Exploiting Citation Contexts for Physics Retrieval

Anna Dabrowska and Birger Larsen

BIR 2015
29th March, 2015

AALBORG UNIVERSITY
DENMARK

Motivation

- Changing publication norms for academic literature
- Anchor text in web retrieval
- Previous research using citation contexts in retrieval
  - O'Connor (1982), Bradshaw (2003), Ritchie (2009)
- Availability of the iSearch test collection

Objectives

- Determine feasibility:
  - Can we reliably identify and extract citation contexts in the iSearch test collection?

- Introduce citation contexts into retrieval:
  - How do citation context of different fixed window sizes impact retrieval performance?

  - What is the effect of altering the weight assigned to citation contexts?

The iSearch Test Collection

- 434,813 Physics documents (XML files)
  - ~160,000 full text with metadata
  - ~275,000 abstract-only with metadata
  - Acquired from ArXiv.org in 2009

- 65 search tasks (i.e. “topics”) with information need descriptions

- Document relevance assessments on a 4-point scale (0-3)

Additional Data

- Direct citation data from CiteBase
  - 3.7 million direct citations in iSearch
  - 259,093 unique cited documents

- Source files from ArXiv.org (TeX files)
  - Complete repository acquired in 2014

- An id concordance file matching iSearch documents with ArXiv documents
  - PF417005 > astro-ph.9903338

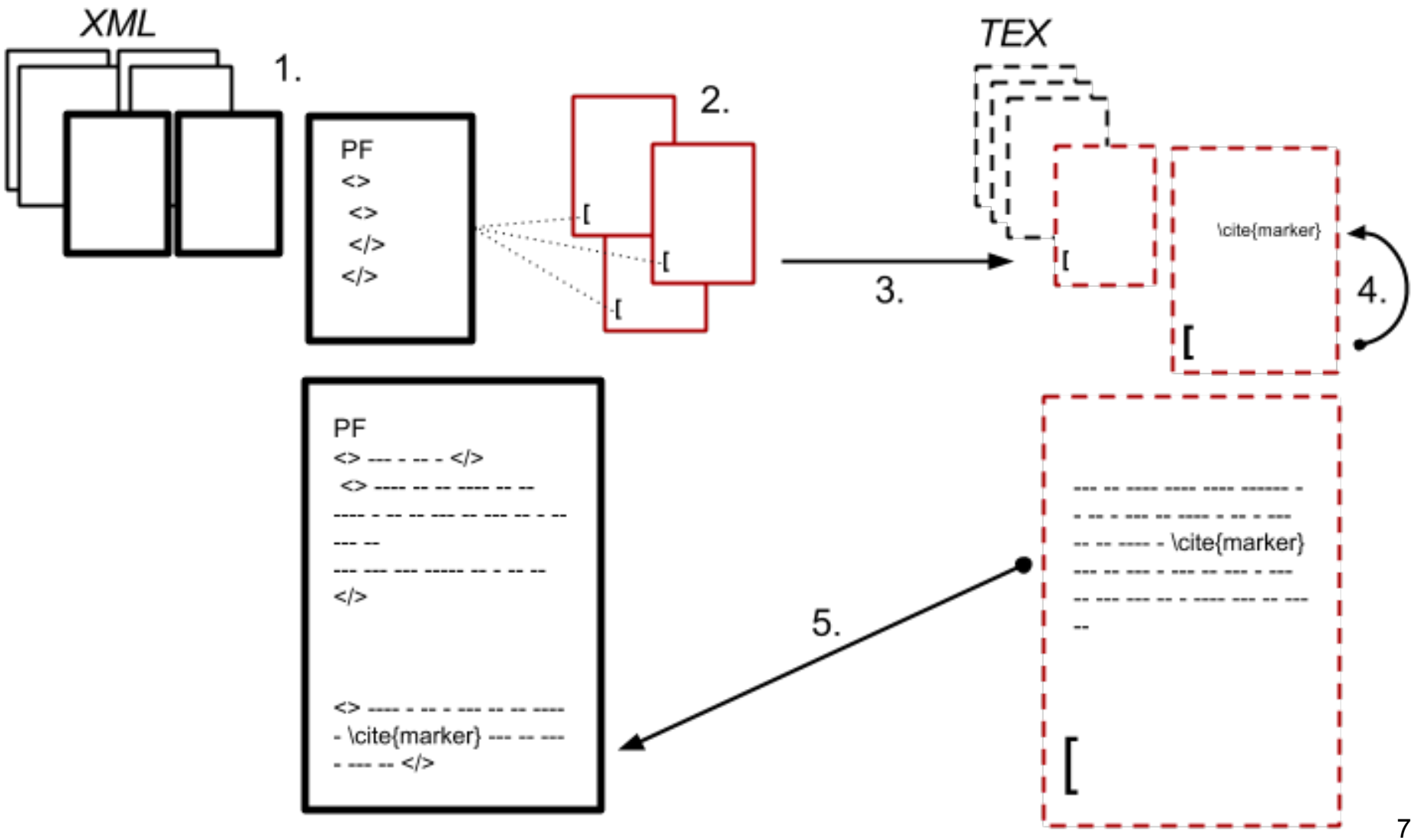
Method Overview

- Conduct initial retrieval for document subset
  - Extract citation contexts from documents citing retrieved documents

- Append citation contexts to cited documents
  - Fixed windows up to 25, 50, 75, and 100 words surrounding in-text citation

- Conduct retrieval experiments
  - Incrementally weight each fixed citation context window (α) relative to document text (1-α)

Document Processing



Retrieval

- Indri Search Engine
  - Language Modelling, Dirichlet smoothing (μ, 0-5000)
  - No stopping, Krovetz stemming
  - Using 64 iSearch topics (topic 5 is excluded)
  - Retrieving 1000 documents per topic

- Evaluation with trec_eval
  - Mean Average Precision (MAP)
  - Normalized Discount Cumulative Gain (nDCG)
  - Upper-bound performance

Indri Query Language

- Ad hoc query:

```
#combine(manipulation nano spheres)
```

- Experimental query, α=0.3:

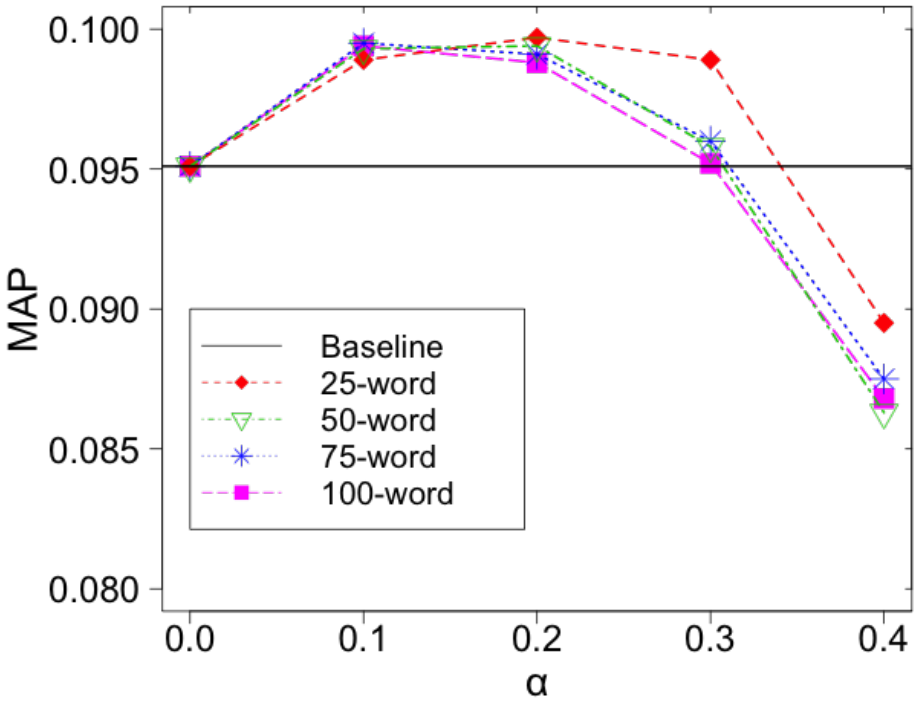
```
#weight(
0.7 #combine(manipulation.(orig_doc) nano.
(orig_doc) spheres.(orig_doc))
0.3 #combine(manipulation.(25word_citContext)
nano.(25word_citContext)
spheres.(25word_citContext)))
```

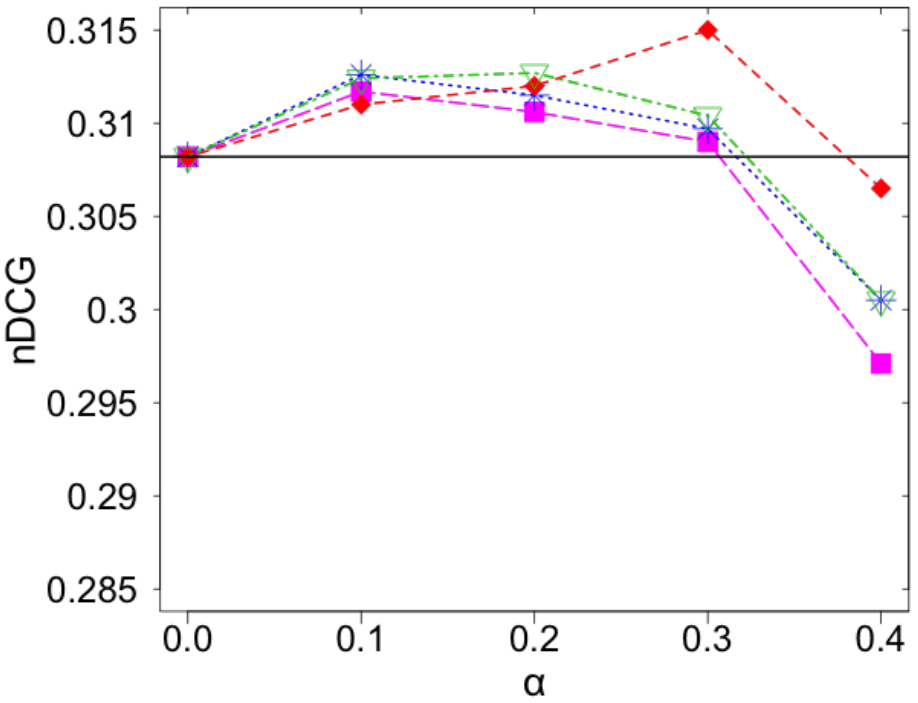
Preliminary Retrieval and Citation Contexts

- 52,586 unique documents ranked by the combination of both MAP and nDCG result sets
  - 25,356 (48%) are cited at least once
  - 1,577 relevant (assessed 1–3)
  - Average citation count: 13.2
  - Highest citation count: 4,904

- 19,248 documents have at least one citation context appended
  - 12,986 with contexts from multiple citing documents
  - 399 relevant (assessed 1–3)
  - 941 irrelevant (assessed 0)

Experimental Retrieval Results





Baseline retrieval scores are relatively low (α = 0.0)

MAP = 0.0951

nDCG = 0.3082

Experimental Retrieval Results

- Citation context runs improved slightly over the baseline, without statistical significance
  - $0.1 \leq \alpha \leq 0.3$ tended to score higher than document text alone
  - $\alpha \geq 0.4$ decreased performance in all window sizes
  - Weighting affected scores differently across window sizes

- Best performance with 25-word windows, but differences were small
  - 4.8% increase over baseline MAP
  - 2.2% increase over baseline nDCG

Current and Future Work

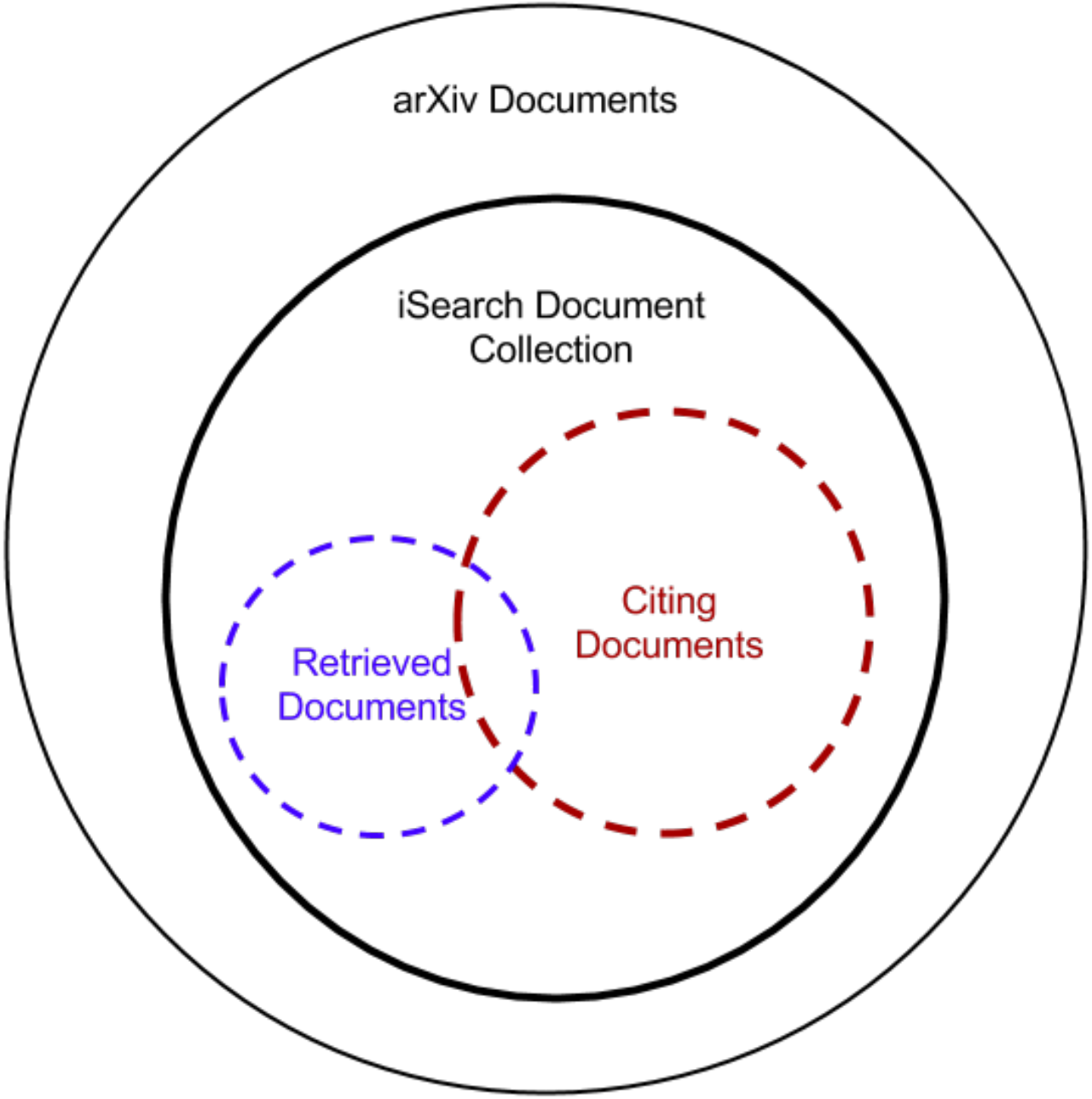
- Full iSearch collection
- Greater coverage for each cited document
- Cleaned citation contexts
- Refined window sizes
- 3-fold cross-validation

- Linguistic analysis
- Query by query analysis
- Comparison between full text and abstract-only documents

Thank you

Questions?

AALBORG UNIVERSITY
DENMARK


arXiv Documents
iSearch Document Collection
Retrieved Documents
Citing Documents