Slide 1

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

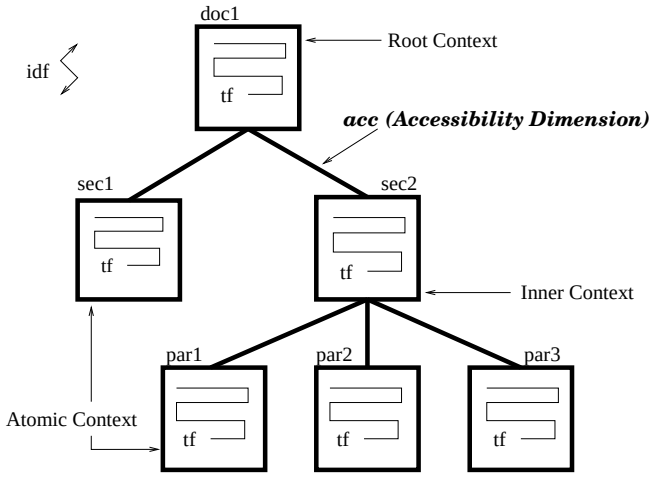
Slide 2

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 3

The Accessibility Dimension



Slide 4

The Theory

term frequency	tf
inverse document frequency	idf
accessibility	acc
restrieval status value	RSV

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

$$RSV(d,q) := \sum_{t \in q} tf(t,q) \cdot tf^*(t,d) \cdot P_{idf}(t)$$
$$tf^*(t,d) := (1 - \Pi_{s \in d}(1 - acc(d,s) \cdot tf(t,s)))$$

Slide 5

Implementation: Probabilistic Relational Algebra (PRA)

qterm (query rep)
(sailing)

term (document rep, based on *tf*)
0.1 (sailing, doc1)
0.7 (sailing, sec1)

termspace (based on *idf*)
0.4 (sailing)

acc (structure)
0.8 (doc1, sec1)

aggregated_term_depth1 = JOIN[$2=$2](term, acc)

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

aggregated_term = UNITE(term, aggregated_term_depth1)

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

Slide 6

Implementation: The Effect of acc

all documents about sailing

wqterm = JOIN[$1=$1](qterm, termspace)

retrieved_*tfidf* = PROJ[$3](JOIN[$1=$1](wqterm, term))

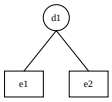
retrieved_*tfidfacc* = PROJ[$3](JOIN[$1=$1](wqterm, aggregated_term))

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

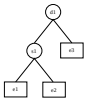
no acc	$0.04 = 0.1 \cdot 0.4$	retrieved_*tfidf*(doc1)
	$0.28 = 0.7 \cdot 0.4$	retrieved_*tfidf*(sec1)
acc = 0.8	$0.2416 = 0.604 \cdot 0.4$	retrieved_*tfidfacc*(doc1)
	$0.2800 = 0.7 \cdot 0.4$	retrieved_*tfidfacc*(sec1)
acc = 1.0	$0.292 = 0.73 \cdot 0.4$	retrieved_*tfidfacc*(doc1)
	$0.280 = 0.7 \cdot 0.4$	retrieved_*tfidfacc*(sec1)

Slide 7

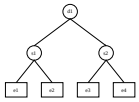
Experiments



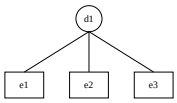
(EE)



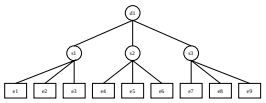
((EE)E)



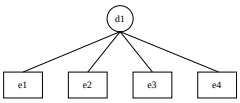
((EE)(EE))



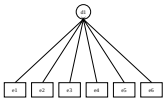
(EEE)



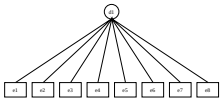
((EEE)(EEE)(EEE))



(EEEE)



(EEEEEE)

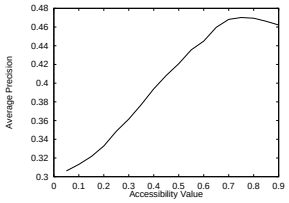


(EEEEEEEE)

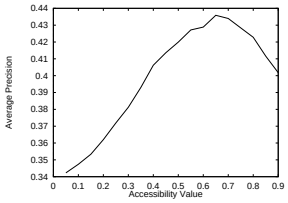
Optimistic / Pessimistic Relevance Assessment Strategies

Slide 8

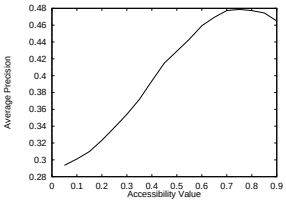
Result: Accessibility Dimension and Average Precision



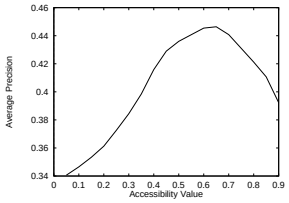
(EE), opt



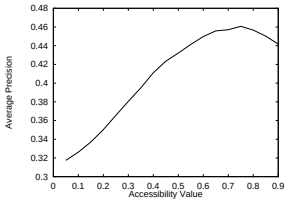
(EE), pess



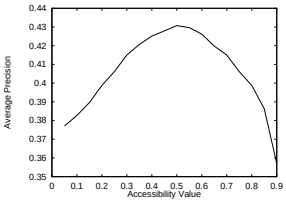
((EE)E), opt



((EE)E), pess



Mix, opt



Mix, pess

Slide 9

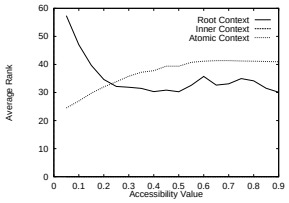
Result: Accessibilty Dimension and Average Precision

collection	Optimistic relevance max. av. precision	*acc*	Pessimistic relevance max. av. precision	*acc*
(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)(EEE)(EEE))	0.4566	0.65	0.4694	0.4
Mix	0.4608	0.75	0.4307	0.5

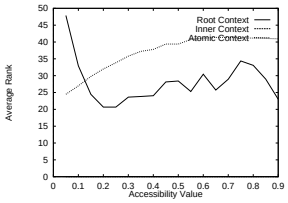
$$acc(d,s) = \frac{1}{\sqrt{n}}, \qquad n = ||\{s|s \in d\}||$$

Slide 10

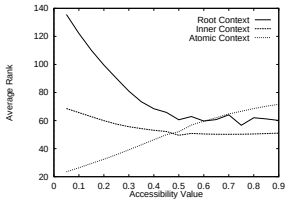
Result: Accessibility Dimension and Retrieved Contexts



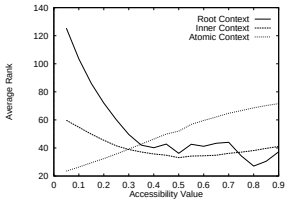
(EEEEEEEE), opt



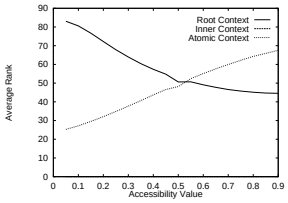
(EEEEEEEE), pess



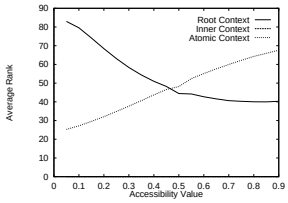
((EEE)(EEE)(EEE)), opt



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



Mix, opt



Mix, pess

Slide 11

Summary and Conclusion

- 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

Slide 12

References

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