

TempWeb 2013

3rd Temporal Web Analytics Workshop



Stuff happens continuously: exploring Web contents with temporal information

Omar Alonso

Microsoft

13 May 2013



Disclaimer

The views, opinions, positions, or strategies expressed in this talk are mine and do not necessarily reflect the official policy or position of Microsoft.

Agenda

- Part I: Time in document collections
- Part II: Social data
- Part III: Exploring the Web using time

Introduction

- Lots of interest in the research community
 - WSDM 2013 tutorial
 - SIGIR TAIA Workshop (2012/2013)
 - WWW Temporal Web (2011/2012/2013)
 - TREC
- Temporal information is more visible now
 - Timeline (Facebook)
 - Trending topics (Twitter)
- Big data

What is time?

- Several ways to express time in a language
- In linguistics: tense and grammatical aspects
- Temporal *reference expressions*
 - On Friday
 - March 31st, 2007
 - In the late 90s
- Several ways to express the *same* time information
 - Christmas Day
 - December 25
 - Navidad
 - Weihnachten

Characteristics of Temporal Information

- Temporal Information is **well-defined**

Characteristics of Temporal Information

- Temporal Information is **well-defined**

- two points/intervals

→ temporal relationship

before



after



overlap



...

Characteristics of Temporal Information

- Temporal Information can be **normalized** to some standard format

Characteristics of Temporal Information

- Temporal Information can be **normalized** to some standard format

- expressions with same semantics

- same value in standard format

March 28, 2011 → 2011-03-28

2011/03/28 → 2011-03-28

today → 2011-03-28

Characteristics of Temporal Information

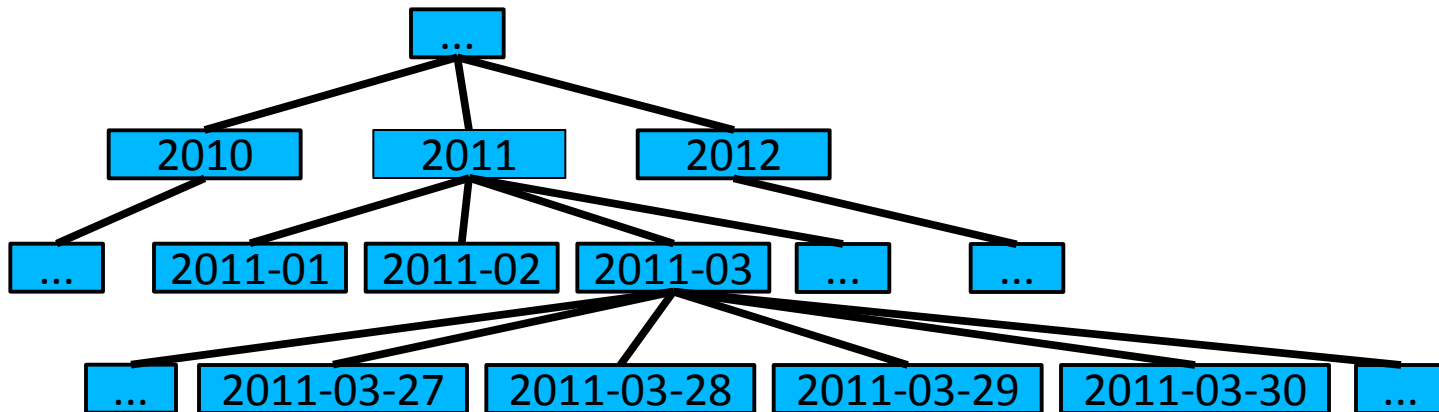
- Temporal Information can be **organized hierarchically**

Characteristics of Temporal Information

- Temporal Information can be **organized hierarchically**

- different granularities (day, month, ...)

→ tree structure



Types of Temporal Information

- 4 types of temporal expressions

- **Dates**

- Monday 13, 2013

- **Times**

- 11 a.m.

- **Durations**

- two weeks

- **Sets**

- twice a week

Types of Temporal Information

- Explicit expressions
- Implicit expressions
- Relative expressions

Types of Temporal Information

□ **Explicit expressions**

- refer to specific point in time
- carry all information explicitly

January 25, 2010

November 2012

the 7th century

- can be normalized without further knowledge

Types of Temporal Information

- **Implicit expressions**

- names of holidays

- names of events

Independency Day 2001

Labor Day 1999

- can be normalized if semantics is known:

Labor Day → first Monday in September

Types of Temporal Information

□ **Relative expressions**

- may be underspecified or deictic

today

January 12

two weeks later

- context information is needed

→ reference time

→ relation to reference time

Types of Temporal Information

□ Example:

Document Creation Time: 1998-04-18
Hungarian astronaut Bertalan Farkas is leaving for the United States to start a new career, he said today.
... On May 22, 1995, Farkas was made a brigadier general, and the following year he was appointed military attache
... However, cited by District of Columbia traffic police in December for driving under the influence of ...

- explicit and relative expressions
with needed context information

Extracting temporal expressions

- Temporal tagger
- Similar to NER but for time only
- TimeML format

Sunday, 09 July 2006 Italy won the World Cup
tonight after beating France on penalties.

<s>

<TIMEX3 tid="t1" TYPE="DATE" VAL="20060709">

Sunday, 09 July 2006 </TIMEX3>

Italy won the World Cup

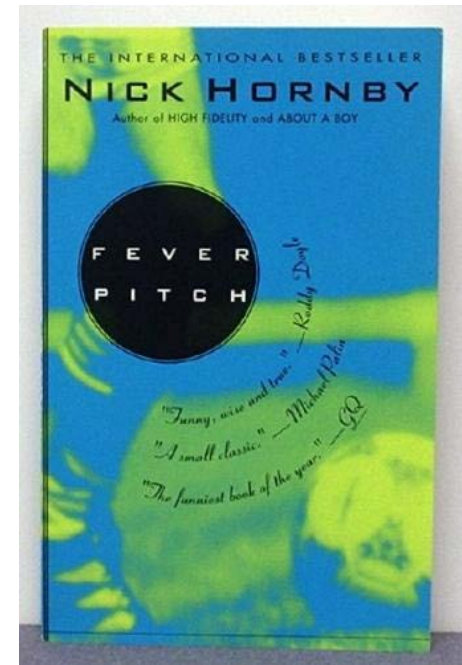
<TIMEX3 tid="t2" TYPE="DATE" VAL="20060709TNI">

tonight</TIMEX3> after beating France on penalties.

</s>

Why we care?

- Temporal landmarks
- Temporal evolution



Wisdom of crowds

- Calendar annotation for free
 - Take all hot topics for a year
 - Annotate events that people care
 - Use that to learn what maybe relevant for that date the next year
- Repeating events
 - Elections, sport championships, etc.
- Unexpected events that will be remembered
 - 9/11, Japan earthquake, Chilean miners, etc.

Document lifecycle

- Document creation time (ts)
- Availability of temporal expressions
- Updates as maintenance
- Topical relevance
- Emphasis on freshness:
 - the extent to which information is recent, current, up-to-date

SOCIAL DATA

Beyond Web pages

- Notion of document changes
- Social data
- Facebook, Twitter, Quora, Foursquare, etc.
- Wikipedia

Twitter

- 140 characters
- Entry barrier is pretty low
- Vitality
- Events
- Sports



Time in Twitter

- Trending topics
- Hashtags
- Cashtags
- Group chats

Trends · Change

Xperia ZR

#TwoThingsThatDontMixWell

#survivorfinale

Happy Mother's Day

#sup5sos

#GOP2016Slogans

IRS

#tco

Tia Sharp

New York



Business Insider @businessinsider

7 May

Microsoft VP Says Smaller Windows 8 Tablets Are Coming At The End Of June **\$MSFT** by @stevekovach read.bi/18s43cL

View summary



Hilgos Foundation @IRememberBetter

4h

#AlzChat 3pmEST/8pmUK Today's topic: Tips for decor & furnishings for a person with dementia @creativitycare

@LivHOMECorp @Alzheimers_NIH

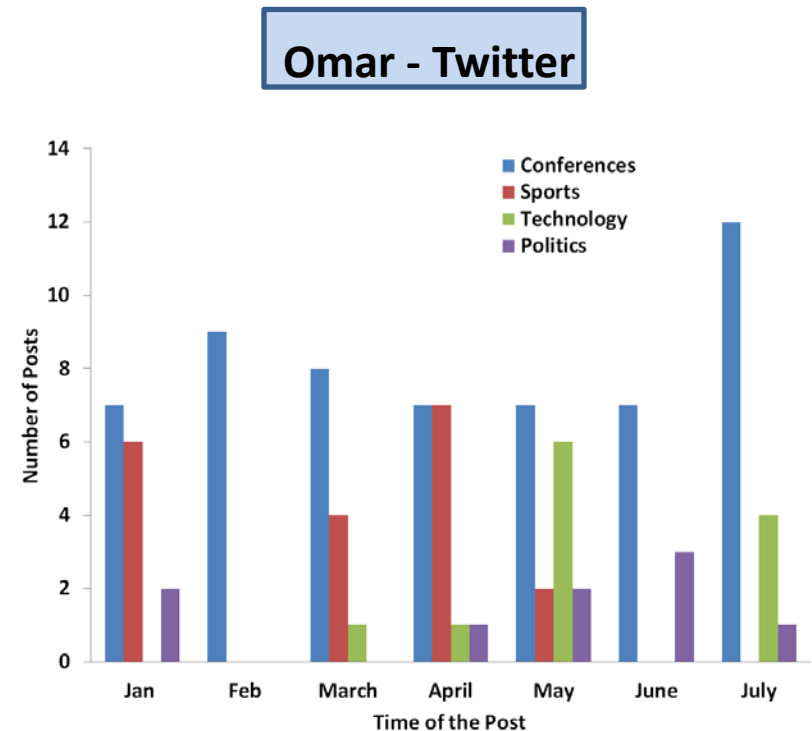
Expand

Facebook

- Your timeline
- Photos over time
- Generic events
- Personal social network events
 - Happy Birthday!
 - Weddings
 - New Baby pictures

Time and social characteristics

- User interests can be time sensitive
- Temporally aware signals (TAS)
 - Recent
 - On-going
 - Seasonal
 - Past
 - Random
- Three levels
 - Personal
 - Friends
 - Universal



Questions and Answers

- Human computation for queries that search engines don't do well
- Latest incarnations: Quora & StackOverflow



Facebook Questions

Learn from people in the know and the friends who know you best.

[Ask Question](#)

YAHOO! ANSWERS



Quora

A continually improving collection of questions and answers created, edited, and organized by everyone who uses it.

Time in CQA

- Some answers don't change
- Sustainable questions

Boston Marathon Terrorist Attacks (April 2013): [Boston](#) [Boston Marathon](#)
[Breaking News](#) [Edit](#)

What happened at the 2013 Boston Marathon? [Edit](#)

I saw something about explosions at the finish line. Looking for more detailed information. [Edit](#)

[1 Comment](#) · [Share \(7\)](#) · [Options](#)

5 Answers

[Ask to Answer](#)

U.S. Presidential Elections: [U.S. Constitution](#) [U.S. Federal Government](#)
[U.S. Political Parties](#) [U.S. Politics](#) [Edit](#)

Should the Electoral College be abolished? Can it be abolished? [Edit](#)

And will any presidential candidate spend much time in Ohio, again, if it is? See also: [Electoral College: Given that the Electoral College appears to be very unpopular, why has there not been more momentum to abolish or change it?](#) [Edit](#)

[2+ Comments](#) · [Share \(1\)](#) · [Options](#)

16 Answers

[Ask to Answer](#)

1
vote

2
answers

Q: Why does the calculation of Cohen's kappa fail across different packages on this contingency...

(conf.mat) The output I get from running this is: `> cohen.kappa(conf.mat)` **Kappa** test for nominally classified data 4 categories - 2 methods **kappa (Cohen)** = 0 , Z = NaN , p = NaN **kappa** (Siegel ... I have a contingency table for which I would like to calculate Cohens's **kappa** - the level of agreement. I have tried using three different packages, which all seem to fail to some degree. The package ...

[r](#) [package](#)

asked aug 9 by [gisol](#)

reddit

- Extremely popular
- Threaded structure
- Community-driven
- AMAs



I am Barack Obama, President of the United States -- AMA (self.IAmA)

submitted 8 months ago* by PresidentObama

Hi, I'm Barack Obama, President of the United States. Ask me anything. I'll be taking your questions for half an hour starting at about 4:30 ET.

Proof it's me: <https://twitter.com/BarackObama/status/240903767350968320>

We're running early and will get started soon.

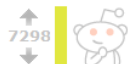
UPDATE: Hey everybody - this is Barack. Just finished a great rally in Charlottesville, and am looking forward to your questions. At the top, I do want to say that our thoughts and prayers are with folks who are dealing with Hurricane Isaac in the Gulf, and to let them know that we are going to be coordinating with state and local officials to make sure that we give families everything they need to recover.

Verification photo: <http://i.imgur.com/oz0a7.jpg>

LAST UPDATE: I need to get going so I'm back in DC in time for dinner. But I want to thank everybody at reddit for participating - this is an example of how technology and the internet can empower the sorts of conversations that strengthen our democracy over the long run. AND REMEMBER TO VOTE IN NOVEMBER - if you need to know how to register, go to <http://gottaregister.com>. By the way, if you want to know what I think about this whole reddit experience - NOT BAD!

<http://www.barackobama.com/reddit> [edit: link fixed by staff]

24230 comments share



I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation. AMA (self.IAmA)

submitted 2 months ago* by thisisbillgates

Hi, I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation. Ask me anything.

Many of you know me from my Microsoft days. The company remains very important to me and I'm still chairman. But today my full time work is with the foundation. Melinda and I believe that everyone deserves the chance for a healthy and productive life - and so with the help of our amazing partners, we are working to find innovative ways to help people in need all over the world.

I've just finished writing my 2013 Annual Letter <http://www.billsletter.com>. This year I wrote about how there is a great opportunity to apply goals and measures to make global improvements in health, development and even education in the U.S.

VERIFICATION: <http://i.imgur.com/vlMjEgF.jpg>

I'll be answering your questions live, starting at 10:45 am PST. I'm looking forward to my first AMA.

UPDATE: Here's a video where I've answered a few popular Reddit questions - http://youtu.be/qv_F-oKvIKU

UPDATE: Thanks for the great AMA, Reddit! I hope you'll read my annual letter www.billsletter.com and visit my website, The Gates Notes, www.gatesnotes.com to see what I'm working on. I'd just like to leave you with the thought that helping others can be very gratifying. <http://i.imgur.com/D3qRaty.jpg>

28250 comments share

reddit

Time in reviews

- Amazon
- Yelp
- Foursquare

★★★★☆ **Kindle Me!** February 25, 2009

By [Robin L. McLaughlin](#)

[Amazon Verified Purchase](#)

I'm a new Kindle 2 owner and I did not own a Kindle 1. I was very interested in the original Kindle, b Since it looked like they made the improvements I was waiting for (one of the others was a bit more

★★★★☆ **New Kindle Buyer? Thinking about Nook? Read on ...** September 12, 2012

By [L. Kirk](#) [TOP 500 REVIEWER](#)

[Amazon Verified Purchase](#)

I got my first Kindle. This is a nice ebook reader. I'm going to jump right in and tell you what I found

Buy your timed ticket over the Internet otherwise you will have to wait in long lines waiting for a van

Denice M. · January 4

[Save](#) [Like](#)

Vale a pena conhecer!

Cleiton L. · December 20, 2012

[Save](#) [Like](#)



Be on top of the world!

Carlo M. · December 19, 2012

[Save](#) [Like](#)

Vista mais bonita do Rio de Janeiro. Mas para ir no final de semana ensolarado tem que ter paciência para enfrentar a fila.

Luisa M. · November 17, 2012

[Save](#) [Like](#)

★★★★★ 3/7/2010

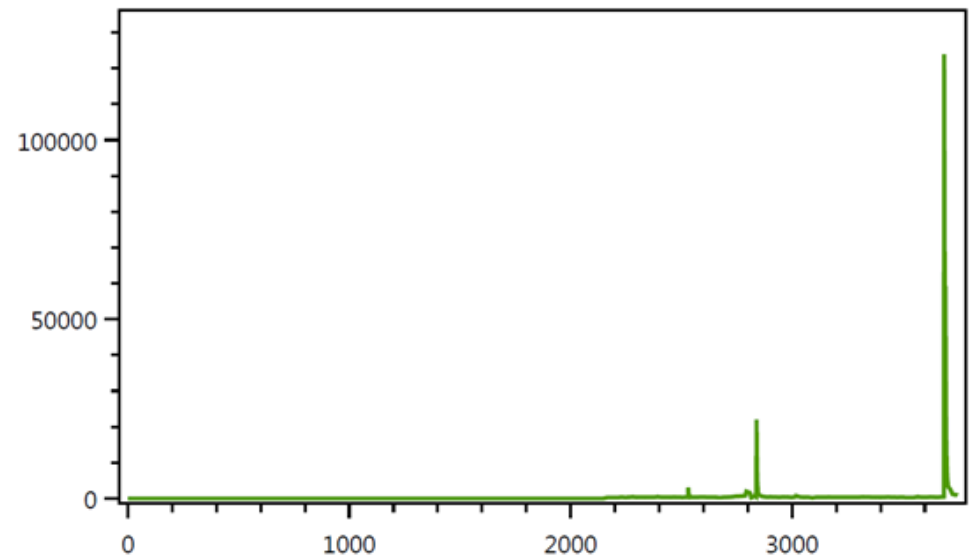
If you walked by here, & looked in, you probably wouldn't think much. However, once through the doors, the dining experience is always superb.

★★★★☆ 5/10/2013

This is my favorite place to have dinner in the city. The food is amazing and the view is pretty sweet. The service could have been a bit better but Ive been back and back so Im not complaining

Time in Wikipedia

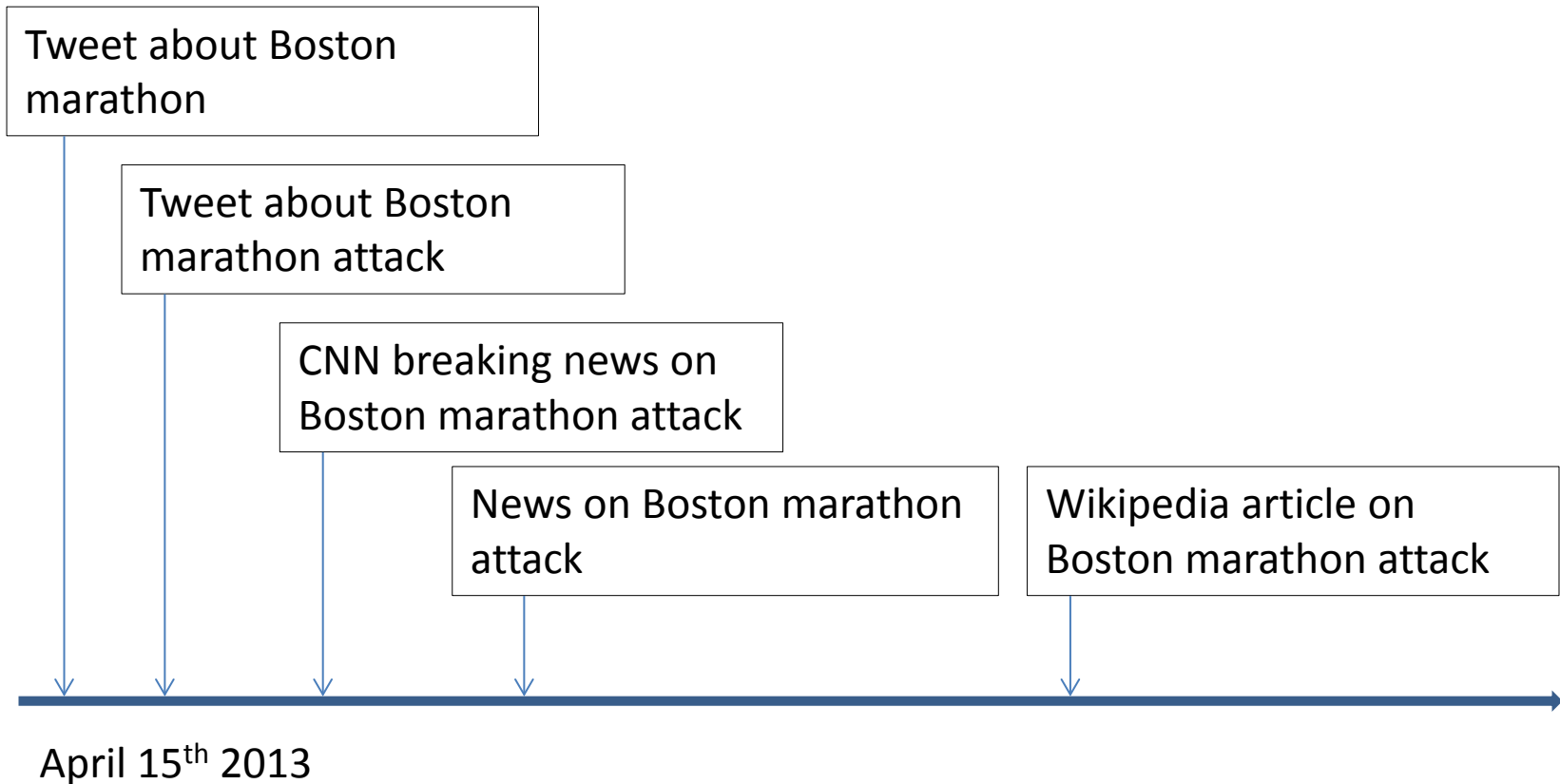
- Crowdsourcing
- Edits
- Whitney Houston
 - TV appearances
 - Her death in February 2012
 - Viewing popularity since article creation on January 8th 2002.



A common pattern

- Tweet -> breaking news -> news article -> Wikipedia page

The current “document” view



EXPLORING DATA WITH TEMPORAL INFORMATION

Temporal value

- Documents/web pages
 - Timestamp + temporal expressions
- Queries
 - Trends & popularity
- Temporal value per source
 - Can we associate a temporal utility value at a source?

Correlating sources

- Identify event-based hashtags, for timeline creation
- Currently using historic/past data
- Find most temporally significant hashtags
- Use multiple signals: Twitter search logs + related Wikipedia article popularity
- Automatic timeline-generation

Hashtags + Wikipedia

Hashtag	Correlation	Wikipedia Article	From Date	To Date
#comeonirene	0.87	Irene	25/08/2011	30/08/2011
#iastrawpoll	0.87	Ames Straw Poll	12/08/2011	15/08/2011
#roke	0.86	Typhoon Roke	20/09/2011	21/09/2011
#chardonshooting	0.85	Chardon High School shooting	27/02/2012	29/02/2012
#westmemphis3	0.82	Free the West Memphis 3	18/08/2011	21/08/2011
#iowacaucus	0.82	Iowa caucuses, 2008	29/12/2011	05/01/2012
#oslobomb	0.73	2011 Norway attacks	22/07/2011	27/07/2011
#lawrencebrewer	0.77	Murder of James Byrd, Jr.	21/09/2011	22/09/2011
#royaltour	0.74	2011 royal tour of Canada	01/07/2011	15/07/2011
#jamesbyrd	0.73	Murder of James Byrd, Jr.	22/09/2011	23/09/2011

Time	Hashtag	Wikipedia Article
07/01/2011	#royaltour	2011 royal tour of Canada
07/06/2011	#cuci	Cuci
07/08/2011	#bettyford	Betty Ford
07/07/2011	#juliogomez	Julio Gomez
08/01/2011	#tropicalstormemily	Tropical Storm Emily (2011)
08/04/2011	#repulsing	Army of the Coasts of Brest
08/12/2011	#iastrawpoll	Ames Straw Poll
8/18/2011	#westmemphis3	Free the West Memphis 3
8/25/2011	#comeonirene	Irene
9/20/2011	#roke	Typhoon Roke
9/21/2011	#lawrencebrewer	Murder of James Byrd, Jr.
9/23/2011	#kimdelaney	Kim Delaney
9/27/2011	#plzbelieve	Singaluru
10/04/2011	#californiamoney	Addis Ababa Fistula Hospital
10/10/2011	#jova	Hurricane Jova
10/23/2011	#simoncelli	Marco Simoncelli
11/26/2011	#goetitanddontlookback	The Legion of Missing Men
11/16/2011	#onlyatecu	ECU'92
11/27/2011	#garyspeed	Gary Speed
01/01/2012	#christmasfeast	Devils on horseback

Combining sources

- CrowdTiles
- Web + Twitter + Wikipedia
- Get a sense of what is happening now

The screenshot shows a Bing search results page for the query "euro 2012". At the top, there are navigation links for WEB, IMAGES, VIDEOS, MAPS, NEWS, and MORE. The search bar contains "euro 2012" and shows 742,000,000 results. An advertisement for "Euro 2012" from www.soccer.com is visible. Below the ad is a "what+when+who+now:" section. On the left, a red box labeled (1) contains a Wikipedia snippet about the UEFA Euro 2012 tournament, noting it changed 20 times in the last hour. On the right, a blue box labeled (2) promotes Sky Sports Euro 2012 news. Below this is a "News about euro 2012" section with a "bing.com/news" link and several news headlines, including "Euro 2012: A Day Off ... and More Views From the Goal Line" from New York Times Blogs and "Croatia faces fresh Euro 2012 racism probe" from CNN. At the bottom, there is a link to the Wikipedia page for "UEFA Euro 2012" and a brief introductory text.

WEB IMAGES VIDEOS MAPS NEWS MORE

bing euro 2012

742,000,000 RESULTS Any time ▾

Euro 2012 Ad
www.soccer.com/Euro · Shop Official Euro 2012 Gear. Free Shipping on Orders \$100 & Up!

what+when+who+now:

(1) **UEFA Euro 2012**
changed 20 times in the last hour
Neither Poland nor Ukraine, the two host countries, made it through the group stage. Thus Euro 2012 becomes the second European Championship so far (after UEFA Euro 2008 held in Austria and Switzerland) to have none of the hosts emerge from the first round, as well as the third big football tournament in a row to do so (after UEFA Euro 2008 and 2010 FIFA World Cup held in South Africa).
on wikipedia

Sky Sports Euro 2012
Get the latest Euro 2012 news with Sky Sports...
on twitter

News about euro 2012 (2)
bing.com/news

Euro 2012: A Day Off ... and More Views From the Goal Line
You really know you have a problem when the prime minister of Ukraine chimes in in favor of more transparency ... on the soccer field. Less than 24 hours after... [New York Times Blogs](#) · 1 hour ago

Croatia faces fresh Euro 2012 racism probe [CNN](#)

Euro 2012 TV Audience Up In North America And Why It Will Only Get Better [Forbes](#)

See also: [More stories](#) · [Top stories](#) · [Related blogs](#)

UEFA Euro 2012 - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/UEFA_Euro_2012 ▾
[Host selection](#) · [Venues](#) · [Qualification](#) · [Final draw](#) · [Match officials](#) · [Squads](#)
The 2012 UEFA European Football Championship, commonly referred to as **Euro 2012**, is the 14th European Championship for national football teams organised by UEFA. ...

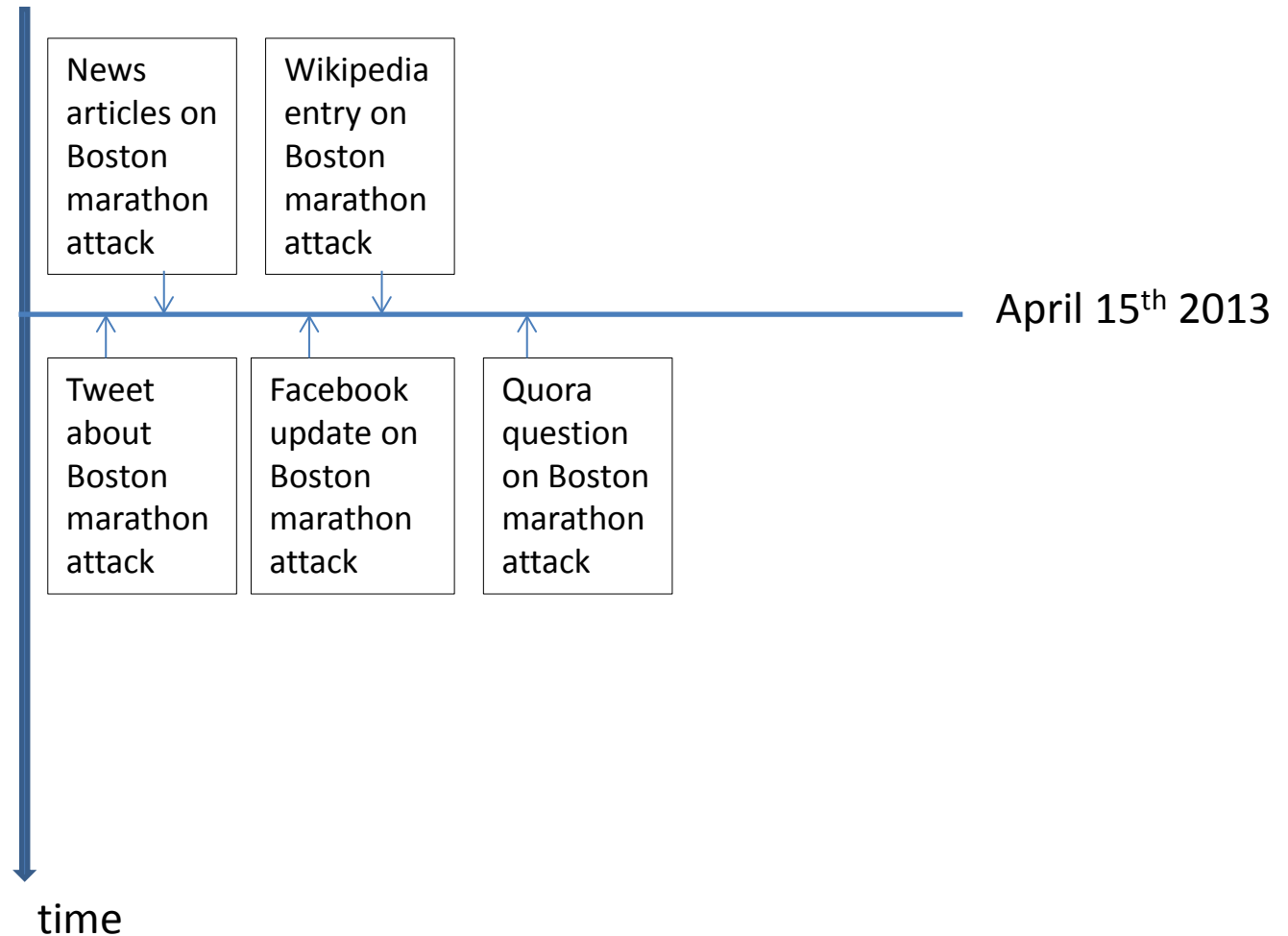
RELATED SEARCHES
[Euro 2012 Draw](#)
[Euro 2012 Groups](#)
[Euro 2012 Stadiums](#)
[Euro 2012 W Polsce](#)
[Euro 2012 Qualifying](#)
[Euro to Dollar](#)
[Euro 2008 Qualifying Matc](#)
[Preliminary Euro 2012](#)

Observation

- Different crowds on different sources
- Two crowds of people
 - Authors/searchers on Twitter
 - Editors/browsers on Wikipedia
- Correlation between signals from the two crowds
 - People search for what is happening
 - People edit Wikipedia with what is happening
- Combining and integrating different sources by time

The new document

- New virtual document
- Time is cross-cutting
- Sources share ts
- Combine all relevant information



Archiving and looking back

- News articles & Wikipedia are fine for main events
- What about stuff that is not so trending?
- What did people say at the Oscars 2011?
- Mario Balotelli at Euro 2012



The #Oscars' woman problem: 98% of nominations went to movies directed by men <http://t.co/TFlw1B94>



Here is the bus in Rome with it's number/destination changed to simply read: SUPER MARIO. #Balotelli #Euro2012 http ...

Problems

- Duplicate and near-duplicate detection across sources
- Temporal summaries
- Temporal precedence by source
- Archiving
- Ranking by temporal value
- Temporal aware search engines
- Synchronization and latency
- Can time help or escape the filter bubble?
- Evaluation

Conclusions

- Temporal information for organizing content is becoming more important
- Exciting area for combining different sources
- Lots of opportunities to improve current search experience
- Time is a cross-cutting aspect

Thank you