RSLIS at INEX 2013
Social Book Search track

Toine Bogers
Birger Larsen

Royal School of Library & Information Science
Copenhagen, Denmark

Outline

- Methodology
  - Pre-processing
  - Indexing & topics
- Content-based retrieval
- What now?

Methodology

Pre-processing

- Retained 19 content-bearing XML fields
  - <isbn>, <title>, <publisher>, <editorial>, <creator>, <series>, <award>, <character>, <place>, <blurber>, <epigraph>, <firstwords>, <lastwords>, <quotation>, <dewey>, <subject>, <browseNode>, <review>, and <tag>

Indexing

- Created six different indexes
  - All fields (all-doc-fields)
  - Metadata (metadata)
  - Content (content)
  - Controlled metadata (controlled-metadata)
  - Tags (tags)
  - User reviews (reviews)

Topics

- Three different topic representations
  - Query (query)
  - Three original topic fields combined (all-topic-fields)
    - Title, group, narrative
  - All four topic fields combined (all-plus-query)
    - Title, group, narrative, query

Content-based retrieval

Approach

- Optimized retrieval parameters using all-topic-fields topic representation on 2012 topic set
  - Query field is new addition in 2013
- Algorithm
  - Language modeling using JM smoothing
  - λ optimized in steps of 0.1 in [0, 1] range
  - Stopword filtering & Krovetz stemming

Optimization results

Document fields	Topic fields
	all-topic-fields
metadata	0.2015
content	0.0115
controlled-metadata	0.0496
tags	0.2056
reviews	0.2832
all-doc-fields	**0.3058**

Optimization results

Submitted runs

Submitted runs

- Three submitted runs
  - Run 1: query.all-doc-fields
  - Run 2: all-topic-fields.all-doc-fields
  - Run 3: all-plus-fields.all-doc-fields

Results

Run #	Run description	NDCG@10	P@10	MRR
1	query.all-doc-fields	0.0401	0.0208	0.0635
2	all-topic-fields.all-doc-fields	0.1295	0.0647	0.2190
3	all-plus-query.all-doc-fields	**0.1361**	**0.0653**	**0.2286**

- Again, combining more representations = better performance!

What now?

Do we have a problem?

as measured by NDCG@10

- Best run does nothing fancy!
  - All topics representations + all document fields outperforms anything else we can throw at this
  - So nothing fancy we do has any effect?
  - What's next...?

What have we done so far?

What have we done so far?

- Standard retrieval
  - Typically using Indri w/ stopword filtering and Krovetz stemming
  - Different combinations of document fields & topic representations
  - Most participants have an all-fields run, but results are not the same!
  - Feature selection techniques show some promise here for determining optimal field set!

What have we done so far?

- Re-ranking of retrieved books based on
  - Book ratings (4 times)
  - Review helpfulness (4)
  - Tag overlap (2)
    - Personalized and non-personalized
  - Never beats the baseline!

What have we done so far?

- Query expansion/pseudo-relevance feedback
  - All document fields
  - Tags (3)
  - Title
  - Subject headings
  - Wikipedia (2)
  - Never beats the baseline!

What have we done so far?

- Linear combination of memory-based collaborative filtering + competitive baseline run
  - Significant improvement over the baseline on 2012 topic set
  - No improvement over the baseline on 2013 topic set (?)

What does this mean?

- Directions for the future
  - Determine optimal collection of fields
  - Stop looking at re-ranking using review scores or helpfulness
  - Investigate the recommendation aspect more!
    - Explore the value of collaborative filtering

Questions?