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 31\textsuperscript{st}, 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 \textit{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 $\rightarrow$ 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:

  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
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
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
Time in CQA

- Some answers don’t change
- Sustainable questions
reddit

- Extremely popular
- Threaded structure
- Community-driven
- AMAs
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
Donece M. - January 4
Save  Love

Yale a pena conhecer!
Cleiton L. - December 20, 2012
Save  Like

Be on top of the world!
Carlo M. - December 19, 2012
Save  Love

Vista mais bonita do Rio de Janeiro. Mas para ir no final de semana ensolarado tem que ter paciência para enfrentar a fila
Lucía M. - November 17, 2012
Save  Like

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

3/7/2010

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 I've been back and back so I'm not complaining

5/10/2013
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

- Tweet about Boston marathon
- Tweet about Boston marathon attack
- CNN breaking news on Boston marathon attack
- News on Boston marathon attack
- Wikipedia article on Boston marathon attack

April 15th 2013
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

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

<table>
<thead>
<tr>
<th>Time</th>
<th>Hashtag</th>
<th>Wikipedia Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>07/01/2011</td>
<td>#royaltour</td>
<td>2011 royal tour of Canada</td>
</tr>
<tr>
<td>07/06/2011</td>
<td>#cuci</td>
<td>Cuci</td>
</tr>
<tr>
<td>07/08/2011</td>
<td>#bettyford</td>
<td>Betty Ford</td>
</tr>
<tr>
<td>07/07/2011</td>
<td>#juliogomez</td>
<td>Julio Gomez</td>
</tr>
<tr>
<td>08/01/2011</td>
<td>#tropicalstormemily</td>
<td>Tropical Storm Emily (2011)</td>
</tr>
<tr>
<td>08/04/2011</td>
<td>#repulsing</td>
<td>Army of the Coasts of Brest</td>
</tr>
<tr>
<td>08/12/2011</td>
<td>#iastrawpoll</td>
<td>Ames Straw Poll</td>
</tr>
<tr>
<td>08/18/2011</td>
<td>#westmemphis3</td>
<td>Free the West Memphis 3</td>
</tr>
<tr>
<td>08/25/2011</td>
<td>#comeonirene</td>
<td>Irene</td>
</tr>
<tr>
<td>09/20/2011</td>
<td>#roke</td>
<td>Typhoon Roke</td>
</tr>
<tr>
<td>09/21/2011</td>
<td>#lawrencebrewer</td>
<td>Murder of James Byrd, Jr.</td>
</tr>
<tr>
<td>09/23/2011</td>
<td>#kimdelaney</td>
<td>Kim Delaney</td>
</tr>
<tr>
<td>09/27/2011</td>
<td>#plzbelieve</td>
<td>Singaluuru</td>
</tr>
<tr>
<td>10/04/2011</td>
<td>#californiamoney</td>
<td>Addis Ababa Fistula Hospital</td>
</tr>
<tr>
<td>10/10/2011</td>
<td>#jova</td>
<td>Hurricane Jova</td>
</tr>
<tr>
<td>10/23/2011</td>
<td>#simoncelli</td>
<td>Marco Simoncelli</td>
</tr>
<tr>
<td>11/26/2011</td>
<td>#gogetitanddontlookback</td>
<td>The Legion of Missing Men</td>
</tr>
<tr>
<td>11/16/2011</td>
<td>#onlyatecu</td>
<td>ECU’92</td>
</tr>
<tr>
<td>11/27/2011</td>
<td>#garryspeed</td>
<td>Gary Speed</td>
</tr>
<tr>
<td>01/01/2012</td>
<td>#christmasfeast</td>
<td>Devils on horseback</td>
</tr>
</tbody>
</table>
Combining sources

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

News articles on Boston marathon attack
Wikipedia entry on Boston marathon attack
Tweet about Boston marathon attack
Facebook update on Boston marathon attack
Quora question on Boston marathon attack

April 15th 2013
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](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