RSLIS at INEX 2011
Social Book Search track

Toine Bogers

Kirstine Wilfred Christensen

Birger Larsen

Royal School of Library & Information Science / DBC
Copenhagen, Denmark

Outline

- Methodology
  - Pre-processing
  - Indexing & topics
- Content-based retrieval
- Social re-ranking
- Submitted runs
- Discussion

Methodology

Pre-processing

- Removed 22 XML fields not likely to contribute to retrieval
  - Example: <image>, <listprice>, <binding>
- 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)
    - All 19 content-bearing XML fields
  - Metadata (metadata)
    - Immutably tied to the book, provided by publisher
    - <title>, <publisher>, <editorial>, <creator>, <series>, <award>, <character>, and <place>

Indexing

- Content (content)
  - Fields that contain some part of the book text
  - <blurber>, <epigraph>, <firstwords>, <lastwords>, and <quotation>
- Controlled metadata (controlled-metadata)
  - Subject descriptions curated by library professionals
  - <browseNode>, <dewey>, and <subject>

Indexing

- Tags (tags)
  - User-generated subject descriptions
  - <tag>
- User reviews
  - Book-centric index reviews (all reviews belonging to the same book aggregated into a single representation)
  - Review-centric index reviews-split (each review indexed separately)

Topics

- Four different topic representations
  - Title (title)
  - Group (group)
  - Narrative (narrative)
  - All three topic fields combined (all-topic-fields)

Content-based retrieval

Approach

- Pairwise combinations of all indexes and topic representations
  - 6 indexes × 4 representations = 24 different runs
- Algorithm
  - Language modeling using JM smoothing
  - λ optimized in steps of 0.1 in [0, 1] range
  - Stopword filtering & Krovetz stemming

Results

Document fields	Topic fields			
	title	narrative	group	all-topic-fields
metadata	0.2756	0.2660	0.0531	0.3373
content	0.0083	0.0091	0.0007	0.0096
controlled-metadata	0.0663	0.0481	0.0235	0.0887
tags	0.2848	0.2106	0.0691	0.3334
reviews	**0.3020**	0.2996	0.0773	0.3748
all-doc-fields	0.2644	**0.3445**	**0.0900**	**0.4436**

Social re-ranking

Approach

- Tags
  - Tag index tags performed well
- Reviews
  - Book-centric index reviews performed well
  - What about the review-centric index reviews-split?

Approach

- Review-centric retrieval

   1. Retrieve individual reviews

   2. Aggregate scores for individual reviews into a single relevance score for each occurring book

      - Similar to results fusion in IR!

      - Can use methods like CombMAX, CombSUM, etc.

Approach

- Unweighted review fusion
  - CombMAX, CombSUM, and CombMNZ
- Weighted review fusion
  - Weighting based on review helpfulness

$$score_{weighted}(i) = score_{org}(i) \times \frac{helpful\ vote\ count}{total\ vote\ count}$$

  - Weighting based on normalized book ratings

$$score_{weighted}(i) = score_{org}(i) \times \frac{r}{5}$$

Results

	Runs	Topic fields			
		title	narrative	group	all-topic-fields
reviews-split	CombMAX	0.3117	**0.3222**	0.0892	0.3457
	CombSUM	**0.3377**	0.3185	**0.0982**	**0.3640**
	CombMNZ	0.3350	0.3193	**0.0982**	0.3462
	CombMAX - Helpfulness	0.2603	0.2842	0.0722	0.3124
	CombSUM - Helpfulness	0.2993	0.2957	0.0703	0.3204
	CombMNZ - Helpfulness	0.3083	0.2983	0.0756	0.3203
	CombMAX - Ratings	0.2882	0.2907	0.0804	0.3306
	CombSUM - Ratings	0.3199	0.3091	0.0891	0.3332
	CombMNZ - Ratings	0.3230	0.3080	0.0901	0.3320
	reviews	0.3020	0.2996	0.0773	0.3748

Submitted runs

Submitted runs

- Four submitted runs
    - Run 1: title.all-doc-fields
    - Run 2: all-topic-fields.all-doc-fields
    - Run 3: title.reviews-split.CombSUM
    - Run 4: all-topic-fields.reviews-split.CombSUM

Results

- Best-performing runs
  - Run 2: all-topic-fields.all-doc-fields
  - Run 4: all-topic-fields.reviews-split.CombSUM
- Means there is hope for the social re-ranking approach...

Discussion

What did we learn?

- Best performance when combining all available information
  - Support for principle of polyrepresentation
    - Ingwersen (1996) and Belkin (1993)
- User-generated metadata ≫ curated metadata
- Book-centric vs. review-centric undecided
  - Helpfulness and ratings do not contribute enough in the current approach

Questions?