Improving the Freshness of Web Collections by Integrating Social Web and Focused Web Crawling

Thomas Risse
L3S Research Center/Leibniz Universität Hannover

IFLA International News Media Conference
Social Media

Properties
• Important change in the communication on the internet
• Easy to create, share, or exchange information
• Easy to connect with family, friends, colleagues, interesting people
• Everybody is able to contribute
• Can be used everywhere
  • Independent of the location
  • Independent of the medium: Web, Smartphone, Smartwatch,

Societal View
• Good representation of our culture and society
• Valuable insights into individuals, groups, and organizations
• Enable an understanding of the public perception of events, people, products, or companies, including the flow of information
• Detailed insights into the day-to-day process of public communication
Twitter – A News Medium for Event-Following

Citizen Journalism
- Everybody can be a journalist by using Smartphone & Twitter
- E.g. Hudson River Plane Crash 2009

Event Discussions
- 2014 FIFA World Cup semi-final between Brazil and Germany on July 8, 2014
  → 35.6 Million tweets

→ Good documentation of the public perception of the event
Growing Interest in Web Archive Content

Journalists, Historians, Social Sciences, Law, …

- Relevant content
  - Official Publications (e.g. Government)
  - Journalistic Resources
  - Important topics and events with a high media coverage
  - Multi-cultural or controversial topics
- Observations of topics and events on major sites or Social Media are good starting points
- Metadata / Context (e.g. Author, Organizations and their interests, gender, location)
- Demographic information about social sites
- Provenance: Transparent and detailed documentation of content selection
Derived Requirements

Topical Dimension
- Crawl intention are mainly focused around events and rarely around entities
- What is the intention of the researcher?
- Easy monitoring by the researcher and possibility to correct

Flexible Crawling Strategies
- Shallow observation crawls (Social Media, Web)
- Focused crawls with prioritization (e.g. PageRank and/or semantics)

Social Web Crawling
- General interest with different media focus
- Integrated with Web crawler to capture the full context

Authenticity
- See a web page as the user saw the page (e.g. including ads and tweets at that time point)

Context and Provenance
- Demographics of sites
- Documentation of crawl specification and history
Is Twitter Content enough?

- A tweet is limited to the most important information
- Can we still understand the meaning and the context in the future?
- We need to make use of all hints we can get to ensure the interpretability
Web changes in response to current events

Internet Archive June 18th, 2015, 3:17 vs. 17:06 (same day)

Source: http://news.yahoo.com/shooting-erupts-church-charleston-south-carolina-021744448.html,
example by Bergis Jules (https://medium.com/on-archivy/the-narrative-of-terrorism-in-charleston-b8bd79d81741)
Current approach: Collect, then crawl

- Social Media: scalable access only through API
  - Requires special client programming and maintenance
  - Not supported by typical crawlers
- Workaround Process
  1. Crawling of Social Media content
  2. Extraction of Links
  3. Crawling of Web Pages
- Result
  - Static integration of Social Media
  - Uni-directional Path: Social Media → Web Content
  - Huge delay between time of post and time of crawling!
  - Missing Path: Web Content → Social Media
Integrated Crawling approach

- **Social Media API**
  - convenient query methods + (in Twitter) real-time stream
    → continuous stream of seeds for Web crawler
  - Social media URLs follow changes in topic
    → keeps crawler on topic even when topic evolves

- **Integrated Crawling**
  - API client and Web crawler cooperate through shared queue
  - URLs in Tweets are inserted early in the queue to ensure timely crawling
  - Suitable prioritization of URLs
  - Crawl continues also from tweeted URLs
Integrated crawling with the L3S iCrawl System

L3S iCrawl System (under development)
- Learning the intention of the crawl
- Integration of Web and Social Media Crawling
- Content based monitoring of the crawl process
Crisis in Ukraine: In depth news, commentary, and analysis ...

Ukraine News + Nachrichten / Ukraine aktuell
http://news.feed-reader.net/886-ukraine.html

Welcome to Ukraine
http://ukraine.org/

The Interpreter auf Twitter: "Ukraine Live: NATO Secretary General Tells Russian Foreign Minister To End Support for Ukrainian Rebels http://t.co/c8EX5TkP6"
https://twitter.com/Interpreter_Mag/status/600661031866758409

Civilian killed in intense Ukrainian army shelling of...
### Example for Integrated Crawling

#### Twitter #Ukraine Feed

- **Lars Olberg** @LarsOlberg: Media: #Ukraine may buy French Exocet anti-ship missiles, missilethreat.com/media-ukraine/... 5h
  - [Link](http://www.foxnews.com/world/2014/11/07/ukraine-accuses-russia-sending-in-dozens-tanks-other-heavy-weapons-into-rebel/)
- **Michael Baskin** @michael3o0e: #Ukraine accuses #Russia of sending in dozens of tanks, weapons into Russian separatists held areas. 5h
  - [Link](http://missilethreat.com/media-ukraine-may-buy-french-exocet-anti-ship-missiles/)
- Fox News: Ukraine accuses Russia of sending in dozens of tanks, weapons into Ukraine on Friday accused Russia of sending dozens of tanks and other heavy weapons into its rebel-controlled easternmost regions.

#### Crawler Queue

<table>
<thead>
<tr>
<th>ID</th>
<th>Batch</th>
<th>URL</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK3</td>
<td>x</td>
<td><a href="http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/">http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/</a></td>
<td>0.40</td>
</tr>
</tbody>
</table>

#### Web Link

- [http://missilethreat.com/articles/](http://missilethreat.com/articles/)
- [http://missilethreat.com](http://missilethreat.com)
Conclusions

Social Media Preservation
- Social Media can provide more than short term views
- Social Media preservation enable long term studies

Social Media Crawling
- Twitter crawls should include the context
- Context of the content
- Visual presentation

Freshness of Content
- Context of an event can evolve over time
- Social Media might point to the wrong context
- Limiting the time gap between Social Media and Web crawling

iCrawl System
- Under development
- Will be integrated into the SoBigData Research Infrastructure
Thank You!

Dr. Thomas Risse
Forschungszentrum L3S
Leibniz Universität Hannover
Appelstrasse 9a
30167 Hannover, Germany

E-Mail: risse@L3S.de
Telefon: +49-511-762 17764
Telefax: +49-511-762 17779