

# Structural Relevance in XML Retrieval Evaluation

M. S. Ali<sup>1</sup>, Mariano P. Consens<sup>1</sup>, Mounia Lalmas<sup>2</sup>

<sup>1</sup> University of Toronto, Canada

<sup>2</sup> Queen Mary, University of London, UK

**3H Workshop on Focused Retrieval**

July 27, 2007



UNIVERSITY of TORONTO

## Outline

- Motivation and Approach
- Structural Relevance
- Results and Comparison
- Conclusions and Future Work



UNIVERSITY of TORONTO

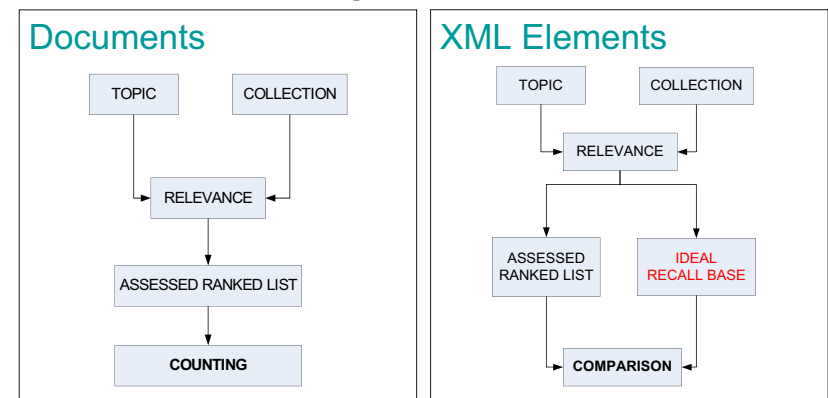
## Motivation and Approach

- Current approaches to evaluation in XML retrieval rely on ideal recall base
- How to evaluate without defining an ideal recall base?
- Our Approach: Differentiate between the relevance of a retrieval element in isolation and the relevance of a retrieval element as a member of a set (i.e. a ranked list) of non-disjoint elements using structure of elements in collection



UNIVERSITY of TORONTO

## Measuring Effectiveness



Precision, Precall<sup>1</sup>

<sup>1</sup>Raghavan (1989)

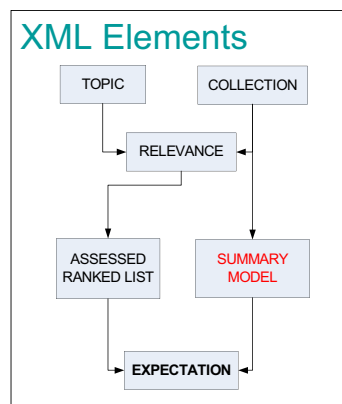
XCG<sup>2</sup>, PRUM<sup>3</sup>, EPRUM<sup>4</sup>

<sup>2</sup>Kazai (2006), <sup>3</sup>Piwarwaski (2007), <sup>4</sup>Piwarwaski (2006).



UNIVERSITY of TORONTO

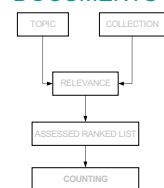
# Proposal: Structural Relevance



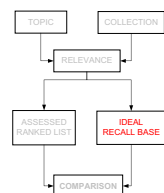
SRP<sup>1</sup>, SRPL<sup>1</sup>

<sup>1</sup> see this paper for these definitions

## DOCUMENTS



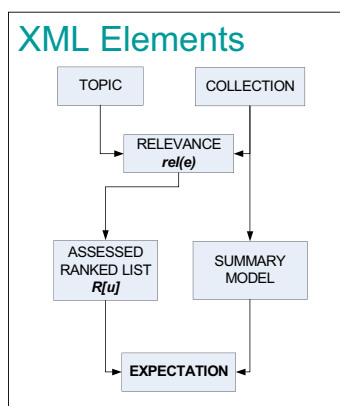
## XML Elements



# Outline

- Motivation and Approach
- **Structural Relevance**
- Results and Comparison
- Conclusions and Future Work

## Relevance



Allow binary, multi-graded or continuous relevance scores  $rel(e)$ .

$R[u]$  is the ranked list up to element  $u$ .

## Expectation

Structural relevance (SR) is an expectation of the number of relevant elements in a ranked list.

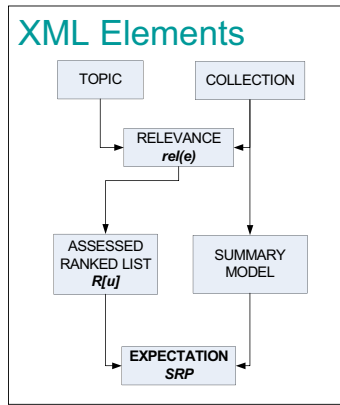
$$E[n_R(u)] = \sum_{e \in R[u]} rel(e) \cdot p(e; R[u])$$

where

$n_R(u)$  is the number of relevant elements up to element  $u$ .

$p(e; R[u])$  is the probability of encountering  $e$  first from the ranked list  $R[u]$ , as opposed to, a different, overlapped element in the list. We call this the *isolation* of  $e$  in ranked list  $R[u]$ .

# Expectation



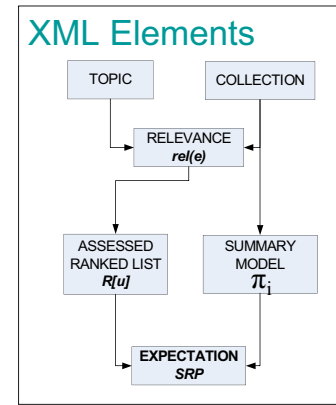
Substitute  $SR$  into a traditional measure for the number of relevant elements.

For precision we get,

$$SRP = E[n_R]/k$$

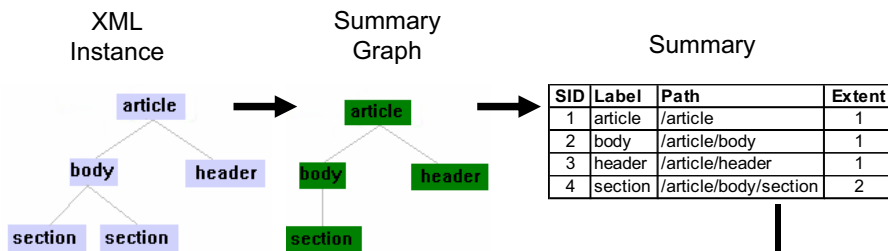
Similarly, this can be done for precall (SRPL).

# Summary Model



In the paper, we show how isolation  $p(e;R[u])$  can be calculated in terms of steady-state probabilities  $\pi_i$  derived from a given summary model.

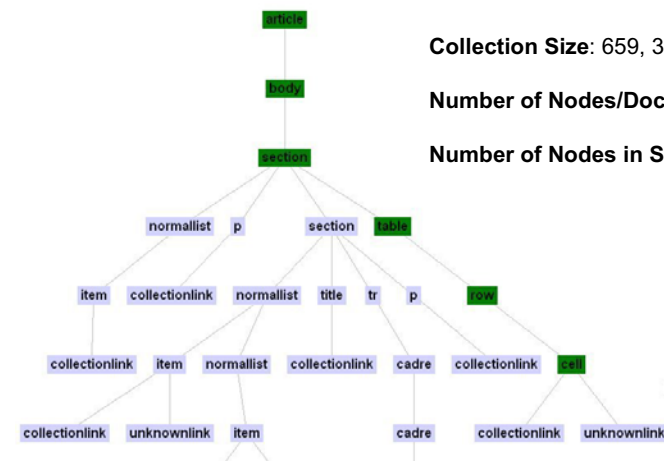
# Summary Model of XML



Weighting Matrix and Steady-state Probabilities

	SID j					
SID i	1	2	3	4	TOTAL	$\pi_i$
1	1	1	1	2	5	0.294
2	1	1	0	2	4	0.235
3	1	0	1	0	2	0.118
4	2	2	0	2	6	0.353

# Incoming Summary for INEX WIKIPEDIA Collection



Collection Size: 659, 388 (English version)

Number of Nodes/Document: 161

Number of Nodes in Summary: 240,000+

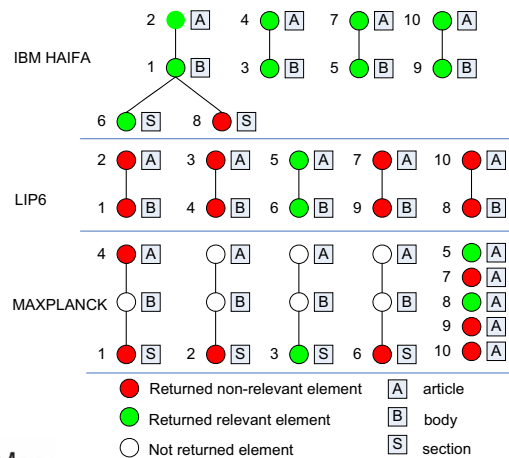
# Outline

- Motivation and Approach
- Structural Relevance
- Results and Comparison
- Conclusions and Future Work

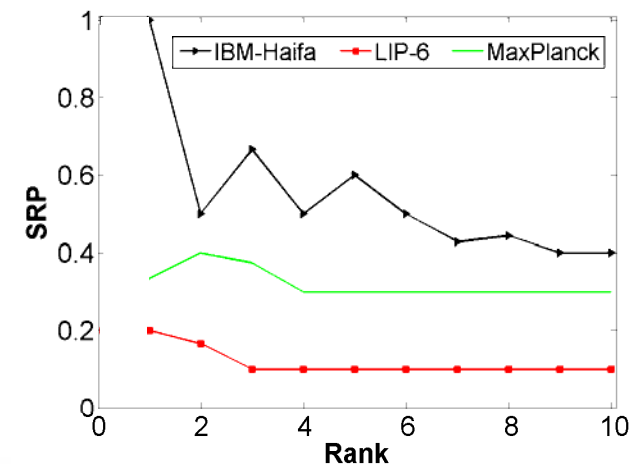
# Comparison

- To compare the evaluation of SR precision (SRP) to extended cumulated gain (XCG)
- Used INEX Wikipedia 2006 topics
- Ad-hoc retrieval for the thorough task
- Compared systems using top-10 results

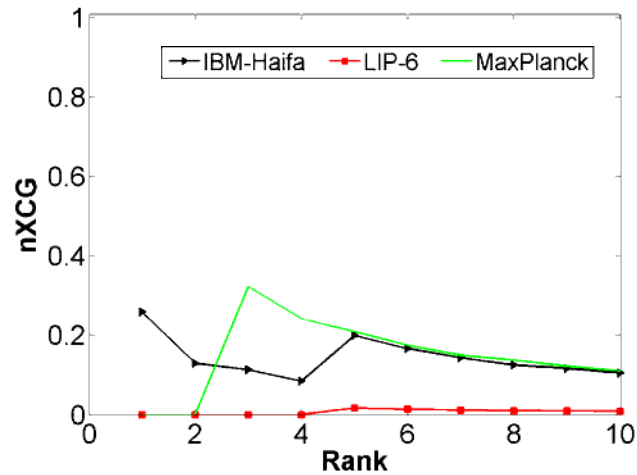
## Top-10 for INEX Topic 295



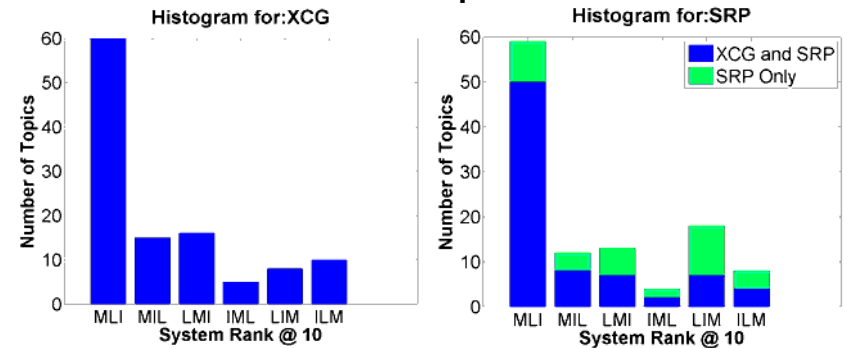
SR Precision (SRP):  
Topic 295 (WIKIPEDIA Top-10)



### Extended Cumulated Gain (XCG): INEX Topic 295 (WIKIPEDIA Top-10)



### Comparing SRP and XCG Across All Topics



(I) IBM HAIFA, (M) Maxplanck, (L) LIP6

Results differ between XCG and SRP because of heuristics in XCG and lack of differentiation in XCG between early- and late-recall.

## Outline

- Motivation and Approach
- Structural Relevance
- Results and Comparison
- Conclusions and Future Work

## Conclusion

- Structural relevance measures effectiveness without an ideal recall-base
  - Motivated by results that show sensitivity to ideal recall-base determination, Kazai (2007)
- SR measure applied to thorough task here, but it can be applied to other tasks (eg, focused task, tasks where overlap is allowed)
- SR can be used with other evaluation measures (eg, using incomplete assessments)

## Future Work

- Stability and Reliability tests
- Further comparison to other measures
- Investigating additional summary models

END OF PRESENTATION

## References

- **Kazai (2006)**, Extended cumulated gain measures for the evaluation of content-oriented xml retrieval. *ACM Trans. Inf. Syst.*, 24(4):503–542.
- **Kazai (2007)**, Choosing an ideal recall-base for the evaluation of the focused task: Sensitivity analysis of the xcg evaluation measures. *INEX 2007*. Springer.
- **Piwarwaski (2007)**, Precision recall with user modeling (prum): Application to structured information retrieval. *ACM Trans. Inf. Syst.*, 25(1):1
- **Piwarwaski (2006)**, Expected precision-recall with user modelling (eprum). In *SIGIR '06: Proc. 29th Ann. Intl ACM SIGIR Conf on R&D in Info Retr.*, pages 260–267, New York, NY, USA. ACM Press.
- **Pehcevski (2005)**, HiXEval: Highlighting XML retrieval evaluation. *Advances in XML Information Retrieval and Evaluation*. *INEX 2005*. Springer-Verlag.
- **Raghavan (1989)**, A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, 7(3):205–229.