Time-Based Exploration of News Archives

Omar Alonso, Klaus Berberich, Srikanta Bedathur, Gerhard Weikum

HCIR 2010
New Brunswick, NJ

August 22, 2010
Motivation

- News archives have **grown in number and size** due to improved digitization techniques

  ![Times Online](http://archive.timesonline.co.uk)
  All issues since **1785** digitized

  ![The New York Times](http://www.nytimes.com/archive)
  All issues since **1851** digitized

- Search on news archives often limited to displaying only a **ranked list of few relevant news articles**

- **Course of history** not easily visible but must be pieced-together by sifting through relevant news articles
Key Ideas

- **NEAT** *(News Exploration Along Time)* prototype

- **Timeline visualization** of relevant news content making use of different kinds of *temporal information*
  - Publication dates of news articles
  - Temporal expressions extracted from articles’ contents

- **Timeline annotation** using *semantic temporal anchors* (e.g., major events) gathered using *crowdsourcing*
Outline

- Motivation
- Search Result Visualization
- Timeline Annotation using Crowdsourcing
- Prototype Implementation
- Conclusion & Future Work
Search Result Visualization

NEAT

Search Box

Main Timeline

Overview Timeline

Timeline Annotations

Time-Based Exploration of News Archives (Klaus Berberich)
Publication-Time Dimension

- 150 most relevant news articles according to Okapi BM25
- Titles placed on timeline based on publication date
- Reflects course of real-world events but is limited to the time interval covered by the document collection (e.g., 1987–2007 for New York Times Annotated Corpus)

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>Jun</td>
<td>MetroStars Like Idea of Having Adu</td>
</tr>
<tr>
<td>2006</td>
<td>Aug</td>
<td>Your Tired and Your Poor But Not Your National Team</td>
</tr>
<tr>
<td>2006</td>
<td>Sep</td>
<td>Second New York Area Team</td>
</tr>
<tr>
<td>2006</td>
<td>Oct</td>
<td>Reason to Rejoice Again in Ukraine</td>
</tr>
<tr>
<td>2006</td>
<td>Nov</td>
<td>There's Still Time To Remember '05</td>
</tr>
<tr>
<td>2006</td>
<td>Dec</td>
<td>Postcards From 2005 Glory and a Ghost</td>
</tr>
<tr>
<td>2006</td>
<td>Feb</td>
<td>World Cup Tickets VIP</td>
</tr>
<tr>
<td>2006</td>
<td>Mar</td>
<td>No New Stadium No San Jose Team</td>
</tr>
<tr>
<td>2006</td>
<td>Apr</td>
<td>No Guarantees For U.S. Veterans</td>
</tr>
<tr>
<td>2006</td>
<td>May</td>
<td>Difficult Draw Doesn't Oweell Arena's Optimism</td>
</tr>
<tr>
<td>2006</td>
<td>Jun</td>
<td>In Berlin Every Cheer Casts An Eerie Echo</td>
</tr>
<tr>
<td>2006</td>
<td>Jul</td>
<td>Blatter's Blather Besmirching Soccer</td>
</tr>
<tr>
<td>2006</td>
<td>Aug</td>
<td>A Last Chance to Reach for World Cup Tickets</td>
</tr>
</tbody>
</table>

Time-Based Exploration of News Archives (Klaus Berberich)
Publication-Time Dimension

- 150 most relevant news articles according to Okapi BM25
- Titles placed on timeline based on publication date
- Reflects course of real-world events but is limited to the time interval covered by the document collection (e.g., 1987–2007 for New York Times Annotated Corpus)

<table>
<thead>
<tr>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time To Remember '05</td>
<td>The Hard Sell At the Cup</td>
<td>A Versatile Rookie Is the Biggest Bargain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In 2005 Glory and a Ghost</td>
<td>MOST BONITO</td>
<td>For Scorer Acceptance Is an Elusive Goal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>World Cup Tickets Via Where</td>
<td>Klimsmann Quits as Coach of Germany</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>San Jose Team</td>
<td>Zidane's Next Team May Be Red Bulls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U.S. Veterans</td>
<td>Top 10 Moments Times Bloggers Hoist the Cup</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quell Arena's Optimism</td>
<td>A Little Traveling Music Some Coaches Get Around</td>
<td>A U.S. Defender With an Attacking Spirit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Casts An Eerie Echo</td>
<td>Cup Empty For Many Members Of Military</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shrinking Soccer</td>
<td>Soccer Fan Makes Big Save For ESPN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>for World Cup Tickets</td>
<td>THEY'RE YOUNG AND READY TO SHINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reference-Time Dimension

- **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)

- **Temporal snippets** as sentences from news articles’ contents that contain at least one temporal expression

- **150 most relevant temporal snippets** according to Okapi BM25 * (# of temporal expressions)

- **Reflects course of real-world events and goes beyond the time interval covered by the document collection**

- **Users can glance at many relevant news articles at once**
Images of world leaders, celebrities and sports heroes are to be seen, along with subjects like the launching of the Queen Mary in 1934, a longshoremen's strike in 1951 and a group of screaming Beatles fans in 1964 (see source).

Wed, 01 Jan 1964 00:00:00 GMT
Fri, 01 Jan 1965 00:00:00 GMT

"The Beatles Anthology 1, " from 1958 to 1964, has sold nearly 10 million copies worldwide...

...the Beatles'appearance on "The Ed Sullivan Show "in 1964

...Today To the Editor: I still remember the day in 1964 when my college roommate and I... He was also the Beatles'publicity agent in 1964 and 1965 and press officer of the group's... 2002 ; the Bee Gees, in 1978 ; and the Beatles, in 1964

...anniversary of Beatles'appearance on Ed Sullivan Show in 1964 ; drawing ( S ) The Powerp...!

The Beatles in the U.S.A. "( 1964 ) ; "Meet Marlon Brando "( 1965 ) ; "A Visit...

...see the group's arrival here on its first visit in 1964 -- he caught both Beatles concerts at S... track lineup and original arrangements of Beatles'1964 American breakthrough album Meet...

..."A Hard Day's Night "and "Beatles for Sale, "from 1964, will be released worldwide as compa...
It's the Beatles, a December 1963 television appearance, which is expected...

...Beatles recorded 88 songs for the BBC between 1962 and 1965...

...Beatles items began turning up at auction as early as 1965, when...

...the American versions of four Beatles albums from 1965, "The Beatles..."

He was also the Beatles' publicity agent in 1964 and 1965 and press...

...Beatles in the U.S.A. "(1964); "Meet Marlon Brando "(1965); "A Visit...

...he caught both Beatles concerts at Shea Stadium, in 1965 and 1966.

1964 can now fetch $450 (without the record..."

A folded 1964 concert program valued at $55 was offered...

...portable once owned by Mick...

...3 and 1964, ranging in estimated price from $340 to...

The December 1962 appearance of the Beatles on "The Ed Sullivan Show" in 1964...

"I still remember the day in 1964 when my college roommate and I -- music..."

...was the Beatles' publicity agent in 1964 and 1965 and press officer of the group's...

...from 1958 to 1964, has sold nearly 10 million copies worldwide...

"On "The Ed Sullivan Show" in 1964...

...a scene on "The Ed Sullivan Show" in 1964...

...I remember the day in 1964 when my college roommate and I -- music..."

"Beatles' appearance on Ed Sullivan Show in 1964; drawing (S) The Powerpuff Girls, those...

"Meet Marlon Brando "(1965); "A Visit...

...Beatles' appearance on Ed Sullivan Show in 1964; drawing (S) The Powerpuff Girls, those...

...U.S.A. "(1964); "Meet Marlon Brando "(1965); "A Visit...

...val here on its first visit in 1964 -- he caught both Beatles concerts at Shea...

...original arrangements of Beatles' 1964 American breakthrough album Meet the Beatles...

..."and "Beatles for Sale," from 1964, will be released worldwide as compact disks..."
Outline

- Motivation
- Search Result Visualization
  - Timeline Annotation using Crowdsourcing
- Prototype Implementation
- Conclusion & Future Work
Timeline Annotation using Crowdsourcing

- Display major events as semantic temporal anchors that help users in contextualizing the relevant news articles and temporal snippets shown.
- Use crowdsourcing (e.g., Mechanical Turk) to let users determine major events that are meaningful to them.

Diagram:
- Thodore Roosevelt
- World War I
- Wall Street Crash
- Atomic Bomb
- Pearl Harbour
- World War II
- John F. Kennedy
- Vietnam War

Timeline:
- 1900
- 1910
- 1920
- 1930
- 1940
- 1950
- 1960
Timeline Annotation using Crowdsourcing

- Display major events as semantic temporal anchors that help users in contextualizing the relevant news articles and temporal snippets shown.
- Use crowdsourcing (e.g., Mechanical Turk) to let users determine major events that are meaningful to them.
Timeline Annotation using Crowdsourcing

- HITS run on Amazon Mechanical Turk in July/August ’09
  - topic-specific (e.g., music, politics, sports) & general
  - different temporal granularities (e.g., dates, years, decades)
  - we paid $0.01 per satisfactorily completed task

- Observations:
  - completion times vary significantly between topics
  - asking users to justify their input keeps work quality high

- From the users’ input we manually compile the list of major events shown based on the consensus seen
Timeline Annotation using Crowdsourcing

Complete a time-related query

We are interested in exploring search scenarios where temporal information is important to satisfy an information need. By temporal information we mean any time reference (e.g., “August 1999”, “20th century”, or “January 1 2002”).

Instructions

You’re given an incomplete query that contains only of a time reference. Please complete the query by adding the missing text that you think makes sense.

Examples

- Given 2001, you could add 9/11 attacks.
- Given 1982, you could add García Márquez Nobel prize or Falklands War.
- Given 1914-1918, you could add World-War I.

Task

Please complete the query by adding the missing text that you think makes sense. Please use your own common sense.

1930

Please provide any justification for your choice. We appreciate your comments!

Submit

major events shown based on the consensus seen
Timeline Annotation using Crowdsourcing

- HITS run on Amazon Mechanical Turk in July/August ’09
  - topic-specific (e.g., music, politics, sports) & general
  - different temporal granularities (e.g., dates, years, decades)
  - we paid $0.01 per satisfactorily completed task

- Observations:
  - completion times vary significantly between topics
  - asking users to justify their input keeps work quality high

- From the users’ input we manually compile the list of major events shown based on the consensus seen
Outline

- Motivation
- Search Result Visualization
- Timeline Annotation using Crowdsourcing
- Prototype Implementation
- Conclusion & Future Work
Always Grounded

...as a homage to George Harrison, the Beatle who died this year.

The True Culprit

...referring to American Ballet Theater's much-mocked 2002 ballet "Within You Without You: A Tribute..."

The True Culprit

This Time a Classical Fit

Personality And Posture Transform A

Cara Aidone John Hughes

Smack a Synergy Of Ballet and Beatle

...George Harrison, " which is to be presented

Yesterday

George Harrison's Gift

Ideas and Trends The Quiet End of the Quietest Beatle

George Harrison World Music Catalyst And Great Souled Man

A Childlike Simplicity Full of Love and

George Harrison World Music Catalyst And Great Souled Man Open to the Influence Of Unfamiliar Cult

QUOTATION OF THE DAY

George Harrison 'Quiet Beatle' And Lead Guitarist Dies At 50

While My Guitar Gently Waits

Just Something in the Way He Moved Them

With a Beatle's Death Mourning for Every Family

SOUNDFACE NAMES

...to music of former Beatle George Harrison, wh died last November, as highlight of its fall season at Olly...

Abbe Ruttenberg Michael Sarphie

Looking in the Sprawl of That Early 70's Sound

He treated in 2001, and coerced him into signing autographs...

Adirondack's Double

Richard M. Nixon

Ronald Reagan

Bill Clinton

George W. Bush

Y2K

Barack Obama

September 11th, 2001

George W. Bush

Lewinsky Scandal

German Reunification

Operation Desert Storm

Fall of Berlin Wall

George H. W. Bush

John F. Kennedy

Vietnam War

Prototype Implementation

- **Dataset**: New York Times Annotated Corpus (1.8M documents published during 1987–2007)
- **Temporal expressions** extracted using TARSQI (and some additional hand-crafted regular expressions)
- **Timeline visualization** from M.I.T.’s SIMILE project
- **Java servlet** processes queries and prepares results (i.e., temporal snippets and titles) for visualization
- **Oracle 11g** relational database keeps indexes and data (e.g., temporal snippets and headlines)
Outline

- Motivation
- Search Result Visualization
- Timeline Annotation using Crowdsourcing
- Prototype Implementation
- Conclusion & Future Work
Conclusion & Future Work

- **NEAT (News Exploration Along Time)** prototype
  - timeline visualization of relevant news content
  - publication dates reveal when relevant content was published
  - temporal snippets show times that relevant content talks about
  - crowdsourced timeline annotations provide historical context

- Future work and open issues
  - systematic empirical evaluation through a user study
  - user-dependent/personalized timeline annotations
  - alternative methods to select news articles and temporal snippets to display (e.g., considering available screen space)
Thanks!

Questions?