Stability of INEX 2007 Evaluation Measures

Sukomal Pal

Mandar Mitra

Arnab Chakraborty

{sukomal_r, mandar}@isical.ac.in

arnabc@stanfordalumni.org

Information Retrieval Lab, CVPR Unit

Indian Statistical Institute

Kolkata - 700108, India.

Outline

- Introduction

Outline

- Introduction
- Test Environment

Outline

- Introduction
- Test Environment
- Experiments & Results

Outline

- Introduction
- Test Environment
- Experiments & Results
- Limitations & Future Work

Outline

- Introduction
- Test Environment
- Experiments & Results
- Limitations & Future Work
- Conclusion

Outline

- Introduction
- Test Environment
- Experiments & Results
- Limitations & Future Work
- Conclusion

Introduction: Content-oriented XML retrieval

- a new domain in IR
- XML as standard document format in web & DL
- growth in XML information repositories
- increase in XML-IR systems
- Two aspects of XML-IR systems
  - content (*text/image/music/video info*)
  - structure (*info about the tags*)

Introduction: Content-oriented XML retrieval

- from whole document → document-part retrieval
- new evaluation framework (*corpus, topic, rel-judged data, metrics*)needed
- Initiative for the Evaluation of XML retrieval, INEX ('02 - ..)
- our stability study on metrics of INEX 07 adhoc focused task

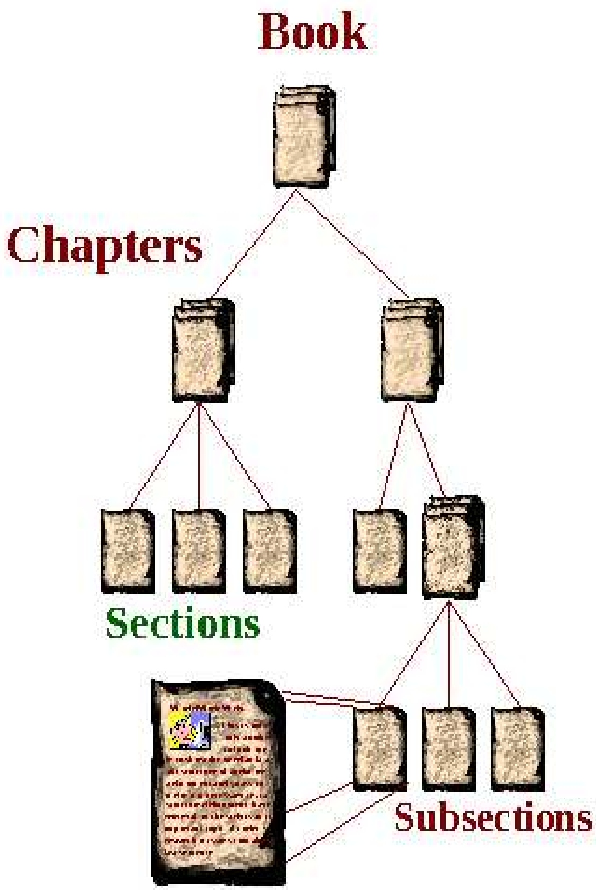


Figure 1: A book example

Introduction: Content-oriented XML retrieval

- from whole document → document-part retrieval
- new evaluation framework (*corpus, topic, rel-judged data, metrics*)needed
- Initiative for the Evaluation of XML retrieval, INEX ('02 - ..)
- our stability study on metrics of INEX 07 adhoc focused task

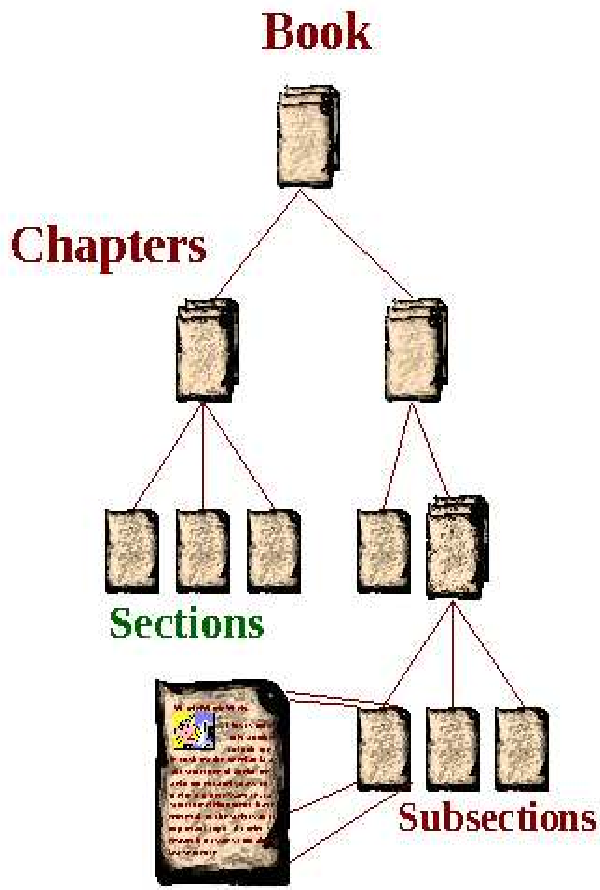


Figure 2: A book example

Outline

- Introduction
- **Test Environment**
- Experiments & Results
- Limitations & Future Work
- Conclusion

Test Environment: Collection

- XML-ified version of English Wikipedia
  - 659,388 documents
  - 4.6 GB
- INEX 2007 topic set
  - 130 queries (*414 - 543*)
- Relevance Judgment
  - 107 queries
- Runs
  - 79 valid runs (*ranked list acc. to relevance-score*)
  - max. 1500 passages/elements per topic

Test Environment: Measures

- Precision

$$\text{precision} = \frac{\text{amount of relevant text retrieved}}{\text{total amount of } \textit{retrieved} \text{ text}} = \frac{\text{length of relevant text retrieved}}{\text{total length of } \textit{retrieved} \text{ text}}$$

- Recall

$$\text{recall} = \frac{\text{length of relevant text retrieved}}{\text{total length of } \textit{relevant} \text{ text}}$$

Test Environment: Measures

$$\begin{aligned} p_r &= \text{document part at rank } r \\ size(p_r) &= \text{total \#characters in } p_r \\ rsize(p_r) &= \text{length of relevant text in } p_r \\ Trel(q) &= \text{total amt of relevant text for topic } q \end{aligned}$$

- Precision at rank r

$$P[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{\sum_{i=1}^{r} size(p_i)}$$

- Recall at rank r

$$R[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{Trel(q)}$$

Test Environment: Measures

- Drawback
  - rank not well-understandable for passages/elements (retrieval granularity not fixed)
  - recall level used instead
- Interpolated Precision at recall level x

$$iP[x] = \begin{cases} \max_{\substack{1 \leq r \leq |L_q| \\ R[r] \geq x}} (P[r]) & \text{if } x \leq R[|L_q|] \\ 0 & \text{if } x > R[|L_q|] \end{cases}$$

(L_q = set of ranked list, $|L_q| \leq 1500$)

e.g.

$$\begin{aligned} iP[0.00] &= \text{int. prec. for first unit retrieved} \\ iP[0.01] &= \text{int. prec. at 1\% recall for a topic} \end{aligned}$$

Test Environment: Measures

- Average interpolated precision for topic t

$$AiP(t) = \frac{1}{101} \sum_{x=\{0.00,0.01,\ldots,1.00\}} iP[x](t)$$

- overall int. precision at reall level x

$$iP[x]_{overall} = \frac{1}{n} \sum_{t=1}^{n} iP[x](t)$$

- Mean Average Interpolated Precision

$$MAiP = \frac{1}{n} \sum_{t=1}^{n} AiP(t).$$

- Reported metrics for INEX 2007 Adhoc focused task
  - $iP[0.00]$, $iP[0.01]$, $iP[0.05]$, $iP[0.10]$ & $MAiP$
  - **official** metric : $iP[0.01]$

Test Environment: Experimental setup

- relevance judgment
  - NOT just *boolean* indicator
  - relevant psg. with start & end-offset in *xpath*
- db of start & end offsets for each element of entire corpus
  - size $\sim$ 14 GB
- a subset of db, representing rel-jdg file, stored
- Out of 79 runs, 62 chosen
  - taken runs ranked 1-21, 31-50, 59-79 acc. to $iP[0.01]$
  - run file consulted with db to get offsets, compared with stored rel-jdg file

Outline

- Introduction
- Test Environment
- **Experiments & Results**
- Limitations & Future Work
- Conclusion

Experiments

3 categories:

- Pool Sampling
  - evaluate using incomplete relevance judgments
  - some rel. passages made irrel. for each topic
- Query Sampling
  - evaluate using smaller subsets of topics
  - complete rel-jdg info for a topic, if selected
- Error Rate
  - offshoot of *query sampling*
  - study of pairwise runs with topic set reduced

Experiments: Pool Sampling

Pool

- generated from the participants' runs
- collaboratively judged by participants
  - relevant passages highlighted
  - no highlighting $\Longrightarrow$ NOT relevant

Qrel

- start and end-points of highlighted passages by *xpath*
- consulted db to get the offsets, stored in a sorted file
- No entries for assessed non-relevant text
- contained 107 topics

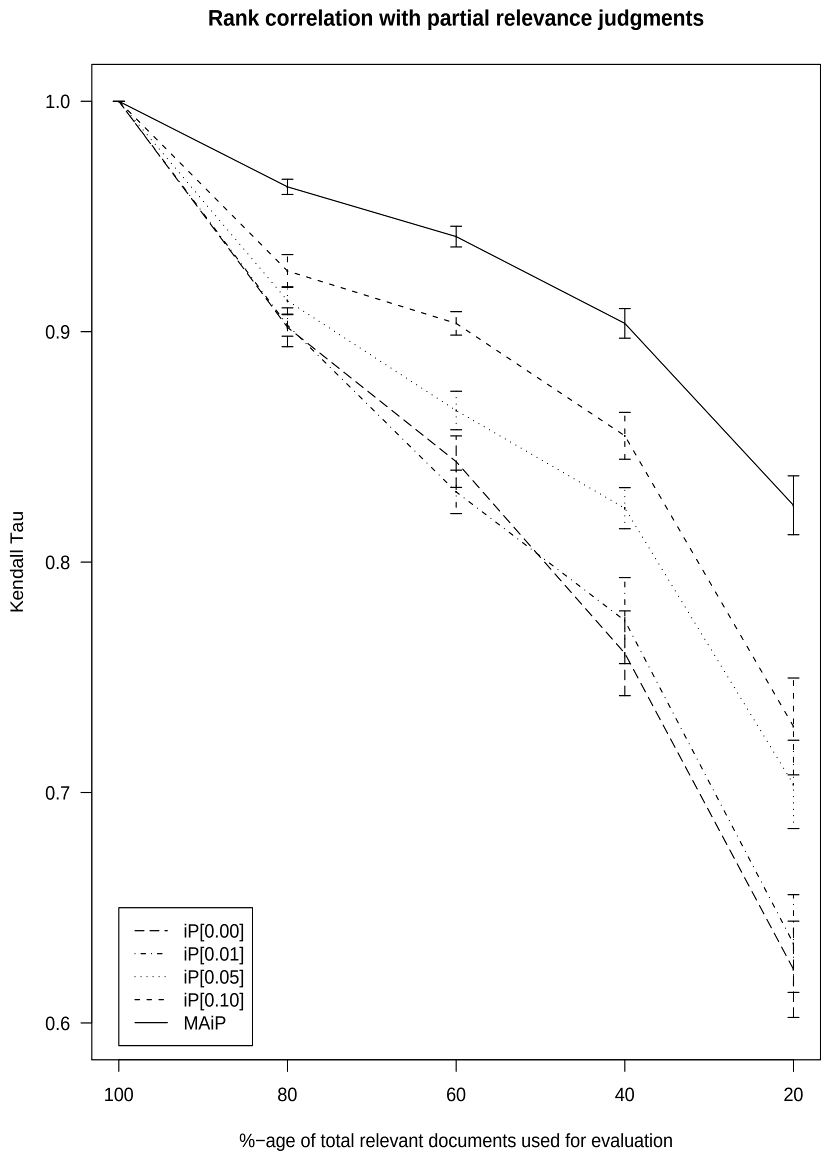
Experiments: Pool Sampling

Alogrithm:

1. 99 topics having $>= 10$ relevant units selected
2. 80% relevant passages SRSWOR for each topic $\rightarrow$ new $qrel$
3. 62 runs evaluated with reduced sample $qrel$
4. Kendall tau (τ) computed betn. 2 rankings for each metric (*i.e. ranking by original* $qrel$ *and reduced* $qrel$)
5. 10-iterations of the above steps 1-4 at 80%-sample

- Steps 1-5 done at 60%, 40%, 20% samples

Results: Pool Sampling



Results: Pool Sampling

- sampling level $\downarrow \rightarrow$ correlation $\downarrow \rightarrow$ curve droops
- precision-score affected non-uniformly across systems
  - *depending upon ranks of retrieved text missing in pool*
- τ drops for $iP[0.00], iP[0.01]$ faster than $iP[0.05]$ or $iP[0.10]$ or $MAiP$
- sampling level $\downarrow \rightarrow$ error-bar $\uparrow$
- sampling level $\downarrow \rightarrow$ overlap among the samples at a fixed $n\% \downarrow \rightarrow$ irregular prec-score

$MAiP$ - least variation in τ

- across different pool-sizes
- across samples at a fixed pool-size

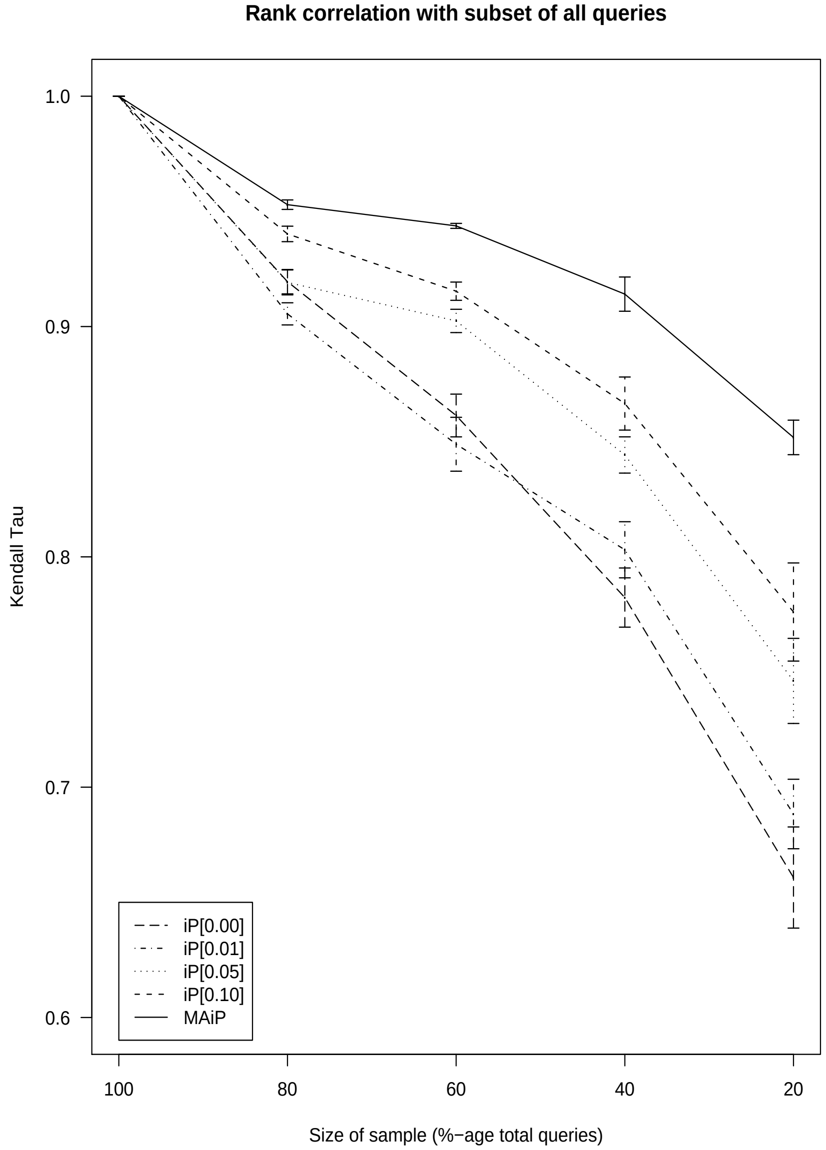
Experiments: Query Sampling

Algorithm:

1. All 107 topics considered
2. 80% of total topics selected at random (SRSWOR)
3. if a topic selected, its entire rel-jdg taken $\rightarrow$ new reduced $qrel$
4. 62 runs evaluated with reduced sample $qrel$
5. Kendall tau (τ) computed betn. 2 rankings for each metric (*i.e. ranking by original* $qrel$ *and reduced* $qrel$)
6. 10-iterations of the above steps 1-4 at 80%-sample

- Steps 1-5 done at 60%, 40%, 20% samples

Results: Query Sampling



Results: Query Sampling

- Similar characteristic comp. to *Pool Sampling*
  τ drops for $iP[0.00], iP[0.01]$ faster than $iP[0.05]$ or $iP[0.10]$ or $MAiP$
  sampling level $\downarrow \rightarrow$ error-bar $\uparrow$

$MAiP$ - best as it has least variation in τ

- across different pool-sizes
- across samples at a fixed pool-size

Curves are more stable than those in *Pool Sampling* (i.e. *system rankings more in agreement with original rankings*)

- if a topic selected, its entire rel-jdgmnt used
- the topic contributes to prec. score uniformly across systems
- τ reduces due to different response of systems to a query

Experiments: Error Rate

Algorithm:

1. Acc. to Buckley & Voorhees 2000 but with modification
   - participants' systems not available
   - results of systems under varying query formulations NOT possible
2. Samples of Query-set with full $qrel$ per topic
   - partitioning of the query-set(SRSWOR) $\rightarrow$ *upper bound of error-rate*
   - subsets of query-set(SRSWR) $\rightarrow$ *lower bound error-rate*
3. 10 samples (SRSWR) at 20%, 40%, 60%, 80% of 107 queries

Experiments: Error Rate

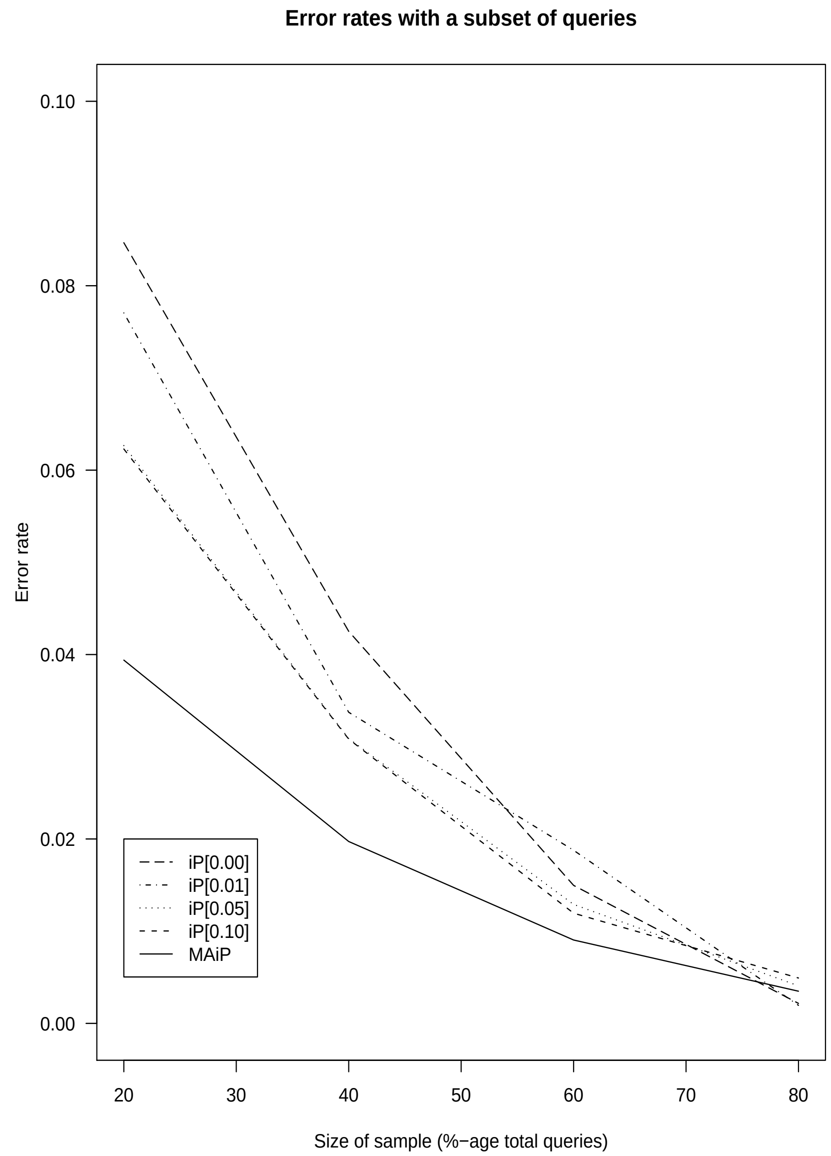
- Error Rate (Buckley et al. '00)

$$\text{Error rate} = \frac{\sum \min(|A > B|, |A < B|)}{\sum(|A > B| + |A < B| + |A == B|)}$$

$|A > B|$ = #times (out of 10) system A better B at a fixed sampling level. Note, $A > B$ by $\geq$ 5%, else $A == B$.

- 62 systems, $\binom{62}{2} = 62.61/2 = 1891$ pairs

Results: Error Rate



Results: Error Rate

Error-rates

- high for small query-sets
- progressively $\downarrow$ as overlap among query samples $\uparrow$
- 40% topics sufficient to achieve less than 5% error
- early-prec. measures more error-prone
- $MAiP$ has least error-rate

$MAiP$ - best as it has least variation in τ

Outline

- Introduction
- Previous Work
- Test Environment
- Experiments & Results
- **Limitations & Future Work**
- Conclusion

Limitation & Future Work

- Observations based only on INEX 2007 test collection
- Not all (79 valid) runs, could consider 62 of them
- Runs from non-random influencing categories
  - passage/element, CO/CAS, short/long, hard/easy queries etc.
- No knowledge of top-n retrieved units used to create pool
  - future task
- Bias of $qrels$ towards participating runs
  - future task
- Error-rates - No idea why steady nature was disturbed
- We considered 5% error rate
  Lot more study needed

$MAiP$

- averages well across topics
- more shock-absorbing than other metrics

Outline

- Introduction
- Previous Work
- Test Environment
- Experiments & Results
- Limitations & Future Work
- **Conclusion**

Conclusion

- XML retrieval evaluation gruelling challenge
- Various metrics tried since INEX ’02 to ’06
- prec-recall based metrics since INEX ’07
- validation of previous findings in XML retrieval domain
- similar results $\longrightarrow$ intrinsic properties of metrics

$MAiP$

- averages well across topics
- more shock-absorbing than other metrics
- most reliable metric for static test environement

Acknowledgments

- work : DIT, Govt. of India
- trip : NTCIR, Japan & Google Inc., USA.

!! THANK YOU !!