Learning Temporal-Dependent Ranking Models

Miguel Costa, Francisco Couto, Mário Silva
LaSIGE @ Faculty of Sciences, University of Lisbon
IST/INESC-ID, University of Lisbon

37th Annual ACM SIGIR Conference, Gold Coast, Australia
July 10, 2014
Our Memory is in Digital Form

- E-books
- Web photo galleries
- Forums
- Blogs
- Online newspapers
- Social networks
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)
Will we face a Digital Dark Age?

The page cannot be found

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the httpd.apache.org home page, and then look for links to the information you want.
- Click the Back button to try another link.
- Click Search to look for information on the Internet.

HTTP 404 - File not found
Internet Explorer
2014: Web Archiving Initiatives

- +68 initiatives in 33 countries
- +534 billions of web contents since 1996 (17 PB)
PWA Search System

- 1.2 billion documents
  - searchable by full-text and URL
  - range between 1996 and 2013
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.
Full-text Search

Find the most relevant results

149.648.512
How to find the best search results for a given query in a Web Archive?

Typical solution: combine a set of proven ranking features using learning-to-rank (L2R) algorithms
We describe how to leverage the **temporal dimension** of web data by:

1. designing novel ranking features that exploit correlations between archived data and relevance

2. designing a novel ranking framework that learns models considering variations of data over time
Temporal Features
Long-term Document Persistence

• Predominant user information need: **navigational**.
• Query-independent ranking features do not work well
  – Much smaller volume of clicks
  – Sparser web-graphs
• We need alternatives

• Are long-term persistent documents more relevant?
• How to measure persistence?
  – lifespan
  – number of versions
Lifespan & Relevance

Documents with higher relevance tend to have a longer lifespan.

14 years of web snapshots analyzed.
# Versions & Relevance

Documents with higher relevance tend to have more versions.

14 years of web snapshots analyzed.
Modeling Document Persistence

\[ f(d) = \log_y(x) \]

Parameters:
\( x = \) \#versions/lifespan of document \( d \)
\( y = \) maximum \#versions/lifespan of a document in the collection
Temporal-Dependent Ranking Models
Temporal-Dependent Ranking

• The web has different characteristics over time:
  – more sites and pages
  – longer contents
  – different technologies
  – slightly different language
  – denser web-graphs

• Should we use a single-model that fits all data?
  – No: [Kang & Kim 2003; Geng et al. 2008; Bian et al. 2010]
Temporal Intervals

- use all data (do not split data by time)
- closer periods are more likely to hold similar web characteristics

Example:
- 3 intervals
- $T = \{ [t1,t2], [t2,t3], [t3,t4] \}$
Temporal-Dependent Models

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

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

\[
TD model = \arg \min_f \sum_{i=1}^{m} L(Y(x_i, Tk) f(x_i, \omega), y_i)
\]

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

\[
\alpha = \text{slope}
\]
Global Loss Function

- Results of temporal models are sub-optimal and hard to combine.
- Minimize a global loss function (correlation and overlap between models are considered).

\[
\text{model}_1, \ldots, \text{model}_n = \arg\min_{f_1, \ldots, f_n} \sum_{i=1}^{m} L \left( \sum_{j=1}^{n} Y(x_i, T_j) f_j(x_i, \omega), y_i \right)
\]

- **Scoring follows the global loss function.**

\[
\text{score}(x_i) = \sum_{j=1}^{n} Y(x_i, T_j) f_j(x_i, \omega)
\]
Experimental Setup
Research Questions

• Do temporal features extracted from web archives improve Web Archive IR?
  – Created a L2R dataset
  – L2R algorithms used: AdaRank, RankSVM, Random Forests.
  – L2R algorithms compared using the dataset with and without temporal features.

• Does the temporal-dependent ranking framework outperforms L2R single-models?
  – L2R algorithms used: RankSVM and TD RankSVM.
  – Temporal-dependent models compared with single-models.
Dataset for L2R in Web Archives

- 39,608 quadruples <query, version, grade, features>
  - 50 queries randomly sampled from logs
  - 843 versions assessed on average per query
  - 3-level scale of relevance
  - 68 ranking features extracted (including temporal)

- LETOR file format:

<table>
<thead>
<tr>
<th>Rel.</th>
<th>Query</th>
<th>Features</th>
<th>Doc. Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>qid:21</td>
<td>1:0.70 2:0.34 3:0.27 ... 68:0.86</td>
<td># id114746079</td>
</tr>
<tr>
<td>0</td>
<td>qid:22</td>
<td>1:0.05 2:0.18 3:0.14 ... 68:0.43</td>
<td># id172346033</td>
</tr>
<tr>
<td>1</td>
<td>qid:22</td>
<td>1:0.75 2:0.33 3:0.84 ... 68:0.54</td>
<td># id456334535</td>
</tr>
</tbody>
</table>
Evaluation Methodology

• Test Collection (based on Cranfield Paradigm):
  – **Corpus**: 6 web collections, 255M contents, 8.9TB
    • broad crawls, selective crawls, integrated collections
  – **Topics**: 50 navigational (with date range)
    • e.g. the page of Publico newspaper before 2000.
  – **Relevance Judgments**: 3 judges, 3-level scale of relevance, 267,822 versions assessed
  – **Metrics**: (NDCG@k, P@k | k=1,5,10)

• 5-fold cross-validation
  – 3 folders for training, 1 for validation, 1 for testing
Results
All results show a statistical significance of $p<0.05$ with a two-sided paired t-test.
Temporal-dependent models vs. Single-models (without temporal features)

NDCG@10

<table>
<thead>
<tr>
<th>time intervals (using 14 years of web collections)</th>
<th>α = 0.25</th>
<th>α = 0.5</th>
<th>α = 0.75</th>
<th>α = 1</th>
<th>α = 1.25</th>
<th>α = 1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>14 years</td>
<td><img src="image" alt="Graph" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 years</td>
<td><img src="image" alt="Graph" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 years</td>
<td><img src="image" alt="Graph" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 years</td>
<td><img src="image" alt="Graph" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 year</td>
<td><img src="image" alt="Graph" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

too large contribution

too small contribution

slope

+ 5%
Temporal-dependent models vs. Single-models
(with temporal features)

NDCG@10 over time intervals (using 14 years of web collections)

slope
- $\alpha = 0.25$
- $\alpha = 0.5$
- $\alpha = 0.75$
- $\alpha = 1$
- $\alpha = 1.25$
- $\alpha = 1.5$

+ 3.3%
Conclusions
Conclusions

• The evolution of web data over time can be exploited to improve the ranking of search results:
  • by designing novel temporal features
    – Relevant documents tend to have a longer lifespan and more versions.
  • by considering time when learning models
    – A model per period outperforms a single-model.

(Combined techniques produce the best results)

• Web archives are an excellent source to provide temporal information to web search systems.
• Public service since 2010:

• OpenSearch API:

• Test collection to support evaluation:

• L2R dataset for web archive IR research:

• All code available under the LGPL license:
Thank you. Questions?

migcosta@gmail.com