Information Search in Web Archives

Miguel Costa

Advisor: Prof. Mário J. Silva
Co-Advisor: Prof. Francisco Couto

Department of Informatics, Faculty of Sciences, University of Lisbon
PhD thesis defense, Lisbon, Portugal
November 4, 2014
The Web is Ephemeral

• 50 days - 50% of documents are changed (Cho and Garcia-Molina. 2000)

• 1 year - 80% of documents become inaccessible (Ntoulas, Cho and Olson. 2004)

• 27 months - 13% of web references disappear (http://webcitation.org/. 2007)
2014: Web Archiving Initiatives

- +68 initiatives in 33 countries
- +534 billions of web contents since 1996 (17 PB)
• Available since 2010: http://archive.pt
• 1.2 billion documents
Problem:
• it is hard to find past information with current Web Archive Information Retrieval (WAIR) systems

Objective:
• study the problems of WAIR and propose solutions
Contributions

1. Understanding WAIR systems
   - What is the state-of-the-art in WAIR?
   - What is the status of web archiving initiatives?
   - How are web archiving initiatives evolving?

2. Understanding web archive users
   - Does the state-of-the-art in WAIR meet the users’ information needs?
   - Why, what and how do web archive users search?
   - What functionalities would like the users to see implemented?
   - What are the specificities of web archive users?

3. Improving WAIR systems
   - How to improve WAIR?
   - How to evaluate WAIR systems?
   - What is the search effectiveness of the state-of-the-art in WAIR?
Understanding WAIR Systems
Methodology: 2 Surveys

- conducted in 2010 and 2014.
- questionnaires and public information.

What is the State-of-the-Art? URL Search

Did you want to see webpages with the text: http://sapo.pt?

Versions of the archived the Web pages

We archived 1,832 versions of the Web page http://sapo.pt from 1 January, 1996 and 26 August, 2013.

Technology based on the Wayback Machine.
Problem: URLs are hard to remember or unknown.
What is the State-of-the-Art? Full-text Search

- Technology based on **Lucene** extensions (NutchWAX & Solr).
- **Problem**: poor relevance rankings.
Understanding Web Archive Users
Methodology: 3 Data Collecting Methods

- Laboratory Studies
- Online Questionnaires
- Search Log Mining

Data richness

Generalization

[03/02/2012 21:16:11] QUERY fcul
[03/02/2012 21:16:19] CLICK RANK=1
What are the Users’ Information Needs?

- **Navigational** – 53% to 81%
  - seeing a web page in the *past* or how it evolved

- **Informational** – 14% to 38%
  - collecting information about a topic written in the *past*

- **Transactional** – 5% to 16%
  - downloading an old file or recovering a site from the *past*

Problems:
- Search engine technology optimized for different needs.
- Some needs are not supported by current technology.

Good news:
- Some needs may be supported by a high quality full-text search.
Improving WAIR
How to improve WAIR?

Previous studies show that temporal information:
• has been exploited to improve IR systems.
• can be extracted from web archives.

**Hypothesis:** state-of-the-art WAIR systems can be improved by exploiting temporal information intrinsic to web archives.
Exploiting Temporal Information

1. novel ranking features
   **Intuition:** persistent documents are more relevant for navigational queries.

2. novel ranking framework
   **Intuition:** ensemble of models learned for specific periods are more effective than a single ranking model.
Temporal Ranking Features

Documents with higher relevance tend to be more persistent (longer lifespan & more versions)
Temporal-Dependent Ranking Framework

- Learn a ranking model for each period.
- Use all data weighted by their temporal distance to the period.
- Combine models by minimizing a global loss function.

\[ \text{slope } \alpha \text{ (learning contribution)} \]
Temporal-Dependent Models

\[
\text{model} = \text{argmin}_f \sum_{i=1}^{m} L(f(x_i, \omega), y_i)
\]

\[
\omega = \text{parameters} \quad y_i = \text{relevance label}
\]

\[
\gamma = \text{temporal weight function}
\]

\[
\text{TD model} = \text{argmin}_f \sum_{i=1}^{m} L(\gamma(x_i, T_k) f(x_i, \omega), y_i)
\]

\[
\gamma(x_i, T_k) = \begin{cases} 
1 & \text{if } x_i \in T_k \\
1 - \alpha \frac{\text{distance}(x_i, T_k)}{|T|} & \text{if } x_i \notin T_k
\end{cases}
\]

\[
\alpha = \text{slope}
\]
Evaluation Methodology
Evaluation Methodology

• Test Collection (based on Cranfield Paradigm):
  – **Corpus**: 6 web collections, 255M contents, 8.9TB
  – **Topics**: 50 navigational (1/3 with date range)
  – **Relevance Judgments**: 3 judges, 3-level scale of relevance, 267 822 versions assessed
  – **Metrics**: (NDCG@k, P@k | k=1,5,10)

• Dataset for learning to rank (L2R):
  – 39 608 quadruples <query, version, grade, features>
  – 68 ranking features extracted (including temporal)
  – 5-fold cross-validation
Results & Validation of Thesis
State-of-the-Art vs. Learning-to-Rank (L2R)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Lucene</th>
<th>NutchWAX</th>
<th>AdaRank</th>
<th>Rank SVM</th>
<th>Random Forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@1</td>
<td>0.220</td>
<td><strong>0.250</strong></td>
<td>0.380</td>
<td>0.500</td>
<td><strong>0.550</strong></td>
</tr>
<tr>
<td>NDCG@5</td>
<td>0.157</td>
<td>0.215</td>
<td>0.427</td>
<td>0.485</td>
<td>0.610</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.133</td>
<td>0.174</td>
<td>0.470</td>
<td>0.523</td>
<td>0.650</td>
</tr>
</tbody>
</table>

All results show a statistical significance of $p<0.01$ with a two-sided paired t-test.

+ 30%
## Temporal Features vs. Without Temporal Features

All results show a statistical significance of $p<0.05$ with a two-sided paired t-test.

<table>
<thead>
<tr>
<th>Metric</th>
<th>L2R algorithms (without temporal features)</th>
<th>L2R algorithms (with temporal features)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AdaRank</td>
<td>Rank SVM</td>
</tr>
<tr>
<td>NDCG@1</td>
<td>0.380</td>
<td>0.500</td>
</tr>
<tr>
<td>NDCG@5</td>
<td>0.427</td>
<td>0.485</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.470</td>
<td>0.523</td>
</tr>
</tbody>
</table>

+ 10%
Temporal-Dependent Models vs. Single-models (without temporal features)

NDCG@10

slope
α = 0.25  α = 0.5  α = 0.75  α = 1  α = 1.25  α = 1.5

time intervals (using 14 years of web collections)

too large contribution
too small contribution

+ 5%
typical L2R
Conclusions
Conclusions

Answers to all research questions:

1. Understanding WAIR systems
   - Large increase of initiatives and volume of data, but smaller teams.
   - Only a small part of the web has been preserved.
   - State-of-the-art WAIR technology is optimized for different needs.
   - Some needs are not supported by state-of-the-art WAIR technology.

2. Understanding web archive users
   - Users have mostly navigational needs and then informational needs.
   - Users search as in web search engines.
   - Users prefer full-text search and older documents.

3. Improving WAIR systems
   - State-of-the-art WAIR systems have low search effectiveness.
   - An extension of the Cranfield paradigm can be used to evaluate WAIR.
   - State-of-the-art WAIR systems can be improved by exploiting temporal information intrinsic to web archives.
• Public service since 2010:
  – http://archive.pt

• OpenSearch API:

• Test collection to support evaluation:

• L2R dataset for WAIR research:
  – http://code.google.com/p/pwa-technologies/wiki/L2R4WAIR

• All code available under the LGPL license:
  – https://code.google.com/p/pwa-technologies/


Thank you.