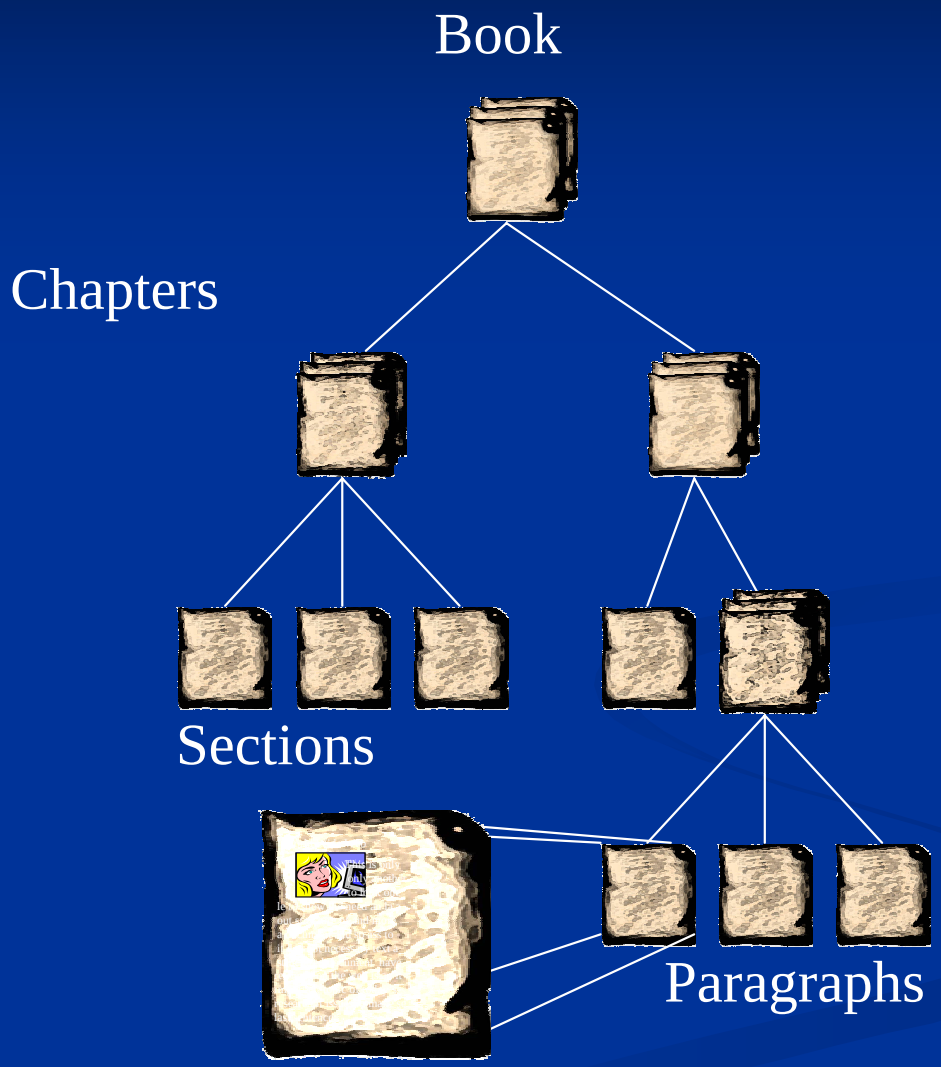
Accessing XML documents: The INEX initiative

Mounia Lalmas, Thomas Rölleke, Zoltán Szlávik, Tassos Tombros
(+ Duisburg-Essen)

XML documents



SEARCHING = QUERYING + BROWSING

Accessing XML documents

Return document components (XML elements) of **varying granularity** (e.g. a book, a chapter, a section, a paragraph, a table, a figure, etc) relevant to the user's information need both with regards to **content** and **structure** criteria.

- **INEX:** most specific component that satisfies the query, while being exhaustive to the query
- **Shakespeare study:** best entry points, which are components from which many relevant components can be reached through browsing (ECIR02)
- …

INEX: Initiative for the Evaluation of XML retrieval

- Evaluating content-oriented XML retrieval approaches
- Collaborative effort ⇒ participants contribute to the development and the evolvement of the collection and its uses
  - **queries**
  - **relevance assessments**
  - relevance assessment interface
  - metrics
  - tracks
  - data
- Similar “methodology” as for TREC, but adapted to XML retrieval
- 57 participants worldwide in 2004
- Workshop in Dagstuhl in December (22 institutions in 2003)

INEX Test Collection

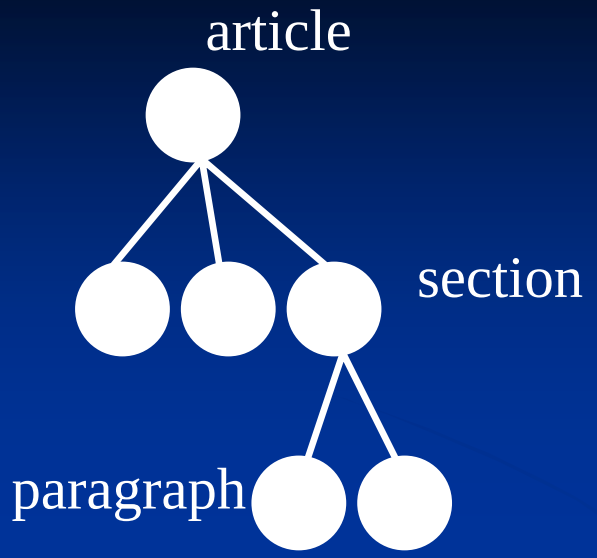
- Documents (~500MB), which consist of 12,107 articles in XML format from the IEEE Computer Society; 8 millions elements!

- INEX 2002 (JASIST04)
  - 30 CO and 30 CAS queries
  - inex2002 metric

- INEX 2003 (SIGIR FORUM 04)
  - 36 CO and 30 CAS queries
  - CAS queries are defined according to enhanced subset of XPath
  - inex2002 and inex2003 metrics

- INEX 2004
  - 40 CO and 34 CAS
  - Official: inex2002 with averaged different “assumed user behaviours”
  - Others: inex2003, CG, T2I, ERR, …

Topics



- Content-only (CO) queries

'open standards for digital video in distance learning'

- Content-and-structure (CAS) queries

//article [about(., 'formal methods verify correctness aviation systems')]
/body//section
[about(.,'case study application model checking theorem proving')]

(NEXI)

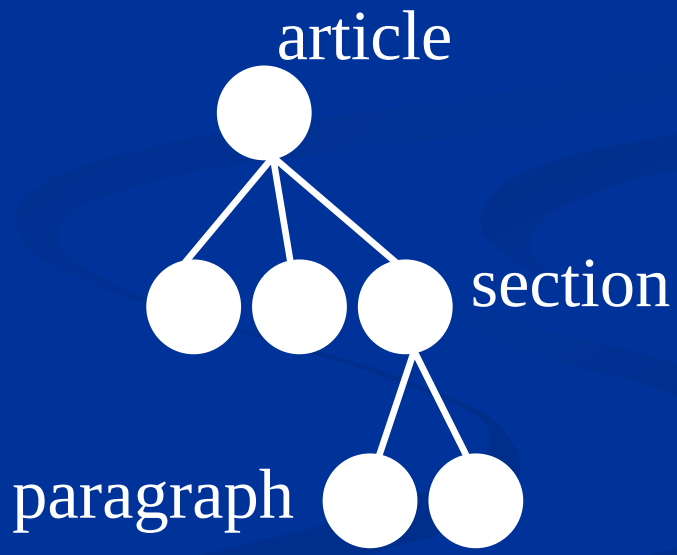
Tasks (ad hoc retrieval)

- **CO**: aim is to decrease user effort by pointing the user to the most specific relevant elements.
- **SCAS**: ~~retrieve relevant elements that exactly match the structure specified in the query~~.
- **VCAS**: retrieve relevant elements even if the result elements do not exactly meet the structural conditions expressed in the query.

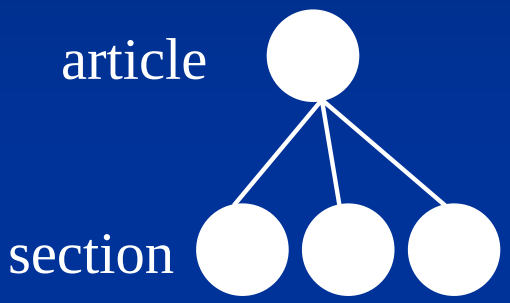
Relevance in XML

- A element is relevant if it “has significant and demonstrable bearing on the matter at hand”

- Common assumptions in IR
  - Objectivity
  - ~~Topicality~~
  - ~~Binary nature~~
  - ~~Independence~~



Relevance in INEX

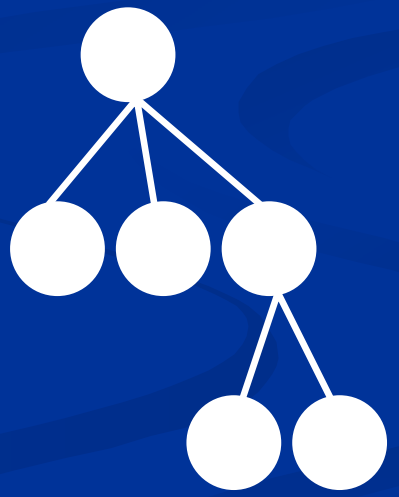


all sections relevant ⇒ article very relevant
all sections relevant ⇒ article better than sections
one section relevant ⇒ article less relevant
one section relevant ⇒ section better than article
…

- **Exhaustivity**
  how exhaustively an XML element discusses the query: 0, 1, 2, 3
- **Specificity**
  how focused an XML element is on the query: 0, 1, 2, 3
- **Relevance**
  (3,3), (2,3), (1,1), (0,0), …

Relevance assessment task

- Completeness
  - Element → parent element, children element
- Consistency
  - Parent of a relevant element must also be relevant, although to a different extent
  - Exhaustivity increase going ↑
  - Specificity decrease going ↑
- Use of an online interface
  - Assessing a query takes a week!
  - Average 2 topics per participants
  - Duplicate assessments



Assessments

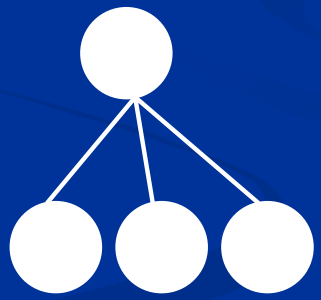
- With respect to the elements to assess
  - 26 % assessments on elements in the pool (66 % in INEX 2002).
  - 68 % highly specific elements (3,3) not in the pool
- 7 % elements automatically assessed
- INEX 2002
  - 23 inconsistent assessments per query for one rule

(ECIR04, CIKM04)

Metrics

Need to consider:

- Two dimensions of relevance
- Independency assumption does not hold
- No predefined retrieval unit
- Overlap



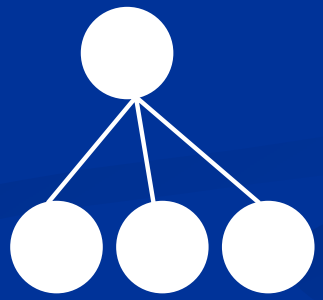
Metrics

- Recall / precision - based (inex2002, inex2003)

  quantisation functions to obtain one relevance value + capture user assumed behaviours

  expected search length

  penalise overlap
  consider size (IR0?)



- Others

  ERR: expected ratio of relevant (INEX03)
  CG: cumulated gain-based metrics (SIGIR04)
  T2I: tolerance to irrelevance (RIAO04)

(Working group report in INEX03)

Overlap problem

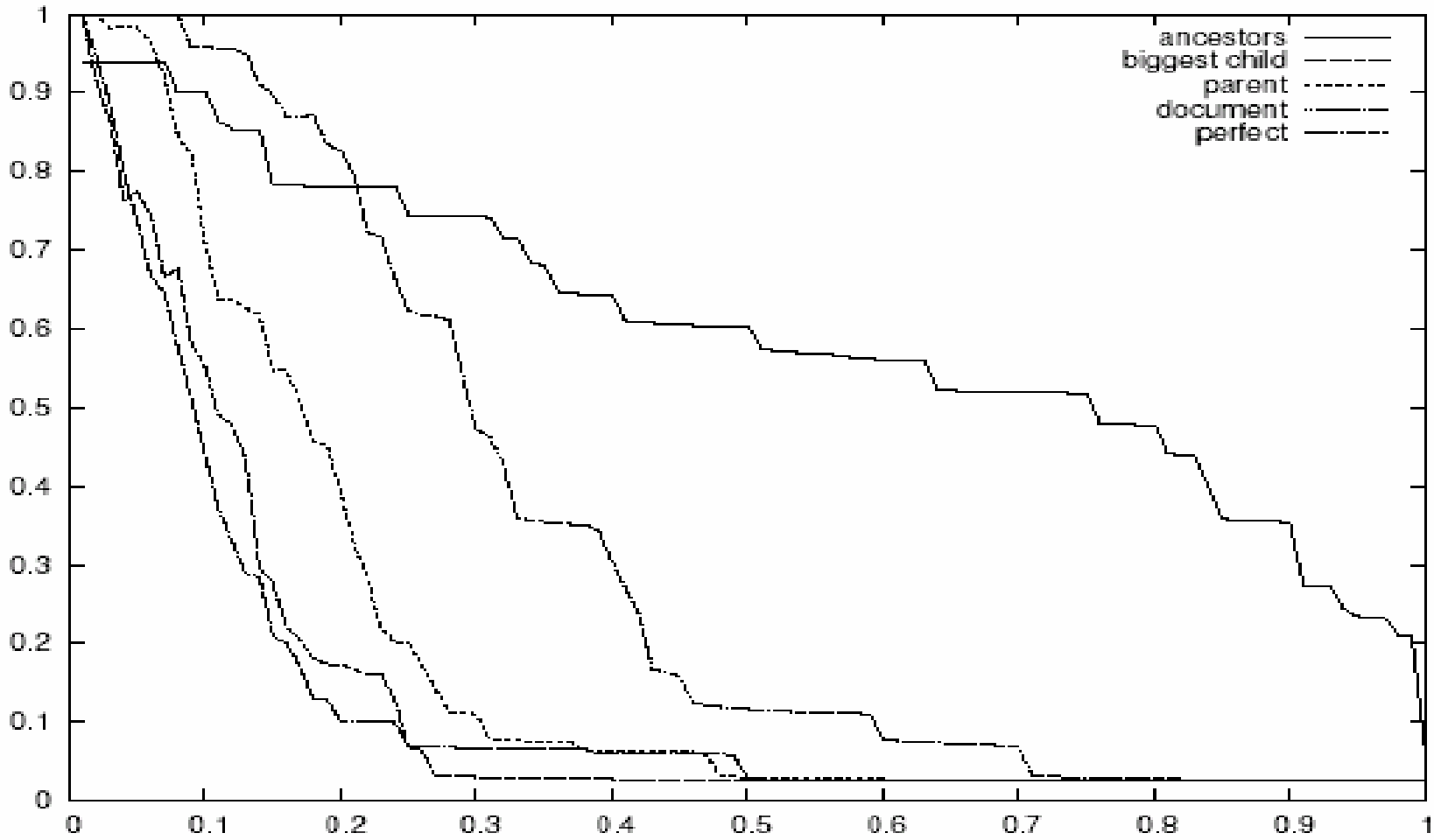


Figure 2. Generalised precision-recall. The axis of abecissas represents recall and the axis of ordinate the precision. Precision are averaged over the queries.

Lessons learnt

- Good definition of relevance
- Expressing CAS queries is not easy
- Relevance assessment process must be “improved”
- Further development on metrics needed
- User studies required

- Real scenarios and environments

INEX 2004 tracks

- Interactive
  - Follow very much interactive TREC but adapted to XML
  - Explorative study of user behaviours when presented with XML elements
  - Baseline interface + fixed tasks
- Heterogeneous collection
  - Berkeley bib, FIZ Karlsruhe, Duisburg-Essen bib, DBLP, HCI resources, QMUL db
  - Small numbers of CO and CAS topics
  - Qualitative rather than quantitative

INEX 2005?

- Metrics - much more work needed

- Multimedia track
- Elsevier, Lonely Planet, Chinese, …

- Interactive - more focussed studies, …
- Heterogeneous - more heterogeneity, …
- Context e.g. digital libraries, intranet, e-learning
- Formal evaluation based on logic-based meta-theories
- …

Acknowledgements

- FERMI ended in 1996 (Glasgow, Dortmund, Grenoble and Pisa)

- INEX participants

  LIP6 Paris, Uni of Otago, CWI Netherlands, Uni of Amsterdam, QUT Australia, LIS Denmark, Uni of Utrecht, CMU, …