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

http://archive.timesonline.co.uk
All issues since 1785 digitized

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** (**N**ews **E**xploration **A**long **T**ime) prototype

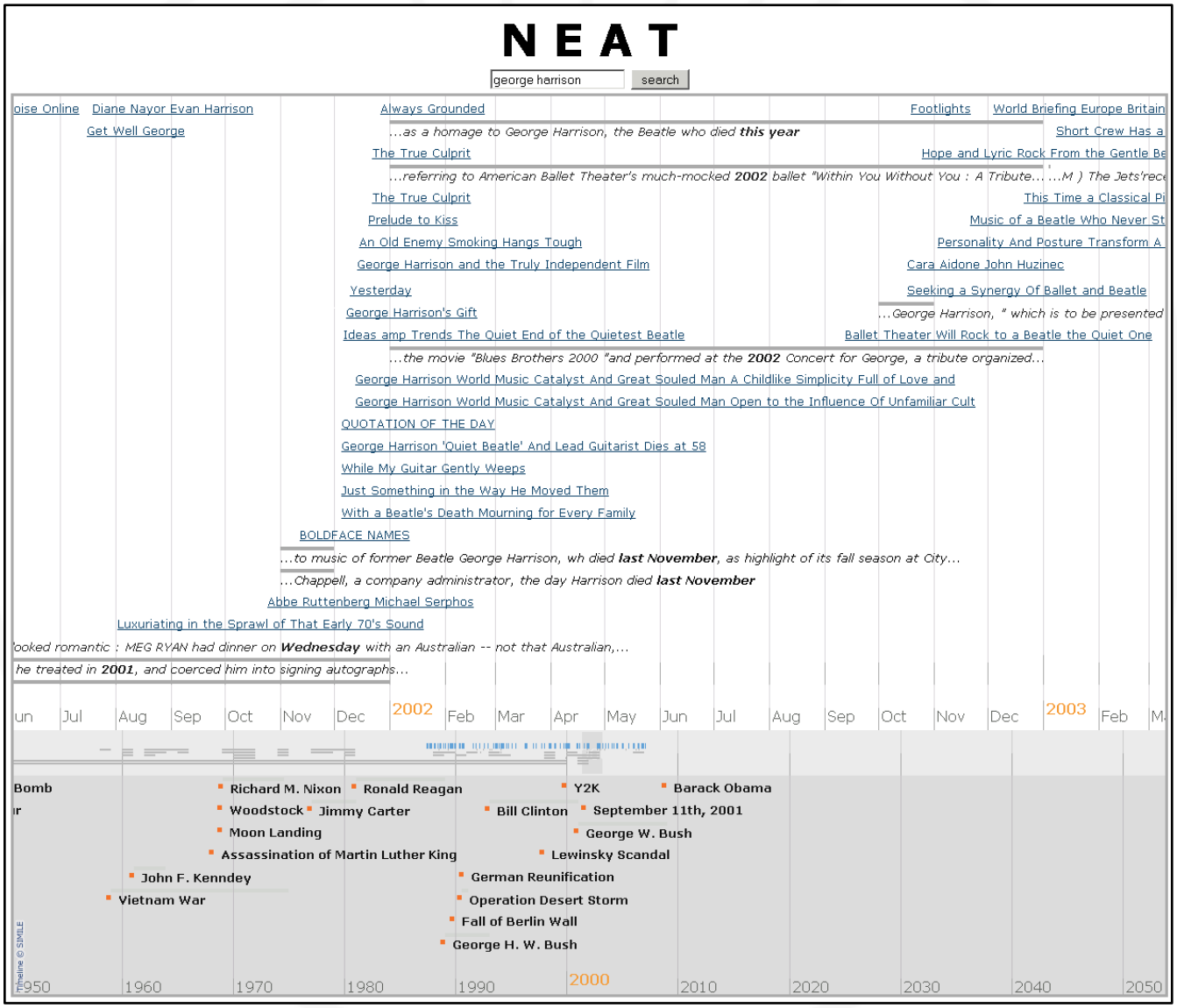
- Timeline visualization of relevant news content making use of different kinds of temporal information
  - Publication dates of news articles
  - Temporal expressions extracted from articles' contents

- 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

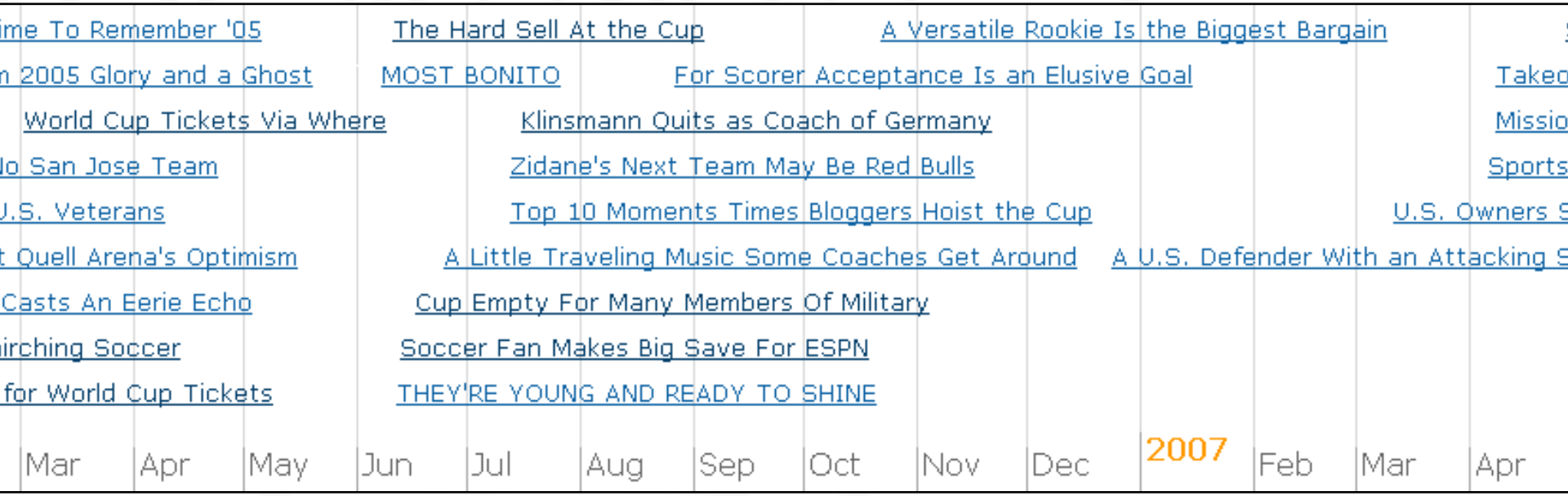


Publication-Time Dimension

- 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)

Publication-Time Dimension

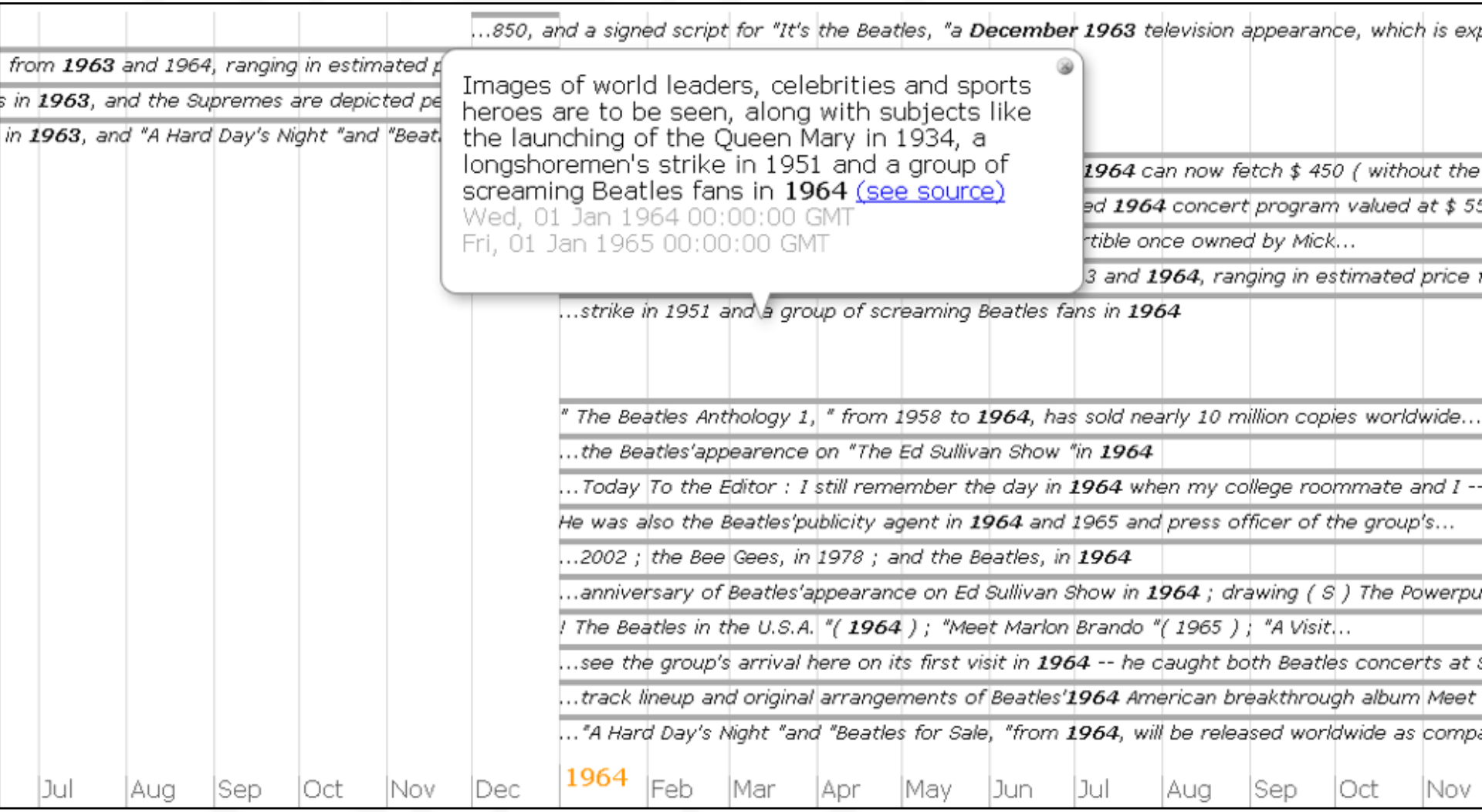
- 150 most relevant news articles according to Okapi BM25
- Titles placed on timeline based on publication date
- Reflects course of real-world events but is limited to the time interval covered by the document collection (e.g., 1987–2007 for New York Times Annotated Corpus)



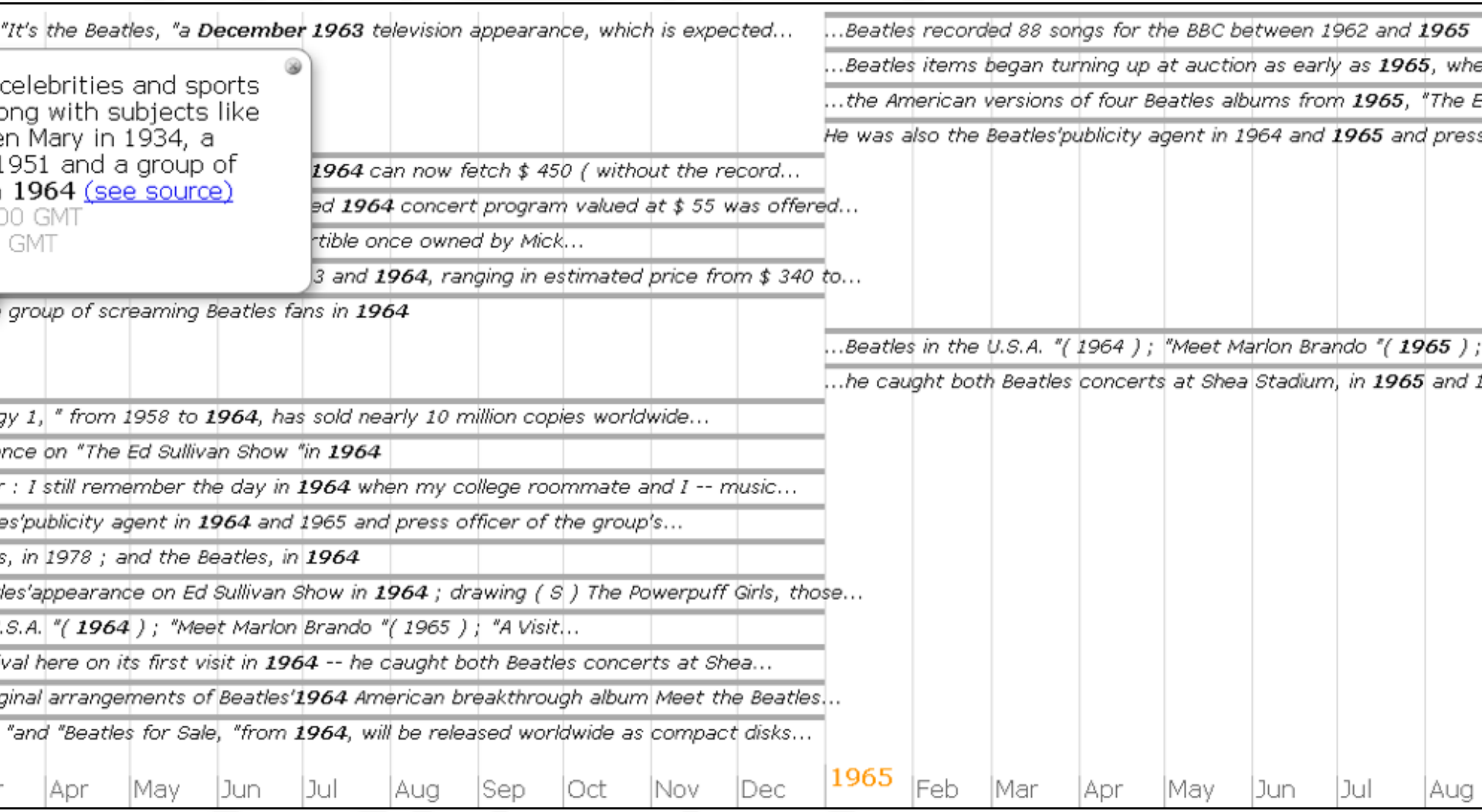
Reference-Time Dimension

- 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

Reference-Time Dimension



Reference-Time Dimension

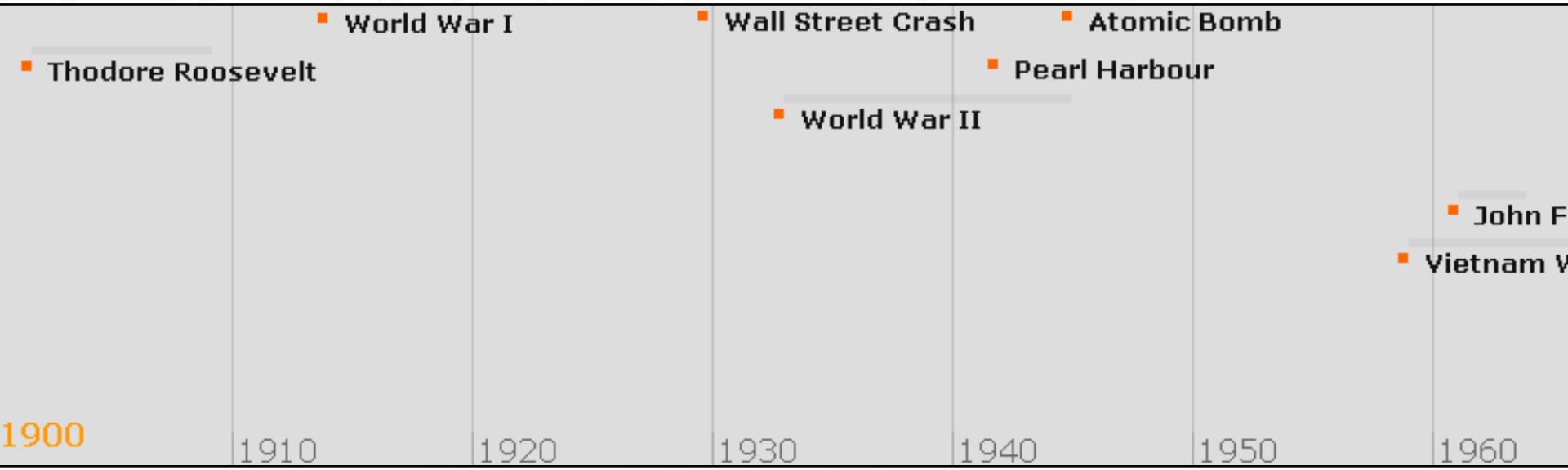


Outline

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

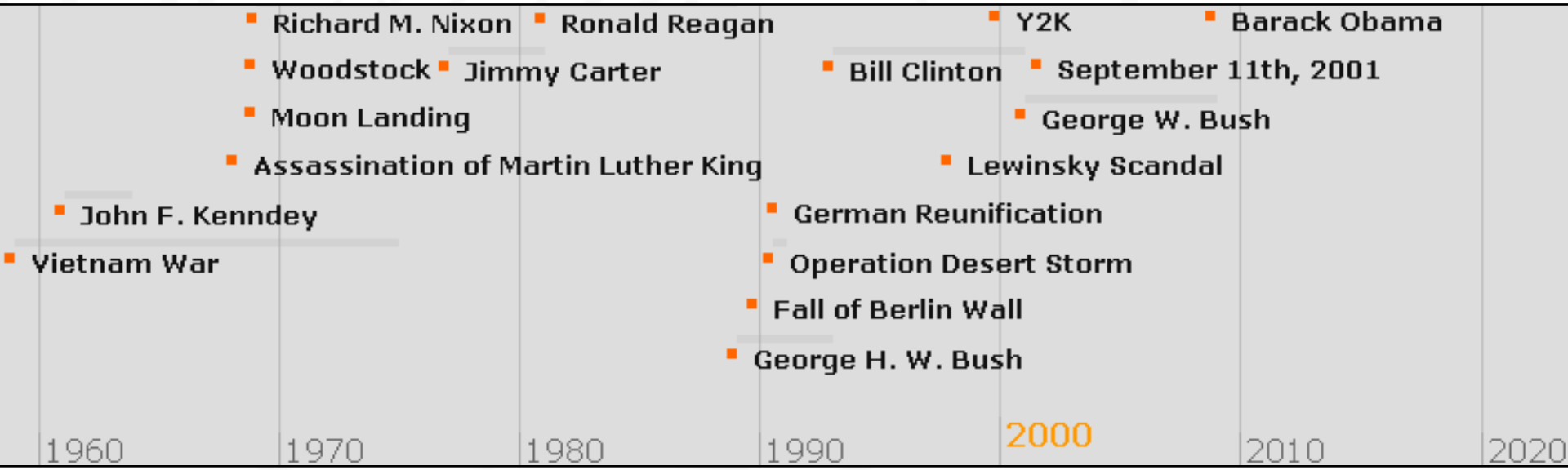
Timeline Annotation using Crowdsourcing

- 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



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 comon 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

N E A T

george harrison search

oise Online Diane Nayor Evan Harrison
Get Well George
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... ...M) The Jets'rec
The True Culprit
Prelude to Kiss
An Old Enemy Smoking Hangs Tough
George Harrison and the Truly Independent Film
Yesterday
George Harrison's Gift
Ideas amp Trends The Quiet End of the Quietest Beatle
...the movie "Blues Brothers 2000 "and performed at the **2002** Concert for George, a tribute organized...
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 58
While My Guitar Gently Weeps
Just Something in the Way He Moved Them
With a Beatle's Death Mourning for Every Family
BOLDFACE NAMES
...to music of former Beatle George Harrison, wh died **last November**, as highlight of its fall season at City...
...Chappell, a company administrator, the day Harrison died **last November**
Abbe Ruttenberg Michael Serphos
Luxuriating in the Sprawl of That Early 70's Sound
Footlights World Briefing Europe Britain
Short Crew Has a
Hope and Lyric Rock From the Gentle Be
This Time a Classical Pi
Music of a Beatle Who Never St
Personality And Posture Transform A
Cara Aidone John Huzinec
Seeking a Synergy Of Ballet and Beatle
...George Harrison, " which is to be presented
Ballet Theater Will Rock to a Beatle the Quiet One
ooked romantic : MEG RYAN had dinner on **Wednesday** with an Australian -- not that Australian,...
he treated in **2001**, and coerced him into signing autographs...

un Jul Aug Sep Oct Nov Dec 2002 Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec 2003 Feb M

Bomb
r
Richard M. Nixon
Ronald Reagan
Woodstock
Jimmy Carter
Moon Landing
Assassination of Martin Luther King
John F. Kenndey
Vietnam War
Y2K
Barack Obama
Bill Clinton
September 11th, 2001
George W. Bush
Lewinsky Scandal
German Reunification
Operation Desert Storm
Fall of Berlin Wall
George H. W. Bush

1950 1960 1970 1980 1990 2000 2010 2020 2030 2040 2050

Timeline © SIMILE

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** (**N**ews **E**xploration **A**long **T**ime) 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?