Defied Expectations Joseph Ratzinger of Germany became Pope Benedict XVI  

[Kanhabua et al., ECML PKDD 2010]
References

References (cont’)