Bias in Learning to Rank Caused by Redundant Web Documents
Bachelor’s Thesis Defence

Jan Heinrich Reimer

Martin Luther University Halle-Wittenberg
Institute of Computer Science
Degree Programme Informatik

June 3, 2020
Duplicates on the Web

Example

Figure: The Beatles article and duplicates on Wikipedia—identical except redirect
Redundancy in Learning to Rank

**query**

**the beatles rock band**

**documents**

**labels**

**features**

**training**

**learning to rank model**

---

**Figure: Training a learning to rank model**

**Problems**

- identical relevance labels (Cranfield paradigm)
- similar features
- double impact on loss functions → overfitting
Duplicates in Web Corpora

- compare fingerprints hashes of documents, e.g., word $n$-grams
  - syntactic equivalence
  - near-duplicate pairs form groups
- 20% duplicates in web crawls, stable in time [Bro+97; FMN03]
  - up to 17% duplicates in TREC test collections [BZ05; Frö+20]
- few domains make up for most near duplicates
  - redundant domains often popular
- canonical links to select representative [OK12], e.g., Beatles $\rightarrow$ The Beatles
  - if no link assert self-link, then choose most often linked
  - resembles authors’ intent
Learning to Rank

- machine learning + search result ranking
- combine predefined features [Liu11, p. 5],
  e.g., retrieval scores, BM25, URL length, click logs, …
- standard approach for ranking: rerank top-\(k\) results
  from conventional ranking function
- prone to imbalanced training data

Approaches

pointwise predict ground truth label for single documents
pairwise minimize inconsistencies in pairwise preferences
listwise optimize loss function ranked lists
Learning to Rank Pipeline

1. deduplicate

2. novelty principle

features

split

train model

test model

evaluate

Figure: Novelty aware learning to rank pipeline for evaluation
Deduplication of Feature Vectors

- reuse methods for counteracting overfitting → undersampling
- active impact on learning
- deduplicate train/test sets separately

Full redundancy (100 %)
- use all documents for training
- baseline

No redundancy (0 %)
- remove non-canonical documents
- algorithms can’t learn about non-canonical documents

Novelty-aware penalization (NOV)
- discount non-canonical documents’ relevance
- add flag feature for most canonical document
Novelty Principle [BZ05]

- deduplication of search engine results
- users don’t want to see the same document twice

Duplicates unmodified

overestimates performance [BZ05]

Duplicates irrelevant

users still see duplicates

Duplicates removed

no redundant content → most realistic
Learning to Rank Datasets

<table>
<thead>
<tr>
<th>Year</th>
<th>Name</th>
<th>Duplicate detection</th>
<th>Queries</th>
<th>Docs. / Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>LETOR 3.0 [Qin+10]</td>
<td>✗</td>
<td>681</td>
<td>800</td>
</tr>
<tr>
<td>2009</td>
<td>LETOR 4.0 [QL13]</td>
<td>✓</td>
<td>2.5K</td>
<td>20</td>
</tr>
<tr>
<td>2011</td>
<td>Yahoo! LTR Challenge [CC11]</td>
<td>✗</td>
<td>36K</td>
<td>20</td>
</tr>
<tr>
<td>2016</td>
<td>MS MARCO [Ngu+16]</td>
<td>✓</td>
<td>100K</td>
<td>10</td>
</tr>
<tr>
<td>2020</td>
<td>our dataset</td>
<td>✓</td>
<td>200</td>
<td>350</td>
</tr>
</tbody>
</table>

- duplicate detection only possible for LETOR 4.0 and MS MARCO
- shallow judgements in existing datasets
- create new deeply judged dataset from TREC Web ’09–’12
- worst-/average-case train/test splits for evaluation
Evaluation

▶ train & rerank common learning-to-rank models: regression, RankBoost [Fre+03], LambdaMART [Wu+10], AdaRank [XL07], Coordinate Ascent [MC07], ListNET [Cao+07]
▶ settings: no hyperparameter tuning, no regularization, 5 runs
▶ remove BM25 = 0 (selection bias in LETOR [MR08])
▶ BM25@body baseline for comparison

Experiments

▶ retrieval performance / nDCG@20 [JK02]
▶ ranking bias / rank of irrelevant duplicates
▶ fairness of exposure [Bie+20]
Retrieval Performance on ClueWeb09
Evaluation with Deep Judgements

Figure: nDCG@20 performance for ClueWeb09, with Coordinate Ascent
Retrieval Performance on GOV2
Evaluation with Shallow Judgements

Figure: nDCG@20 performance for GOV2, with AdaRank
Retrieval Performance

Evaluation

- performance decreases by up to 39% under novelty principle
- improvement with penalization of duplicates, compensates novelty principle impact
- significant changes only for some algorithms, mostly when duplicates irrelevant
- slightly decreased performance when deduplicating without novelty principle
- all learning to rank models better than BM25 baseline
Ranking Bias on ClueWeb09
Evaluation with Deep Judgements

Figure: First irrelevant duplicate rank for ClueWeb09, with Coordinate Ascent
Ranking Bias on GOV2
Evaluation with Shallow Judgements

**Figure:** First irrelevant duplicate rank for GOV2, with AdaRank
 Ranking Bias

Evaluation

▸ irrelevant duplicates ranked higher under novelty principle, often top-10
▸ bias towards duplicate content
▸ removing/penalizing duplicates counteracts bias significantly
▸ more biased than BM25 baseline
▸ implicit popularity bias as redundant domains are most popular
▸ poses risk at search engines using learning to rank
Fairness of Exposure [Bie+20]

Evaluation

Figure: Fairness of exposure for ClueWeb09 and GOV2

- no significant effects
- fairness measures unaware of duplicates
- duplicates should count for exposure, not for relevance
- tune Biega’s parameters → trade-off fairness vs. relevance [Bie+20]
- experiment with other fairness measures
Conclusion

▶ near-duplicates present in learning-to-rank datasets
  ▶ reduce retrieval performance
  ▶ induce bias
  ▶ don’t affect fairness of exposure

▶ novelty principle for measuring impact
▶ deduplication to prevent

Future Work

▶ direct optimization [Xu+08] of novelty-aware metrics [Cla+08]
▶ reflect redundancy in fairness of exposure
▶ experiments on more datasets (e.g., Common Crawl) and more algorithms (e.g., deep learning)
▶ detect & remove vulnerable features

Thank you!


Ohye, Maile et al. (Apr. 2012). The Canonical Link Relation. RFC 6596.
Wikipedia Bias on ClueWeb09
Evaluation with Deep Judgements

Figure: First irrelevant Wikipedia rank for ClueWeb09, with Coordinate Ascent
Fairness of Exposure on ClueWeb09 [Bie+20]

Evaluation with Deep Judgements

Figure: Fairness of exposure for ClueWeb09, with Coordinate Ascent
Fairness of Exposure on GOV2 [Bie+20]
Evaluation with Shallow Judgements

Figure: Fairness of exposure for GOV2, with AdaRank