Efficient Extraction of Event-Centric Sub-Collections from the Web and Large Scale Web Archives

Dr. Elena Demidova

Data Science Seminar
May 11, 2016
University of Southampton
Vision

Information related to events and topics is currently scattered across different sources on the Web, in Web archives, digital libraries, news sources, open datasets, Linked Data and Social Media.

This information should evolve in a virtual informational space, providing different e.g. language-specific views, provenance and temporal context.
The iCrawl team

Gerhard Gossen

In the iCrawl project we develop methods and tools that enable collection of fresh and relevant Web, Social Web and Web archive content for a topic of interest through seamless integration of Web, Social Media and archived Web collections in a novel integrated focused crawler.

http://icrawl.l3s.uni-hannover.de/
Overview

Extraction of fresh and relevant event-centric collections from the Web

... and temporally & topically focused collections from Web archives.

Light-weight NLP and IR techniques for Web archive search.
Extraction of fresh and relevant event-centric collections from the Web

- How to extract event-centric fresh sub-collections from large-scale distributed Web sources?

[JCDL 2015, ECIR 2015]
Web changes in response to current events

Mystery London Marathon couple saying 'I do' revealed  
25 April 2016  
http://www.bbc.co.uk/news/uk-36128695
Motivation: Web archives about current events

- **Target users:**
  - Web Science researchers
  - Digital Humanities researchers
  - Journalists

- **Content:**
  - Activities on the Web around global events (e.g. Refugee crisis, Ebola outbreak, etc.) and local events.
  - People's views about politics, events, persons and popular topics shared on Social Media platforms.

- **iCrawl goal:**
  - Creation of event-centric collections of fresh and relevant Web documents.
Events on the Web and in Social Media: Challenges

- Newly created Web content is not yet well-interlinked and thus unlikely to be found by Web crawlers.
  → Continuous supply of fresh entry points is needed.

- Social Media are integral part of the Web
- Stories/conversations jump back and forth
- Web provides context to Social Media discussions
  → Necessary to collect both for full understanding!

- Rapid change of linked URLs (~few hours)
- Social Media collection can change based on crawl time!
  → URLs in Social Media need to be crawled soon after posting.
Social Media as a source for Web crawls: Opportunities

- Social Media typically link to new Web content.
  - E.g. news agencies often offer entry points to new content via Social Media (e.g. Twitter).
  - Also users pick up and post fresh content.
- Social Media offers convenient query methods (APIs).
- Twitter offers a real-time stream.
  - Continuous stream of seeds for Web crawler.
  - Especially for topic-focused crawls.
- Social Media URLs follow changes in the topic.
  → The stream can keep the crawler on topic even when the topic evolves.
Related approaches to Web archive creation

- Crawling with topical focus (1999 [1])
- iCrawl: Integrated crawling
- Crawling with temporal focus (2014 [2])
- Twitter API (2006)

Improving freshness of Web collections by integrating Social Web and focused Web crawling

**Approach:**

- Use Twitter stream to identify fresh and relevant Web content
- Integrate Web- and Twitter input to guide the focused crawler
  - Start with topic description
  - Follow outgoing links (only) to relevant pages
  - Integrate with Social Media crawling

http://icrawl.l3s.uni-hannover.de/
The iCrawl Wizard: Interactive crawl specification

London elections

Search sources: Web Twitter

Web

London Elects | Your London. Your vote.
https://www.londonelects.org.uk/
The official website of the upcoming 2016 Mayor of London and London Assembly elections.

https://www.london.gov.uk/elections/Pages/default.aspx

Keywords: Business, permit, property, Planning, licensing, site, page, information, Municipal Property, Rapid Transit
Entities: London, Windermere Road, Canada, Adelaide Street

Information on this website is available in alternate formats. Upon request. Contact Us: Elections Office

London mayoral election, 2016 - Wikipedia, the free encyclopedia

Keywords: page, links, Tube, councillor, - articles, funding, costs, media, terms


London mayoral election, 2012 - Wikipedia, the free encyclopedia

Keywords: page, Views, journalist, navigation, links, PDF, statement, tax, friend, radio
Entities: London, Ken Livingstone, Boris Johnson, Livingstone, UKIP

Twitter

Labour’s Khan becomes first Muslim mayor of London -

Keywords: UK, London, LondonMayor2016, Muslim, Twitter, housing, May, Party-officer, election, voters, campaign, PM, speakers
Entities: London, Khan, Saadi Khan, Scotland Labour, Corbyn Britain, Goldsmith, Panama, Labour Party

RT @ARYNEWSOFFICIAL: #London set for #Muslim mayor as elections shake UK parties https://londonmayor2016 Read more: https://t.co/V5vdXMtJaJ

Women’s Equality Party gets more than 350,000 votes and beats George Galloway in London | UK Politics | News | The Independent

Keywords: travel, business, money, family, ticket, years, comments, cookies, stay, job

RT @WEP_UK: Read @independent’s coverage of our success in Thursday’s elections https://t.co/WIk2sMvxbQ RT @spikedonline: Did the Women’s Equality Party @WEP_UK_ play the fear card in the elections? Brendan O’Neill says yes, party says no. http... One in every 20 Londoners voted for @SophieRunning as Mayor! Read more about the amazing @WEP_UK_ election results: https://t.co/7Jv4ZRT @catherine_mayer: Good @independent piece on @WEP_UK_’s election impact https://t.co/0mr3BcP7Rz

What should be top of Mayor Khan’s to do list | Red Box
http://www.thetimes.co.uk/redbox/topic/london-mayoral-elections-what-should-

http://icrawl.l3s.uni-hannover.de:8090/campaign/1/add

Dr. Elena Demidova
Integrated crawling approach

- API client and Web crawler cooperate through shared queue
- URLs in Tweets are inserted early in the queue to ensure timely crawling
- Crawl continues also from tweeted URLs
Focusing for freshness and relevance

- **Freshness**: \( \approx \) time between page modification and crawl time
  - Page modification date seldom explicit in Web pages
  - HTTP headers can refer irrelevant changes
    \( \rightarrow \) estimated from page content, fallback to HTTP headers

- **Topical relevance**: Is this Web page relevant to crawl topic?
  - Computed as similarity of word vector to reference vector

- Uncrawled URLs are assigned priority based on relevance of documents containing link

- Document priority: Weighted sum of freshness and topical relevance
ID | Batch | URL |
---|-------|-----|
UK 1 | 1 | http://www.foxnews.com/world/2014/11/07/ukraine-accuses-russia-sending-in-dozens-tanks-other-heavy-weapons-into-rebel/ | 1.00 |
UK 2 | 1 | http://missilethreat.com/media-ukraine-may-buy-french-exocet-anti-ship-missiles/ | 1.00 |
UK 3 | x | http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/ | 0.40 |
UK 4 | y | http://missilethreat.com/turkey-missile-talks-france-china-disagreements-erdogan/ | 0.05 |

Twitter #Ukraine Feed

Lars Olberg @MissileMonitor - 5h
Media: #Ukraine may buy French #Exocet anti-ship missiles, missilethreat.com/media-ukraine-may-buy-french-exocet-anti-ship-missiles

Michael Baskin @michael000e - 5h
#Ukraine accuses #Russia of sending in dozens of tanks, weapons into Russian separatists held areas fn.ws/snmv6lp

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

View in web

US-led strikes hit IS group oil sites for 2nd day

Associated Press published on http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/

Twitter #Ukraine Feed

Dr. Elena Demidova
Implementation

- Extension of Apache Nutch (open source web crawler)
- Twitter API fetcher component
- Prioritization module
- (+ easy to use UI)
- Available as open source software:

  http://icrawl.l3s.uni-hannover.de/
Evaluation

- Comparison of freshness and topical relevance on two event crawls
  - Ebola epidemic
  - Ukraine conflict
- Uses integrated approach: Twitter + Time + Topic focus (INT)
- Baselines:
  - Unfocused
  - Only Twitter (TB)
  - Only topic focus (FO)
- Crawls run in Nov 2014, configurations in parallel on separate machines
- On average 16K (Ebola)/13.8K (Ukraine) pages per configuration
Most frequent sites for the Ebola Epidemic crawls

<table>
<thead>
<tr>
<th>TB</th>
<th>UN</th>
<th>FO</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>直线</td>
<td>instagram.com</td>
<td>who.int</td>
<td>instagram.com</td>
</tr>
<tr>
<td>直线</td>
<td>linkis.com</td>
<td>reuters.com</td>
<td>vine.co</td>
</tr>
<tr>
<td>直线</td>
<td>newslocker.com</td>
<td>twitter.com</td>
<td>gnnliberia.com</td>
</tr>
<tr>
<td>直线</td>
<td>nytimes.com</td>
<td>thejournal.ie</td>
<td>ria.ru</td>
</tr>
<tr>
<td>直线</td>
<td>huffingtonpost.com</td>
<td>google.com</td>
<td>aerzte-ohne-grenzen.at</td>
</tr>
</tbody>
</table>

Twitter-based crawl: Social sites, newspapers and news aggregators
Focused Web crawl: Official sites (WHO, CDC - Centers for Disease Control and Prevention, ...), news
Freshness in the Ebola Epidemic crawls

Twitter input enables integrated crawler to:

- Improve freshness of the crawled collection
- Find relevant pages beyond the graph parts covered by seeds

(a) Ebola Epidemic Relevance

(a) Ebola Epidemic Freshness
Sub-collections on the Web: Contributions

- An efficient method to combine Social Media and focused crawling to create fresh event-centric sub-collections.
- Collects more fresh and relevant Web pages, with higher diversity compared to baselines.
- Open source prototype: http://icrawl.l3s.uni-hannover.de
Extraction of topically and temporally focused event-centric sub-collections from Web archives

• How to extract event-centric sub-collections from Web archives to include temporally and topically and relevant information.

[prelim. Version in WebSci 2016]
Extraction of event-centric datasets from large scale temporal Web collections

**German Web archive dataset:** Web pages from the .de top-level domain as captured by the Internet Archive from 12/1994 – 09/2013. Server content type of text/*. 31.13 TB, 4.8 billion captures, 1 billion URLs.

<table>
<thead>
<tr>
<th>Event</th>
<th>Type</th>
<th>Time span</th>
<th>Lead time</th>
<th>Cool-down</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olympic games</td>
<td>recurring</td>
<td>2 weeks</td>
<td>weeks</td>
<td>days</td>
</tr>
<tr>
<td>Federal election</td>
<td>recurring</td>
<td>1 day</td>
<td>months</td>
<td>weeks</td>
</tr>
<tr>
<td>Fukushima accident</td>
<td>nonrecurring</td>
<td>1 week</td>
<td>—</td>
<td>months</td>
</tr>
<tr>
<td>Snowden leaks</td>
<td>nonrecurring</td>
<td>1 day</td>
<td>—</td>
<td>years</td>
</tr>
</tbody>
</table>

http://alexandria-project.eu/
Collection specification for Snowden leaks

Topical Scope:

Reference documents:

https://de.wikipedia.org/w/Edward_Snowden
https://de.wikipedia.org/w/Globale_Überwachungs-und_Spionageaffäre

Keywords: NSA scandal, data monitoring, freedom of speech, Citizenfour, Edward Snowden

Temporal Scope:

The time interval of the event: [2013-06-01; 2013-06-30]
Duration of the lead time ($t_l$) = 0
Duration of the cool-down period ($t_r$) = 1 year
Estimating topical and temporal relevance

Q1: How can we effectively estimate topical and temporal relevance of Web documents with respect to the Collection Specification?

Combining topical and temporal relevance:

\[
TTR(d, CS) = \alpha \times TempR(d, CS) + (1 - \alpha) \times TopicR(d, CS)
\]

Topical relevance: Cosine similarity, entity and phrase boosting
Temporal relevance: Weibull function, exponential relevance decay with an increasing temporal distance.
Efficient sub-collection extraction

Q2: How can we efficiently create topically and temporally relevant and coherent event-centric sub-collections from broad scope temporal collections in Web archives?

- Problem:
  - Full-text indexing of entire collection is impractical
    - Time, storage space; large index is inefficient
    - Too many documents, no overview, no interlinking

- Approach:
  - Focused re-crawling of temporal collections: Crawler recursively follows the links in the Web archive starting from seed documents. Only documents with TTR over a threshold are selected.
  - Light-weight full – text index on URLs only to obtain seeds efficiently. Particularly useful in the news domain.
Manual evaluation of the relevance estimation

- A random sample containing 3.9 million documents (0.1%) of the entire dataset, topical relevance computation.
- Further sample of 50 documents to ensure inclusion of documents in both the positive and the negative class.
- 11 event collections including recurring events, Bundestag elections in Germany and Olympic games in different years as well as nonrecurring events such as the Fukushima nuclear disaster.
- 5 annotators (graduate CS students).
Evaluation: Nonrecurring vs. recurring events

Precision/recall diagrams for the relevance estimation of nonrecurring (on the left) vs. repeated (on the right) events.
Re-crawling: Harvest rate examples

Figure 2: Harvest rate of different crawl topics (accumulated)
Web archives sub-collections: Contributions

- Effective relevance estimation with respect to topical and temporal dimension.
- An efficient method to extract focused sub-collections from Web archives using focused re-crawling to create event-centric sub-collections for the past events.
Light-weight NLP and IR techniques for Web archive search

• How to find relevant seed pages in Web archives.

[IKC 2015]
URL analysis pipeline in Web archives

<table>
<thead>
<tr>
<th>URL</th>
<th>Entities</th>
</tr>
</thead>
</table>

Pre-processing ➔ Stop words removal ➔ Language detection ➔ Named entity extraction
The Popular German Web: A dataset description

Provided in the context of ALEXANDRIA project
- We generated a subset named Popular German Web.
- The subset contains 17 categories from 2000 to 2012 according to Alexa ranking.
- URLs and capture ids stored as CDX files.
Domain and temporal coverage of NER

- Overall 42,547,734 captures containing named entities have been identified by the extractor
- Frequency range: from 2,301,917 to 3

<table>
<thead>
<tr>
<th>Label</th>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>deutschland</td>
<td>location</td>
<td>2,301,917</td>
</tr>
<tr>
<td>berlin</td>
<td>location</td>
<td>628,300</td>
</tr>
<tr>
<td>hamburg</td>
<td>location</td>
<td>557,000</td>
</tr>
<tr>
<td>nordrhein</td>
<td>location</td>
<td>430,939</td>
</tr>
<tr>
<td>muenchen</td>
<td>location</td>
<td>405,845</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>michael jackson</td>
<td>person</td>
<td>30,210</td>
</tr>
<tr>
<td>tommy hilfiger</td>
<td>person</td>
<td>25,943</td>
</tr>
<tr>
<td>harald schmidt</td>
<td>person</td>
<td>25,176</td>
</tr>
<tr>
<td>heidi klum</td>
<td>person</td>
<td>21,291</td>
</tr>
<tr>
<td>merkel</td>
<td>person</td>
<td>17,835</td>
</tr>
</tbody>
</table>

Dr. Elena Demidova
Distribution of entities by domain category

![Bar chart showing distribution of entities by domain category from 2000 to 2012.]
URL analytics summary

- URL analytics towards providing efficient semantic annotations to large-scale Web archives
- Named entity recognition techniques can be effectively applied to URLs of the Web documents in order to provide an efficient way of initial document annotation
- Further aspects:
  - Good correlation between the URLs and document content
  - Temporal expressions in URLs
  - Precise retrieval results on URLs only
Summary

Use Social Media to create fresh event-centric Web collections. (JCDL 2015)

Support users during collection specification. (ECIR 2015)

Create temporal event-centric collections from Web archives using focused re-crawling. (WebSci 2016)

Analyse URLs for efficient seed generation. (IKC 2015)
Questions, Comments?

Dr. Elena Demidova

Senior Research Fellow
Web and Internet Science
University of Southampton

Contact: E.Demidova@soton.ac.uk

May 11, 2016