RSLIS at INEX 2012
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
- Social re-ranking
- Submitted runs
- Discussion

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>
- Merged the BL and LoC metadata with the relevant fields

Indexing

- Created eight different indexes
  - All fields (all-doc-fields)
    - Separate version including the BL/LoC data (all-doc-fields-plus)
  - Metadata (metadata)
  - Content (content)

Indexing

- Controlled metadata (controlled-metadata)
  - Separate version including the BL/LoC data (controlled-metadata-plus)
- Tags (tags)
- User reviews (reviews)

Topics

Two

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

Content-based retrieval

Approach

- Pairwise combinations of all indexes and topic representations on 2011 test topics
  - 8 indexes × 2 representations = 16 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	all-topic-fields
metadata	0.0915	0.2015
content	0.0108	0.0115
controlled-metadata	0.0406	0.0496
controlled-metadata-plus	0.0514	0.0691
tags	0.0792	0.2056
reviews	0.1041	0.2832
all-doc-fields	**0.1129**	**0.3058**
all-doc-fields-plus	0.1120	0.3029

Social re-ranking

Two approaches

- Book similarity re-ranking
  - Similarity between books helps move similar books closer together in the results list
- Personalized re-ranking
  - Take into account the past preferences of the topic creator → books similar to past reads are pushed upwards

Book similarity re-ranking

- Two books retrieved at wildly different ranks can still be very similar in other aspects
  - Can including these different types of book similarities help improve results?
    - Relevant books are similar in many aspects
    - Ideally, relevant books are a contiguous block at the top of the results list
    - Solution: move similar books closer together in the results list

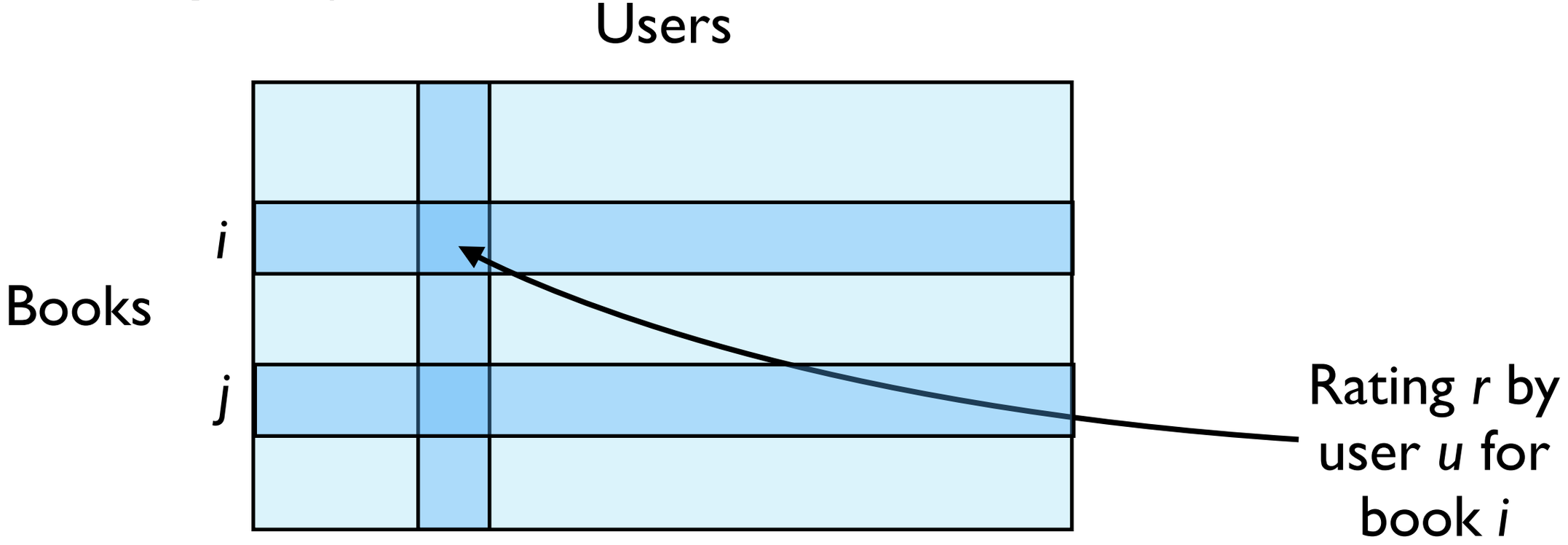
Book similarity re-ranking

- Every retrieved book *i* borrows a bit of the retrieval score of every other retrieved book *j*
- More similar books should borrow more from each other
- Original retrieval score should continue to play a role in this → parameter α controls this

$$score_{re-ranked}(i) = \alpha \cdot score_{org}(i) + (1-\alpha) \cdot \sum_{j=1, i \neq j}^{n} score_{org}(j) \cdot sim(i,j)$$

Book similarities

- Five different types of book similarities
  - IU-similarity is cosine similarity of two book rating vectors *i* and *j* from user reviews (inspired by CF)



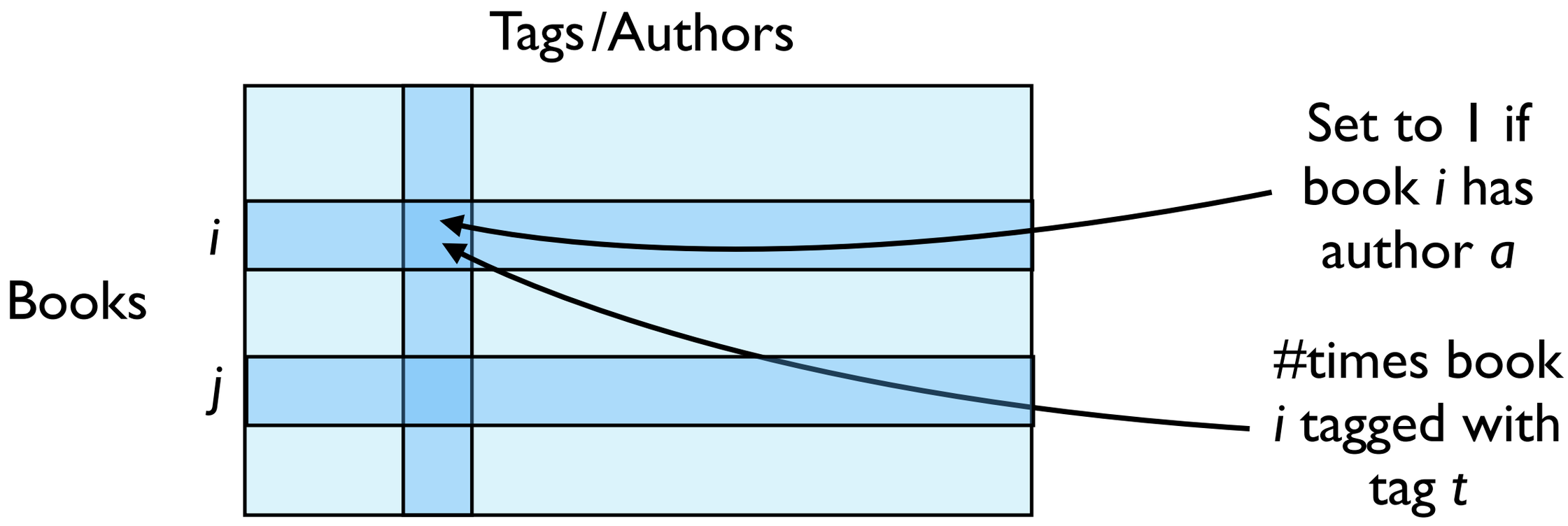
Book similarities

- ll-similarity is derived from Amazon's *"similar products"* data

- Set to 1 if a book pair is included in the collection
- Based on CF on all of Amazon

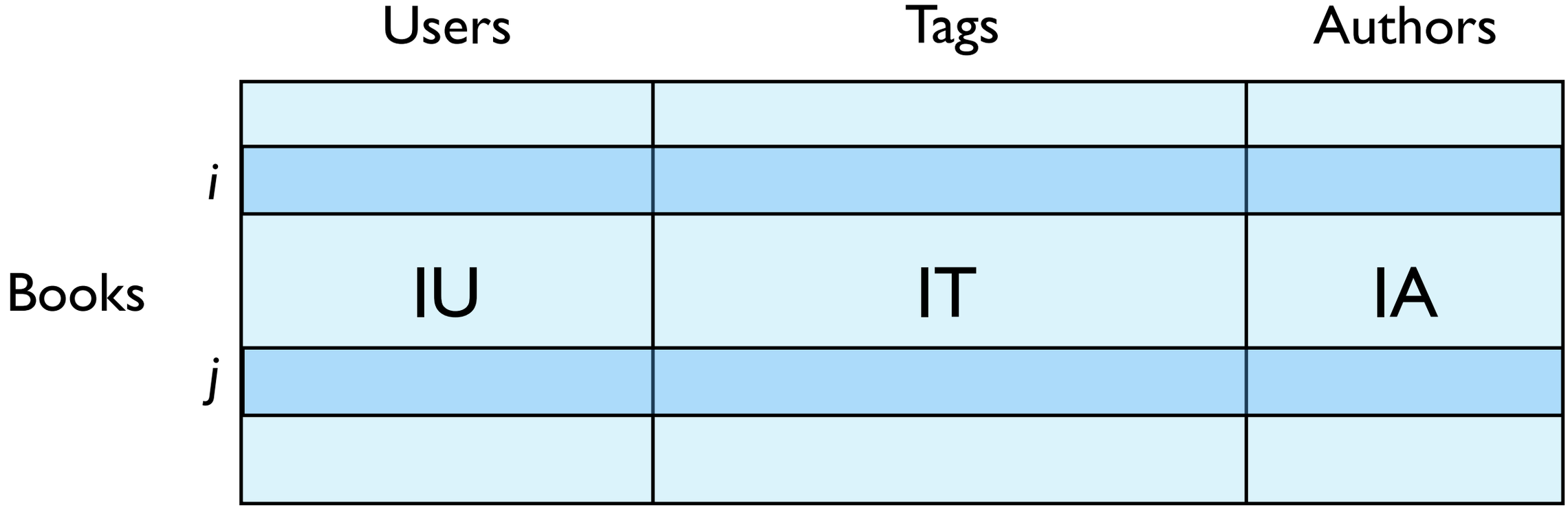
Book similarities

- IT-similarity is cosine similarity of two book-tag vectors *i* and *j*
- IA-similarity is cosine similarity of two book-author vectors *i* and *j*



Book similarities

- IUTA-similarity is cosine similarity on fused IU, IT, and IA matrices



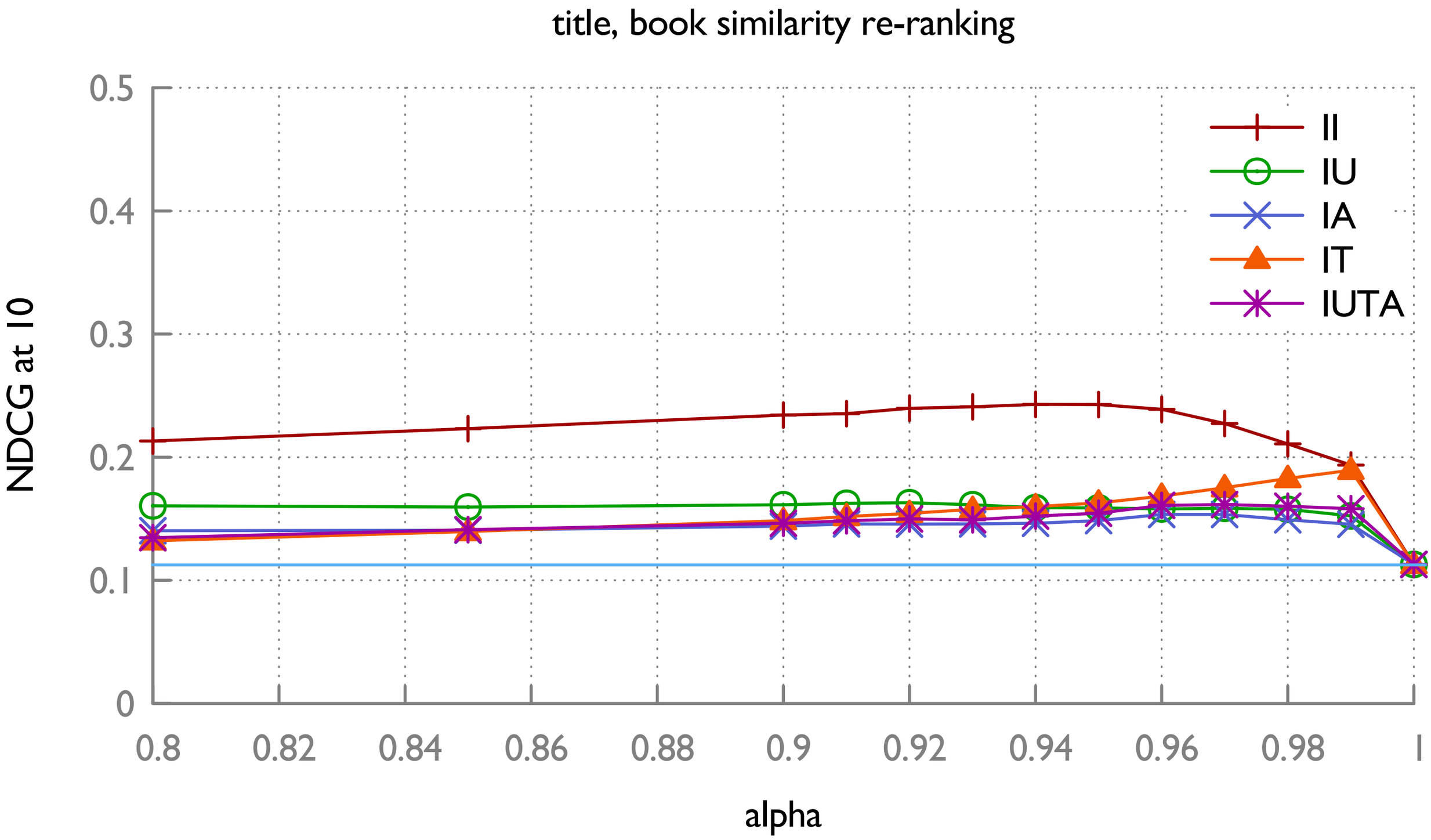
Personalized re-ranking

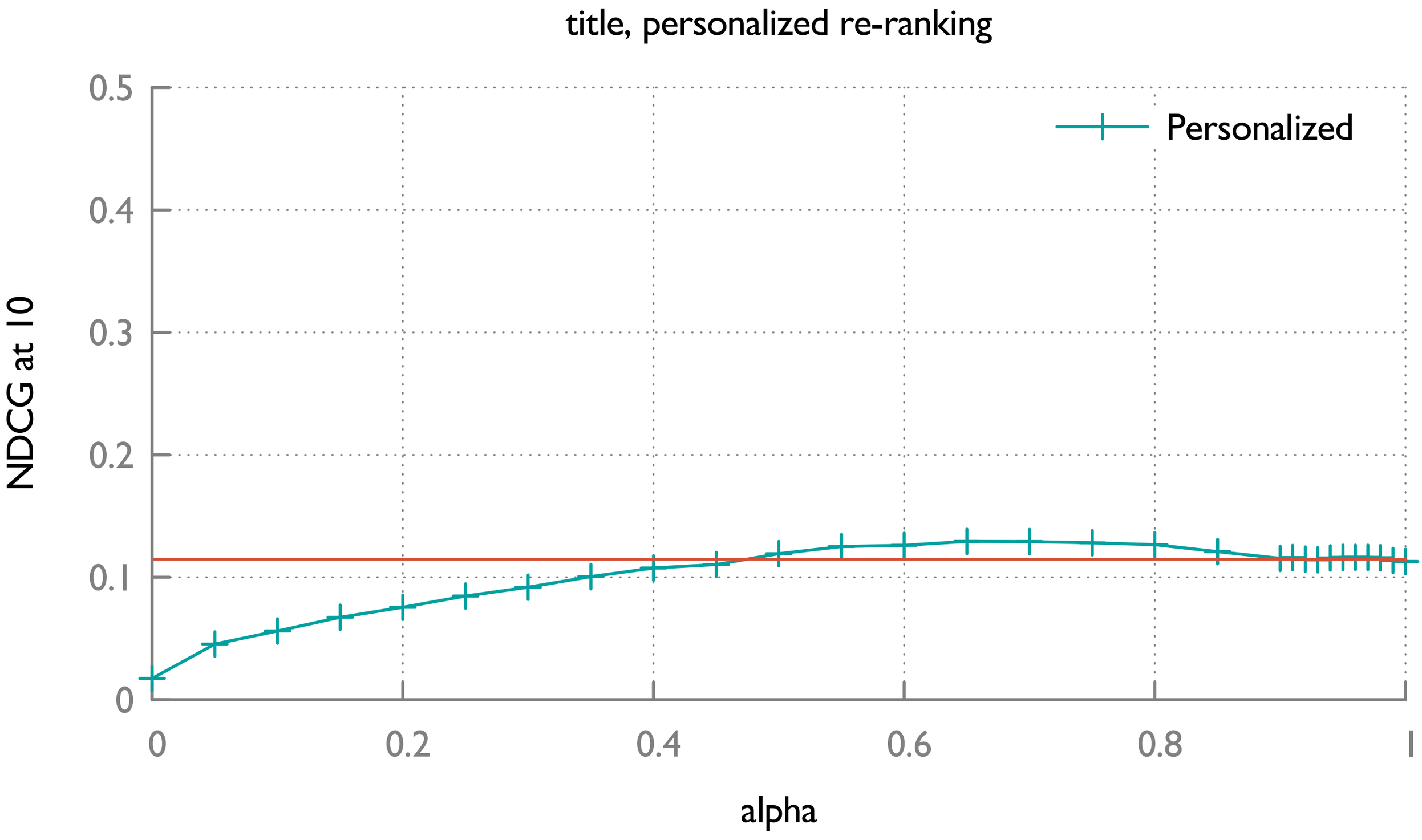
- Can we personalize the results list for each topic creator?
  - Take into account the past preferences of the topic creator
    - Books similar to past reads are pushed upwards
  - Similarity based on Jaccard overlap between tags in user *u*'s library and book *i*, controlled by α

$$score_{personalized}(u,i) = \alpha \cdot score_{org}(i) + (1-\alpha) \cdot sim_{tag}(u,i)$$

Results

Runs	Topic fields			
	title		all-topic-fields	
	NDCG@10	α	NDCG@10	α
Baseline	0.1129	-	**0.3058**	-
IU-similarity	0.1631	0.92	0.3058	1.0
II-similarity	**0.2429**	0.94	0.3058	1.0
IT-similarity	0.1895	0.99	0.3058	1.0
IA-similarity	0.1535	0.96	0.3058	1.0
IUTA-similarity	0.1615	0.97	0.3058	1.0
pers-similarity	0.1293	0.65	0.3058	1.0


title, book similarity re-ranking
NDCG at 10
0.5
0.4
0.3
0.2
0.1
0
0.8
0.82
0.84
0.86
0.88
0.9
0.92
0.94
0.96
0.98
1
alpha
II
IU
IA
IT
IUTA


title, personalized re-ranking
Personalized
NDCG at 10
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
alpha

Discussion

What did we learn?

- Best performance when combining all available information
  - Support for principle of polyrepresentation
  - Best submitted run (NCDG@10)
- Social re-ranking
  - Works great on short, Web-search-like queries
  - Does not work at all on longer queries

Future work?

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...?

Questions?