

The Accessibility Dimension for Structured Document Retrieval

Thomas Rölleke, Mounia Lalmas and Gabriella Kazai
Queen Mary University of London

Ian Ruthven
University of Strathclyde

Stefan Quicker
University of Dortmund

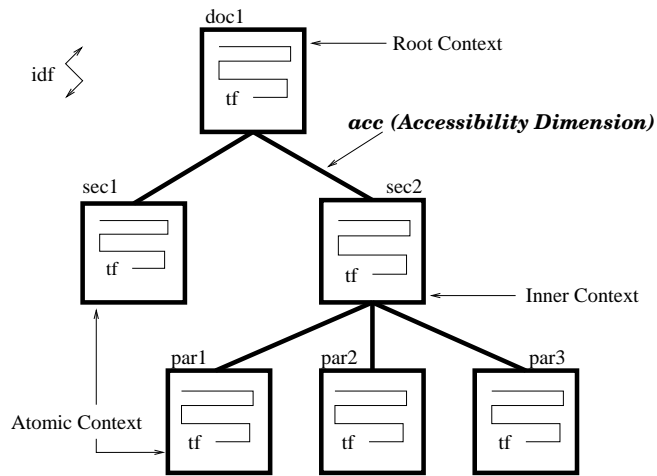
Slide 1

Outline

1. The Accessibility Dimension
2. The Theory
3. Implementation: Probabilistic Relational Algebra (PRA)
4. Implementation: The Effect of *acc*
5. Experiments
6. Result: Accessibility Dimension and Average Precision
7. Result: Accessibility Dimension and Retrieved Contexts
8. Summary and Conclusion
9. References

Slide 2

The Accessibility Dimension



Slide 3

The Theory

term frequency	<i>tf</i>
inverse document frequency	<i>idf</i>
accessibility	<i>acc</i>
retrieval status value	<i>RSV</i>

Slide 4

$$RSV(d, q) := \sum_{t \in q} tf(t, q) \cdot tf(t, d) \cdot idf(t)$$

$$RSV(d, q) := \sum_{t \in q} tf(t, q) \cdot tf^*(t, d) \cdot P_{idf}(t)$$

$$tf^*(t, d) := (1 - \prod_{s \in d} (1 - acc(d, s) \cdot tf(t, s)))$$

Implementation: Probabilistic Relational Algebra (PRA)

Slide 5

qterm (query rep)
(sailing)

term (document rep, based on <i>tf</i>)
0.1 (sailing, doc1)
0.7 (sailing, sec1)

termspace (based on <i>idf</i>)
0.4 (sailing)

acc (structure)
0.8 (doc1, sec1)

$$\text{aggregated_term_depth1} = \text{JOIN}[\$2=\$2](\text{term}, \text{acc})$$

$$0.56 = 0.7 \cdot 0.8 \text{ aggregated_term_depth1}(\text{sailing}, \text{doc1})$$

$$\text{aggregated_term} = \text{UNITE}(\text{term}, \text{aggregated_term_depth1})$$

$$0.604 = 0.1 + 0.56 - 0.1 \cdot 0.56 \text{ aggregated_term}(\text{sailing}, \text{doc1})$$

Implementation: The Effect of acc

all documents about sailing

$$\text{wqterm} = \text{JOIN}[\$1=\$1](\text{qterm}, \text{termspace})$$

$$\text{retrieved_tfidf} = \text{PROJ}[\$3](\text{JOIN}[\$1=\$1](\text{wqterm}, \text{term}))$$

$$\text{retrieved_tfidfacc} = \text{PROJ}[\$3](\text{JOIN}[\$1=\$1](\text{wqterm}, \text{aggregated_term}))$$

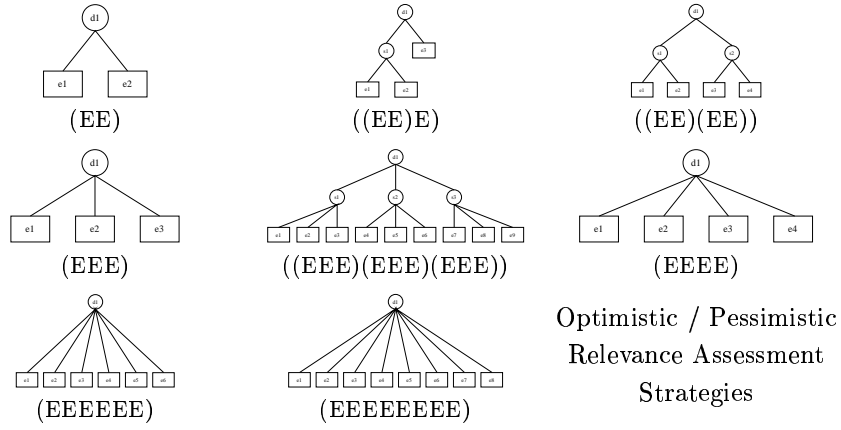
Slide 6

0.1 term(sailing, doc1), 0.7 term(sailing, sec1), 0.4 termspace(sailing)

no acc	0.04 = 0.1 · 0.4	retrieved_tfidf(doc1)
	0.28 = 0.7 · 0.4	retrieved_tfidf(sec1)
acc = 0.8	0.2416 = 0.604 · 0.4	retrieved_tfidfacc(doc1)
	0.2800 = 0.7 · 0.4	retrieved_tfidfacc(sec1)
acc = 1.0	0.292 = 0.73 · 0.4	retrieved_tfidfacc(doc1)
	0.280 = 0.7 · 0.4	retrieved_tfidfacc(sec1)

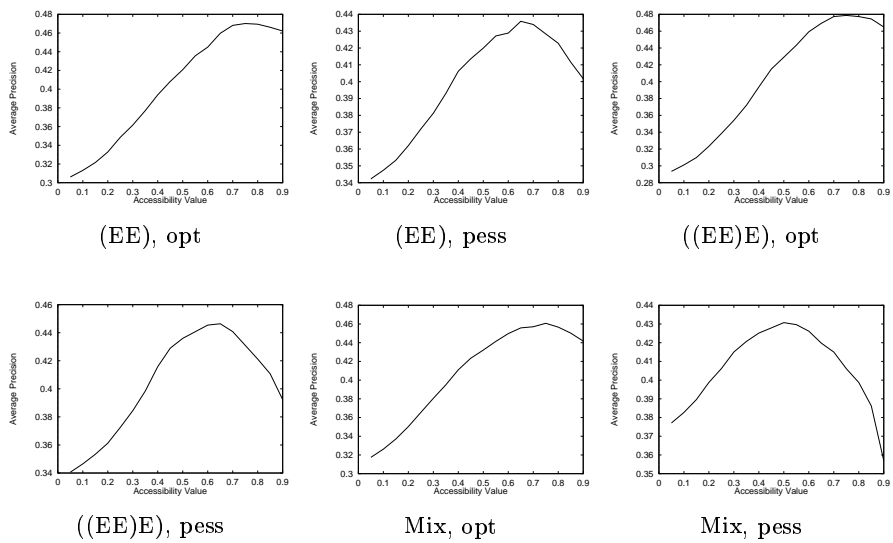
Experiments

Slide 7



Result: Accessibility Dimension and Average Precision

Slide 8



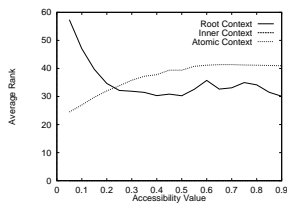
Result: Accessibility Dimension and Average Precision

collection	Optimistic relevance		Pessimistic relevance	
	max. av. precision	<i>acc</i>	max. av. precision	<i>acc</i>
(EE)	0.4702	0.75	0.4359	0.65
(EEE)	0.4719	0.6	0.4479	0.45
(EEEE)	0.455	0.55	0.4474	0.35
(EEEEEE)	0.4431	0.45	0.4507	0.25
(EEEEEEEE)	0.4277	0.35	0.4404	0.2
((EE)(EE))	0.4722	0.8	0.4556	0.6
((EE)E)	0.4787	0.75	0.4464	0.65
((EEE)(EE)(EEE))	0.4566	0.65	0.4694	0.4
Mix	0.4608	0.75	0.4307	0.5

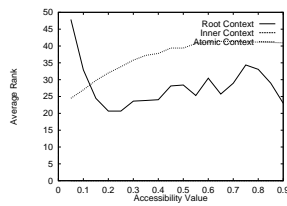
$$acc(d, s) = \frac{1}{\sqrt{n}}, \quad n = ||\{s | s \in d\}||$$

Slide 9

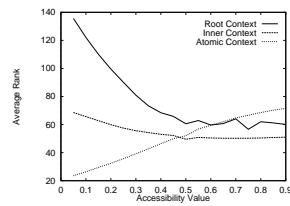
Result: Accessibility Dimension and Retrieved Contexts



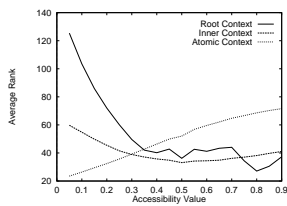
((EEEEEEEE), opt)



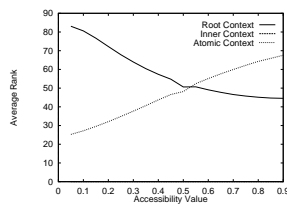
(EEEEEEEE), pess



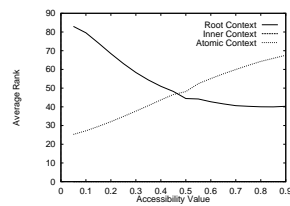
((EEE)(EE)(EEE), opt)



((EEE)(EE)(EEE), pess)



Mix, opt



Mix, pess

Slide 10

Summary and Conclusion

Slide 11

- Accessibility dimension: *tf-idf-acc* approach for structured document retrieval
 - *acc* depends on number of contexts and relevance assessment strategies for structured documents
 - *acc* controls retrieval of “higher” (root and inner) and “lower” (inner and atomic) contexts
- PRA (HySpirit framework) models parameter settings and aggregation strategies
- Automatic test collection building for structured document retrieval, optimistic and pessimistic relevance
- Name “accessibility” from Kripke structures

References

Slide 12

1. Mounia Lalmas and Thomas Rölleke. Four-valued Knowledge Augmentation for Structured Document Retrieval, 13th International Symposium on Methodologies for Intelligent Systems (ISMIS02), Lyon, France, June 2002.
2. Norbert Fuhr and Thomas Rölleke. Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems, ACM Transaction on Information Systems 14(1), 1997.
3. <http://qmir.dcs.qmul.ac.uk>
4. <http://www.hyspirit.com>