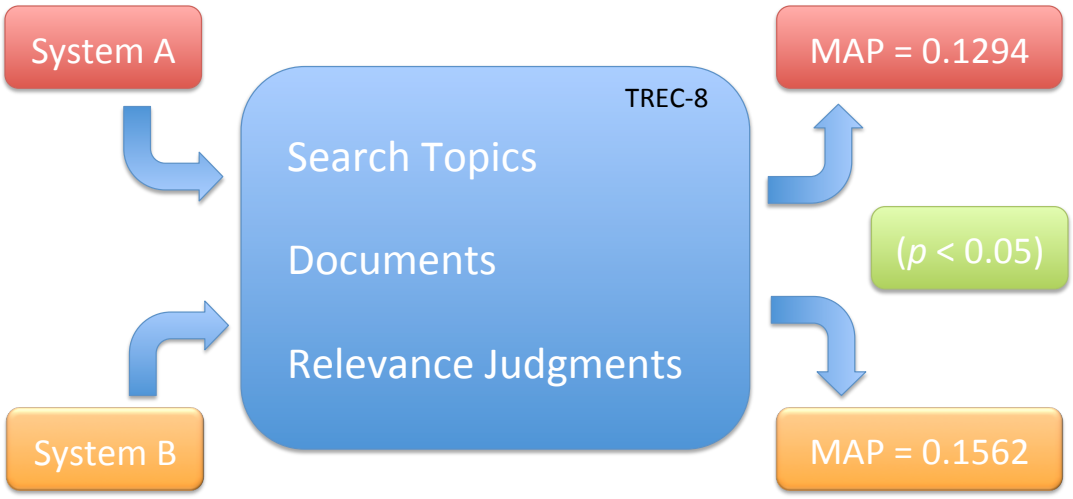
Differences in Effectiveness Across Sub-collections

Mark Sanderson
Andrew Turpin
Ying Zhang
Falk Scholer

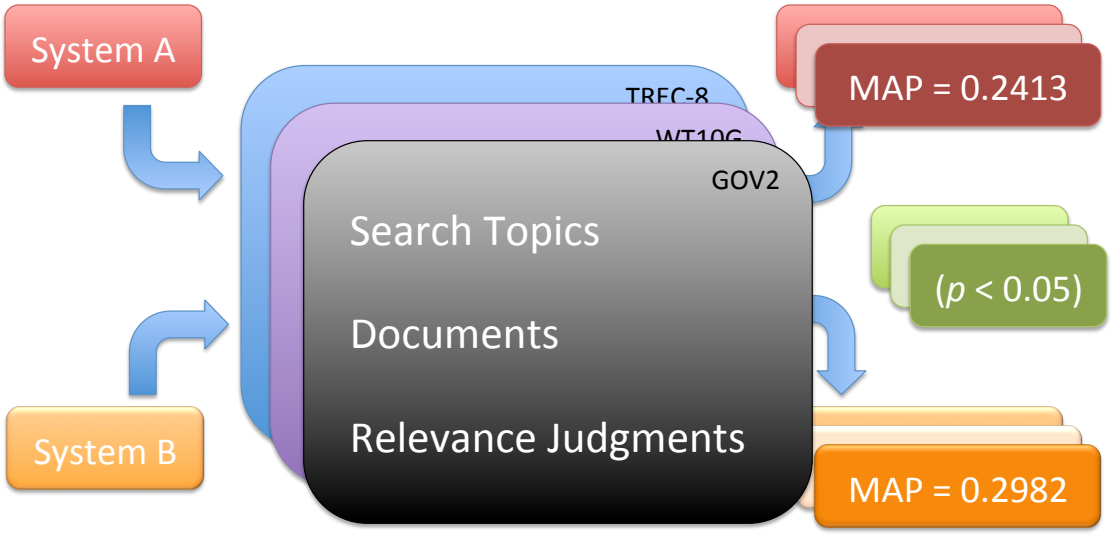
RMIT UNIVERSITY

THE UNIVERSITY OF MELBOURNE

Test Collection-based Evaluation

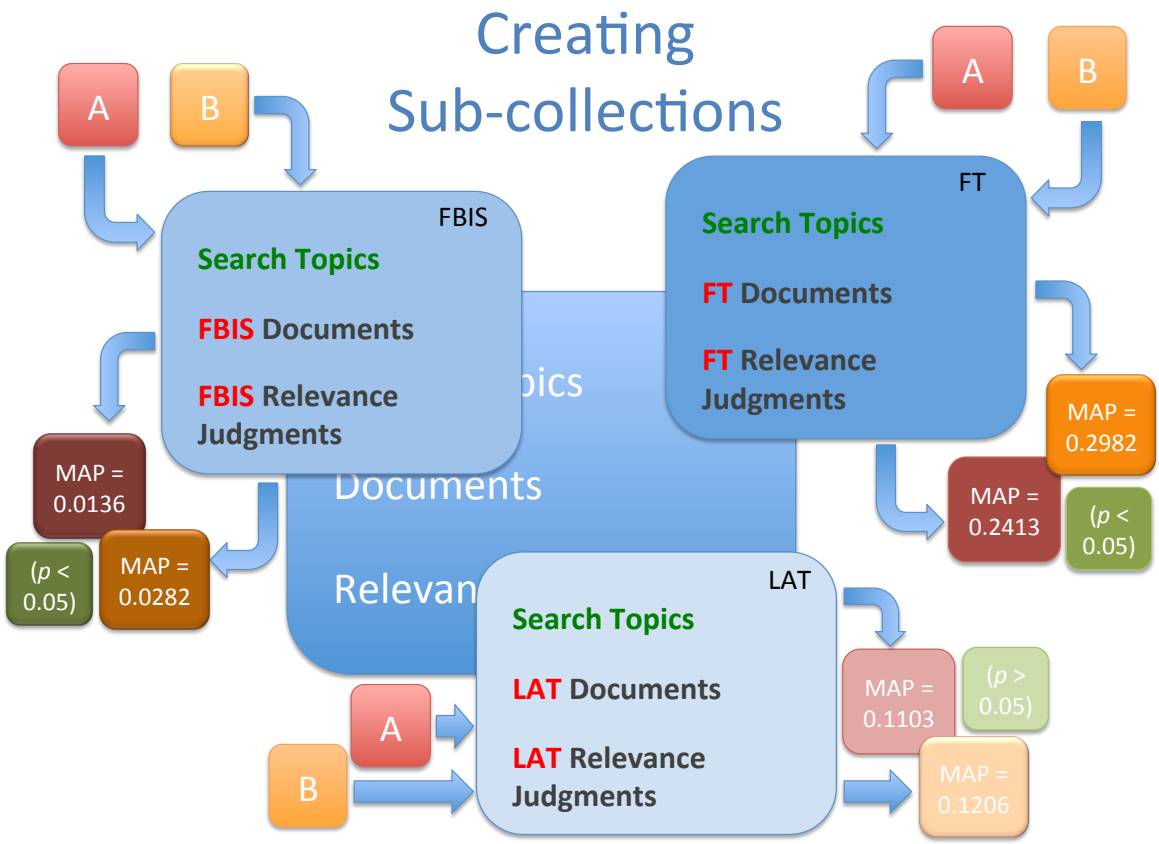


Still Skeptical?



Testing on Multiple Collections

- Helps to ensure that ranking functions are robust
- Shows performance in different environments
- Is often done in IR research by unspoken convention
- Does evaluating over different collections lead to substantially different outcomes?
- If so, what causes the differences?

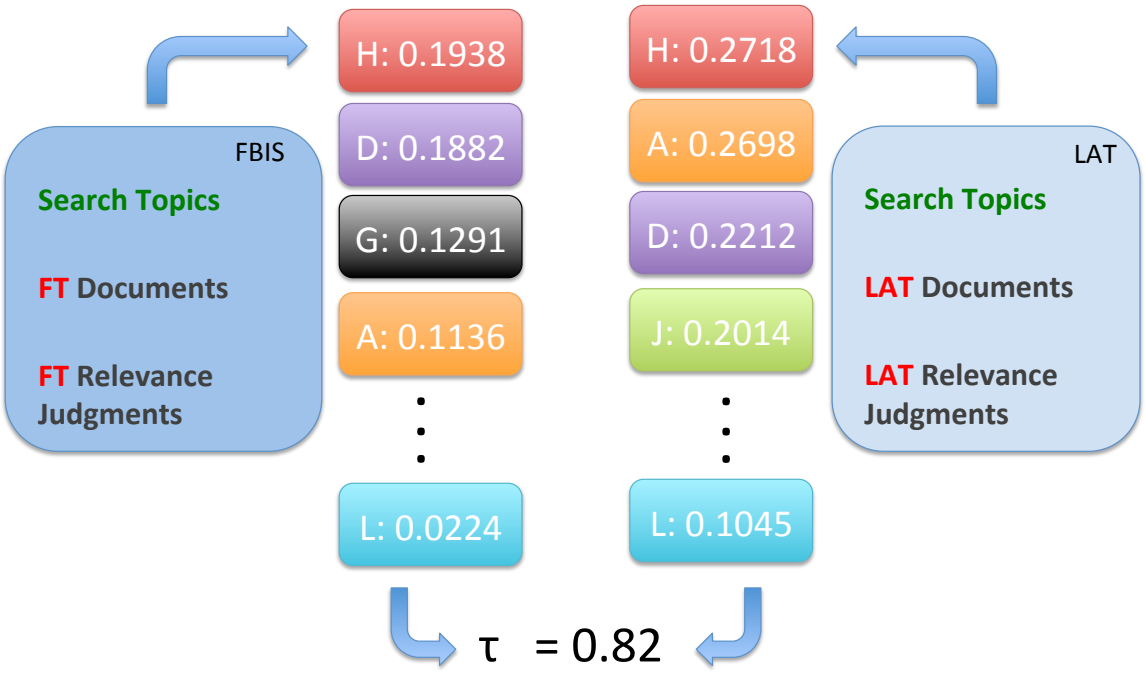


Sub-collection Splitting

Sub-collections were created using various criteria:

- Publication source (TREC 2—8)
  - FT, LAT, FBIS, …
- Top-level domain (GOV2)
  - .gov and .us domains
- MIME type (GOV2)
  - text/html, application/pdf

Comparing Sub-collections



Measuring Run Rankings

- If there was no effect from splitting, relative system orders would be consistent on each SC
  - Since qrels are different, can't expect perfect agreement
- To understand the impact of this noise, *randomly* split a collection into SCs
  - Doing this many times gives a distribution of τ values

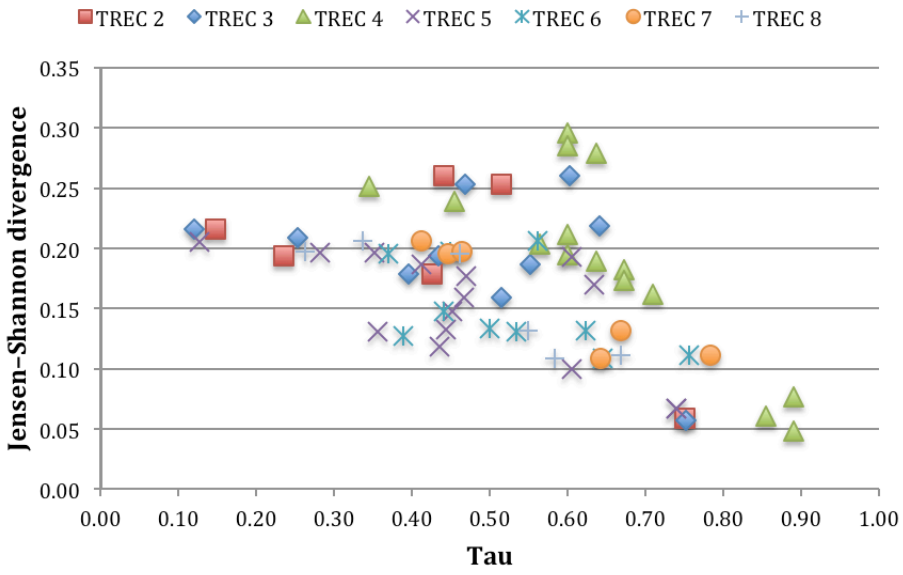
Comparisons for TREC-8

Pair	τ	Random range from	Random range to	p-value
FBIS-FR	0.46	0.51	0.85	< 0.001
FBIS-FT	0.58	0.65	0.87	< 0.001
FBIS-LA	0.55	0.61	0.85	< 0.001
FR-FT	0.26	0.55	0.85	< 0.001
FR-LA	0.34	0.53	0.84	< 0.001

Reasons for the Effect

- When two sub-collections have properties that are "similar", would expect agreement between system orderings to be higher
- Can measure correlations between collection properties, and consistency of system orderings

Language Similarity



R = -0.509 ($p < 0.0001$)

Other Reasons?

- Document length
  - Not significantly correlated: r=-0.172, p>0.1
- Length of relevant documents
  - Significantly correlated: r=-0.406, p<0.001
- Number of relevant documents
  - Not significantly correlated: r=-0.090, p=0.464

Conclusions

- Generic ranking functions don't search consistently over SCs
  - Relative system effectiveness differs substantially on different SCs
- Certain properties of SCs seem to be related to the level of divergence of ranking behavior
  - Language use
  - Length of relevant documents

Future Work

- What is the impact of other partitioning approaches?
- Can IR effectiveness be improved through purposeful partitioning?

Questions?