Accessing Web Archives

Maximilian Kaulmann
Outline

• Recap: Web Archives

• Publication Date Prediction through Reverse Engineering of the Web

• The Importance of Anchor Text for Ad Hoc Search Revisited
Recap: Web Archives

• What is a Web Archive?
  • Consists of all kinds of web resources
  • Stored in big files in the standardized WARC format
  • Preserve our history as documented on the web

• Three different “views” on Web Archives
  • User-centric
  • Data-centric
  • Graph-centric
Outline

• Recap: Web Archives

• Publication Date Prediction through Reverse Engineering of the Web
  • Introduction
  • Algorithm
  • Experiments and results
  • Conclusion

• The Importance of Anchor Text for Ad Hoc Search Revisited
Publication Date Prediction through Reverse Engineering of the Web

- Published
  - *On* February 8th, 2016
  - *By* Liudmila Ostroumova Prokhorenkova, Petr Prokhorenkov, Egor Samosvat, Pavel Serdyukov (all from Yandex)
  - *In* Proceedings of the Ninth ACM International Conference on Web Search and Data Mining
Introduction 1/2

- Why publication date prediction?
  - Temporal information retrieval
  - Web crawling policies

- Common ways to determine publication dates
  - Content-based methods
  - Metadata from web crawler
Introduction 2/2

• Limitations
  • Several candidates
  • Different formats
  • Documents without text (e.g. videos)
  • Too infrequent re-crawls
  • Expand index

• Related work
  • Link-based approaches

➢ Supplement content-based approach with link-based approach
Outline

• Recap: Web Archives

• Publication Date Prediction through Reverse Engineering of the Web
  • Introduction
  • Algorithm
  • Experiments and results
  • Conclusion

• The Importance of Anchor Text for Ad Hoc Search Revisited
Algorithm – General Scheme

• Initialize with no dates
• 3-stage algorithm
  • Stage 1: Content-based extraction
  • Stage 2: Date propagation
  • Stage 3: Likelihood optimization
Algorithm – Stage 1: Content-Based Extraction

1. Extract publication date candidates
   - Regular expressions
2. Prune corrupted dates
3. Select best candidate
   - Based on where the candidate was found (e.g. URL, page title)

- Result
  - Anchor dates
  - Seed dates

June 5th, 2018
Maximilian Kaulmann
Algorithm – Stage 2: Date Propagation

- Estimate publication date based on neighbors
- Various functions

Result
- Propagated dates
- Constant dates
Algorithm – Stage 2: Date Propagation (Example)
Algorithm – Stage 3: Likelihood Optimization

- Improve gathered dates
  - Seed dates
  - Propagated dates
  - Constant dates
- Based on *Evolution of the Media Web*
  - *attr*: Attractiveness of e.g. d1 for d4

\[
\text{Likelihood} = \prod_{t \in T} \text{attr}(d1, d4)
\]
Outline

• Recap: Web Archives

• Publication Date Prediction through Reverse Engineering of the Web
  • Introduction
  • Algorithm
  • Experiments and results
  • Conclusion

• The Importance of Anchor Text for Ad Hoc Search Revisited
Experiments – Datasets

Crawled by Yandex

- 70 hosts, 40M pages
- January 2013 to May 2014

- For each page $p$
  - URL
  - Document body
  - Outgoing links
    - $t_p^{\text{crawl}}$
    - $t_p^{\text{visit}}$

Memetracker

- 250k hosts, 12M pages, 29M links
- August 2008 to April 2009
- Blog posts and news articles

- For each page $p$
  - URL
  - Links
  - Publication date $t_p$
## Experiments – Date Extraction

### MAE for individual locations

<table>
<thead>
<tr>
<th>Location</th>
<th>MAE [days]</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>1.0</td>
</tr>
<tr>
<td>Before main content</td>
<td>4.3</td>
</tr>
<tr>
<td>Title</td>
<td>4.8</td>
</tr>
<tr>
<td>After main content</td>
<td>34.0</td>
</tr>
<tr>
<td>Inside main content</td>
<td>74.0</td>
</tr>
<tr>
<td>Elsewhere</td>
<td>88.0</td>
</tr>
</tbody>
</table>

### MAE for complete dataset

<table>
<thead>
<tr>
<th>Location</th>
<th>MAE [days]</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>124.0</td>
<td>0 %</td>
</tr>
<tr>
<td>Every</td>
<td>59.1</td>
<td>47.2 %</td>
</tr>
</tbody>
</table>

Add dates found in different locations (e.g. URL, title)

Source: Liudmila Ostroumova Prokhorenkova et al. Publication Date Prediction through Reverse Engineering of the Web. WSDM 2016 (Table 1)
## Experiments – Date Propagation and Likelihood Optimization

### Comparison on Yandex crawl

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE [days]</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed and anchor dates</td>
<td>57.1</td>
<td>47.7 %</td>
</tr>
<tr>
<td>1-step in-avg</td>
<td>55.9</td>
<td>49.3 %</td>
</tr>
<tr>
<td>1-step all-avg</td>
<td>55.5</td>
<td>64.4 %</td>
</tr>
<tr>
<td>model-0.6</td>
<td>51.2</td>
<td>69.7 %</td>
</tr>
<tr>
<td>Likelihood Optimization</td>
<td>49.9</td>
<td>69.7 %</td>
</tr>
</tbody>
</table>

### Comparison on Memetracker

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE [days]</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor dates</td>
<td>77.6</td>
<td>50.0 %</td>
</tr>
<tr>
<td>1-step out-avg</td>
<td>66.8</td>
<td>65.5 %</td>
</tr>
<tr>
<td>all-avg</td>
<td>58.0</td>
<td>75.3 %</td>
</tr>
<tr>
<td>model-0.4</td>
<td>58.0</td>
<td>75.3 %</td>
</tr>
<tr>
<td>Likelihood Optimization</td>
<td>57.4</td>
<td>75.3 %</td>
</tr>
</tbody>
</table>

### Source:
Liudmila Ostroumova Prokhorenkova et al. Publication Date Prediction through Reverse Engineering of the Web. WSDM 2016 (Table 2 and Table 3)
Outline

• Recap: Web Archives

• Publication Date Prediction through Reverse Engineering of the Web
  • Introduction
  • Algorithm
  • Experiments and results
  • Conclusion

• The Importance of Anchor Text for Ad Hoc Search Revisited
Publication Date Prediction – Conclusion

- Suggestion and comparison of methods for publication date estimation
- Improvement of the mean average error by 10% and 14% (depending on the used dataset)
Outline

- Recap: Web Archives
- Publication Date Prediction through Reverse Engineering of the Web
- The Importance of Anchor Text for Ad Hoc Search Revisited
  - Introduction
  - Experiments and results
  - Conclusion
The Importance of Anchor Text for Ad Hoc Search Revisited

• Published
  • *On* July 19th, 2010
  • *By* Marijn Koolen, Jaap Kamps (both from UVA, The Netherlands)
  • *In* Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval
Introduction

• The anchor text
• Anchor text as a short summary of the target page
• History
  • 2000: No improvements over content-only baselines
  • 2002: Effective for home page finding tasks
  • 2009: Effective for diversity in search results

• Experiences of search engines differ from results at TREC
• New collection: ClueWeb09
Outline

- **Recap**: Web Archives

- Publication Date Prediction through Reverse Engineering of the Web

- The Importance of Anchor Text for Ad Hoc Search Revisited
  - Introduction
  - Experiments and results
  - Conclusion
Experiments – Goals

• Compare full-text search with anchor text search

• Impact of link density and collection size

• Diversity of anchor text search results
Experiments – Preparation

- ClueWeb09 category B
  - (First) 50M pages of ClueWeb09 crawl

- Search engine Indri
  - Preprocessing
    - Stopword removal
    - Stemmed with Krovetz stemmer
  - Extract anchor text with *harvestlinks* (feature of Indri)

- Evaluation with TREC Web Track 2009
  - Task: Return 1000 relevant document for each of the 50 topics (queries)
## Experiments – Full-Text vs Anchor Text Search

<table>
<thead>
<tr>
<th>Run</th>
<th>Full collection</th>
<th>No Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>statMAP</td>
<td>MPC(30)</td>
</tr>
<tr>
<td>Text</td>
<td>0.1442</td>
<td>0.3079</td>
</tr>
<tr>
<td>Anchor</td>
<td>0.0567</td>
<td>0.5558</td>
</tr>
<tr>
<td>Mix</td>
<td>0.1643°</td>
<td>0.4812°</td>
</tr>
<tr>
<td>UDWAxQEWeb</td>
<td>0.1999</td>
<td>0.5010</td>
</tr>
<tr>
<td>uogTrdphCEwP</td>
<td>0.2072</td>
<td>0.4966</td>
</tr>
<tr>
<td>ICTNETADRun4</td>
<td>0.1746</td>
<td>0.4368</td>
</tr>
</tbody>
</table>

Source: Marijn Koolen and Jaap Kamps. The Importance of Anchor Text for Ad Hoc Search Revisited. SIGIR ‘10 (Table 1)
Experiments – Impact of Link Density

- Even few links improve search results
- High quality pages are robust against sampling

Source: Marijn Koolen and Jaap Kamps. The Importance of Anchor Text for Ad Hoc Search Revisited. SIGIR ‘10 (Figure 2)
Experiments – Impact of Collection Size

Collection size has higher impact on performance of anchor text index

Source: Marijn Koolen and Jaap Kamps. The Importance of Anchor Text for Ad Hoc Search Revisited. SIGIR ‘10 (Figure 4)
Anchor Text – Conclusion

• Can improve search results in combination with a full-text index

• Link density has small impact on anchor text effectiveness

• Collection size has big impact on anchor text effectiveness
Thank you!

Questions?
References

• Liudmila Ostroumova Prokhorenkova et al. Publication Date Prediction through Reverse Engineering of the Web. WSDM ’16
• Marijn Koolen and Jaap Kamps. The Importance of Anchor Text for Ad Hoc Search Revisited. SIGIR ’10