Searching XML Documents Preliminary Work

INEX 2005 Workshop
28.— 30. November 2005

Marcus Hassler
Abdelhamid Bouchachia

ALPEN-ADRIA
UNIVERSITÄT
KLAGENFURT

Overview

- Introduction
- Architecture
  - Representation
  - Storage
  - Indexing
  - Retrieval
- First Results
- Conclusion

Introduction

- Context of this work
  - eLearning
  - Infrastructure PlaNet-ET

PlaNet-ET
Platform and Network for Educational Technologies

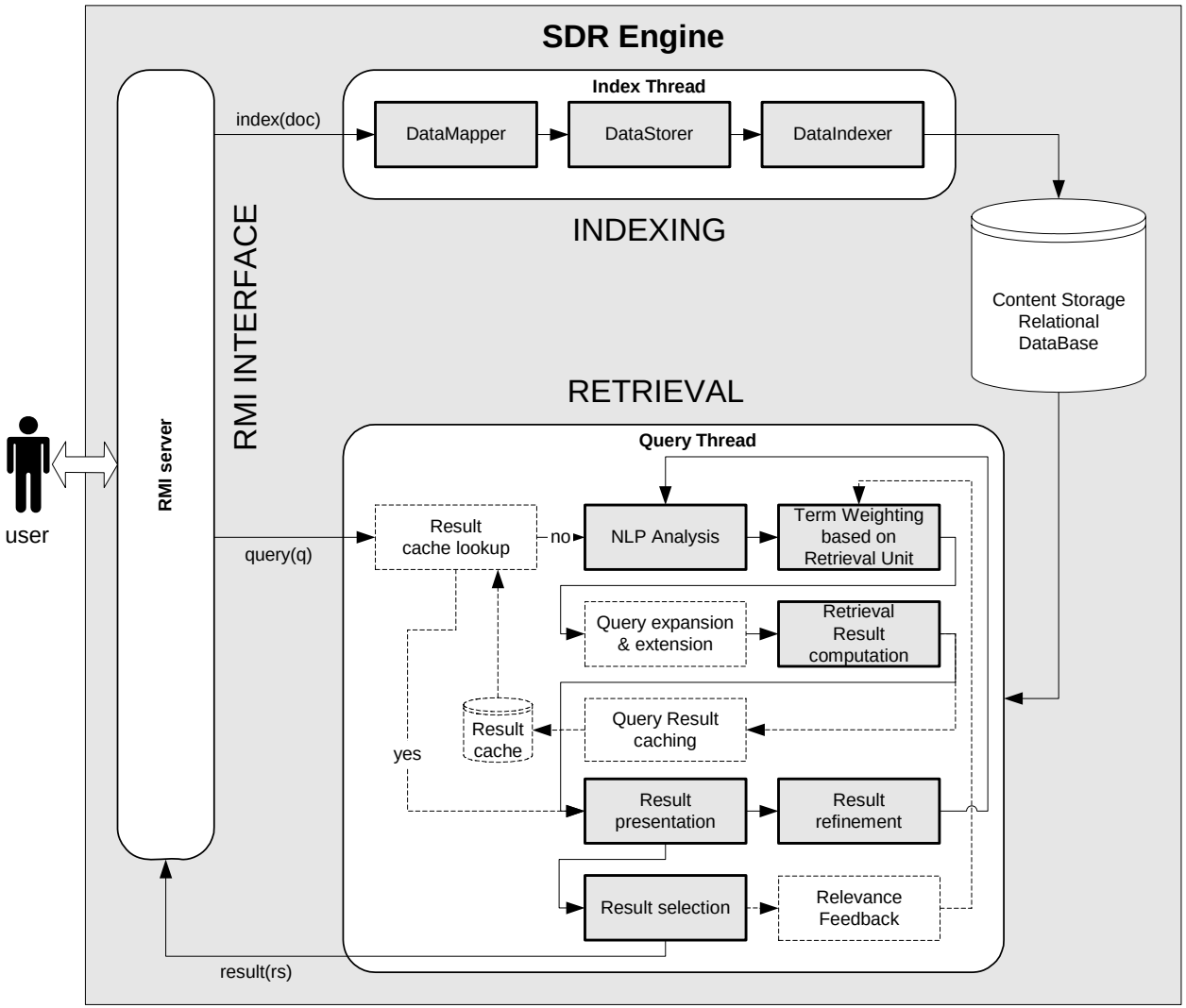
- What is PlaNet-ET
  - Online eLearning – portal [http://planet-et.at/]
  - HyperWave Information Server
    - courses
    - chunks (learning objects)

  ➔ structured content + metadata (attributes of objects, LOM)

- Relevance of XML representation
  - Extract XML documents from HIS

  ➔ independent search engine

Architecture

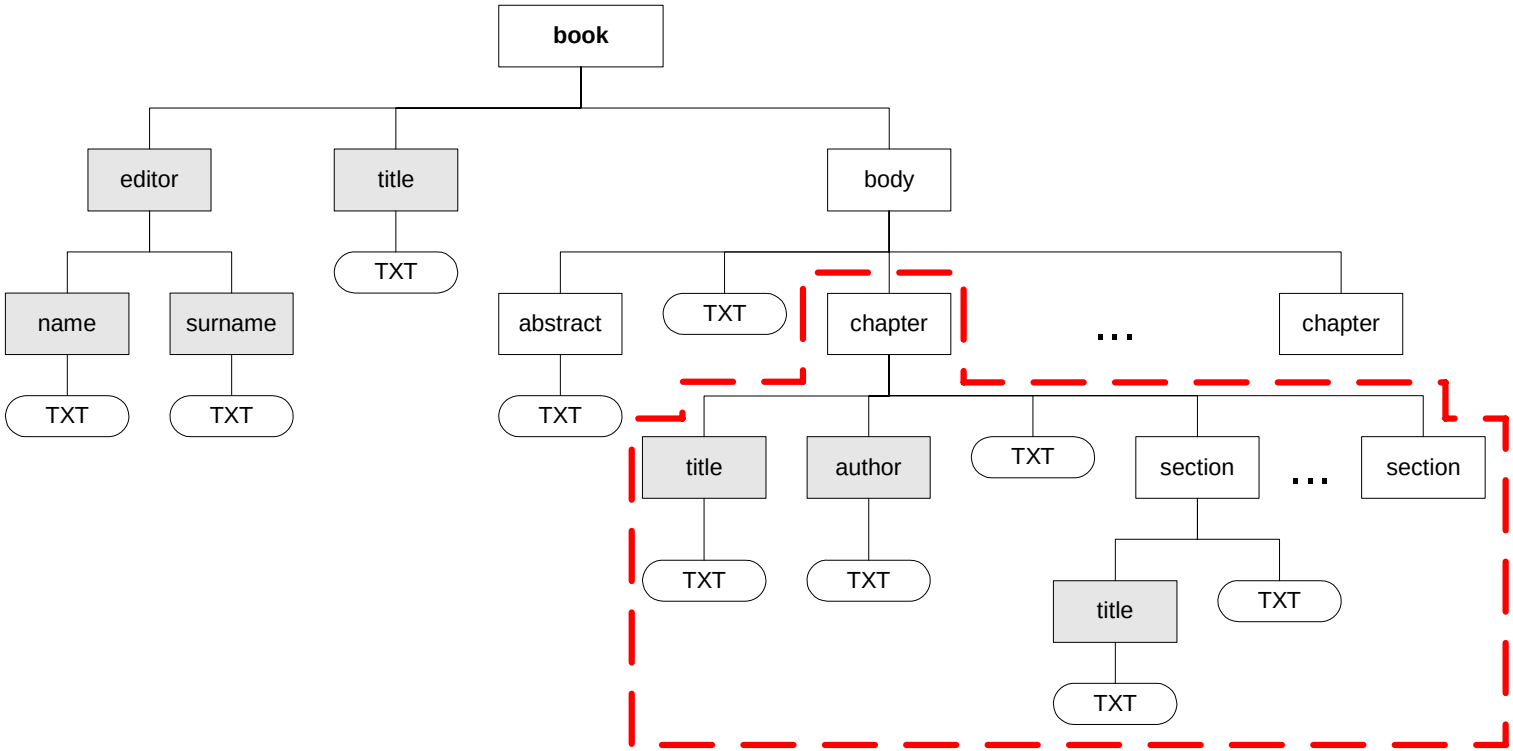


Representation

- XSLT transformation → general schema
- Consists of 3 main elements
  - **DOC**
  - **SEC**
  - **FRA** (smallest retrievable unit)
- Each Element
  - **Metadata** - block
  - **Content** - block
- Pros
  - no mixed-content nodes (txt + structure)
  - Semantic relativism (**sec=ss1=ss2=ss3**, **p=p1=ip=ilrj**)
  - Efficient storage & retrieval

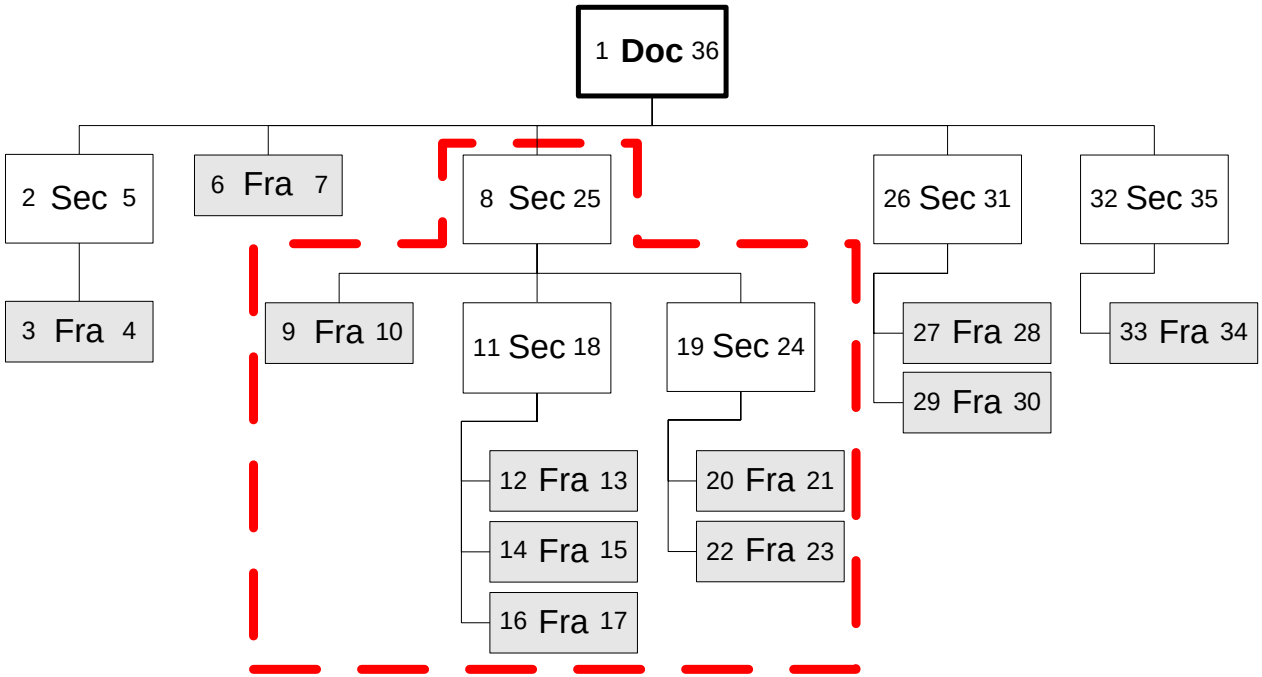
Representation

- Example



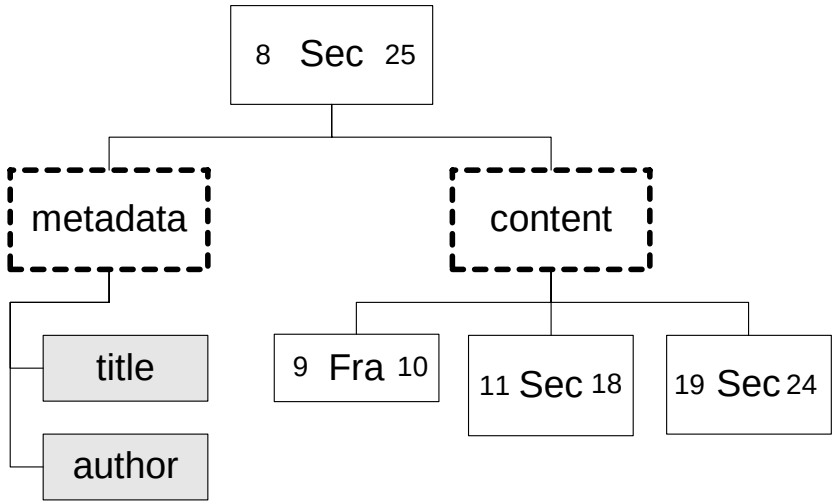
Representation

- Example



Representation

- Example



Storage

- Based on Grust's work

  [T. Grust, Accelerating XPath location steps, ACM SIGMOD 2002, pp. 109--120]

  - preorder and postorder

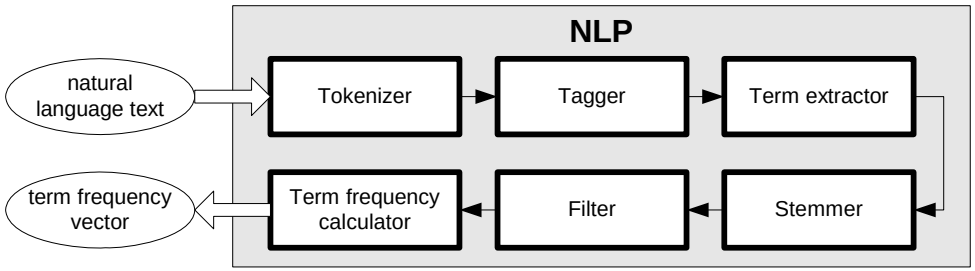
docID	pre	post	parentID	tag	path
1	1	36	0	DOC	/DOC
1	2	5	1	SEC	/DOC/SEC
1	3	4	2	FRA	/DOC/SEC/FRA
1	6	7	1	FRA	/DOC/FRA
1	8	25	1	SEC	/DOC/SEC
1	9	10	8	FRA	/DOC/SEC/FRA
1	11	18	8	FRA	/DOC/SEC/SEC
1	12	13	11	FRA	/DOC/SEC/SEC/FRA
1	14	15	11	FRA	/DOC/SEC/SEC/FRA
...	...	...	...	...	...

docID	pre	cdata
1	3	In this Paper we describe the ...
1	6	In Sec. 1 the architecture is ...
1	9	In our approach we distinguish ...
...	...	...

docID	pre	sec_title	sec_author
1	2	Abstract	R. Smith
1	8	Introduction	J. Alf
1	11	Architecture	NULL
...	...	...	...

Indexing

- Content only in leaf nodes (**FRA** element)
- Representation: term-frequency vector

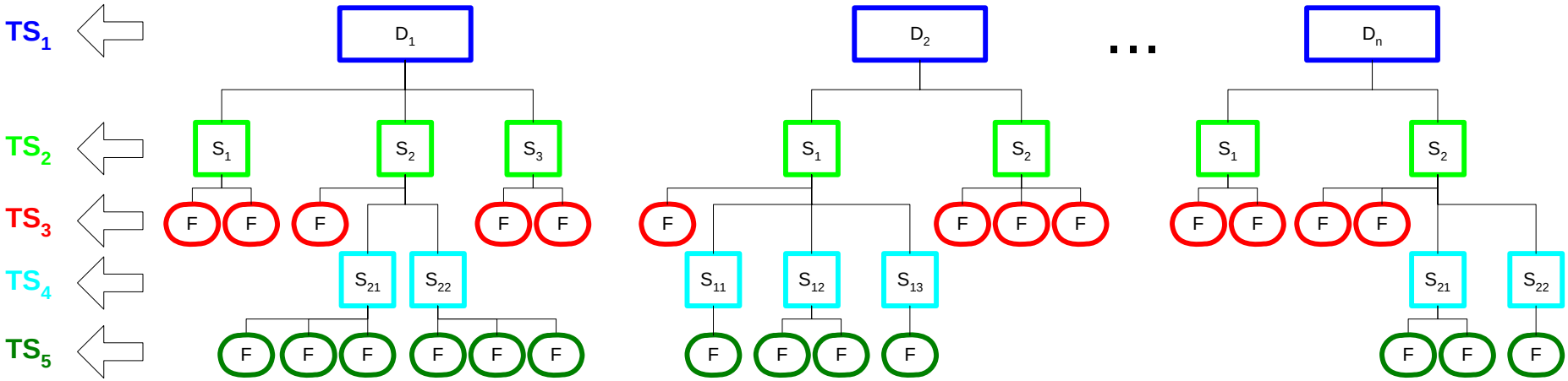


- Pros:
  - Dynamic granularity (dyn. TS + dyn. IDF)
  - Dynamic document environment
    - add, remove, re-index → on-the-fly

Retrieval

- Term weighting (VSM)
  - *Context* of a node:
    - all nodes in the same path

    e.g., **`/DOC`**, **`/DOC/SEC/SEC`**

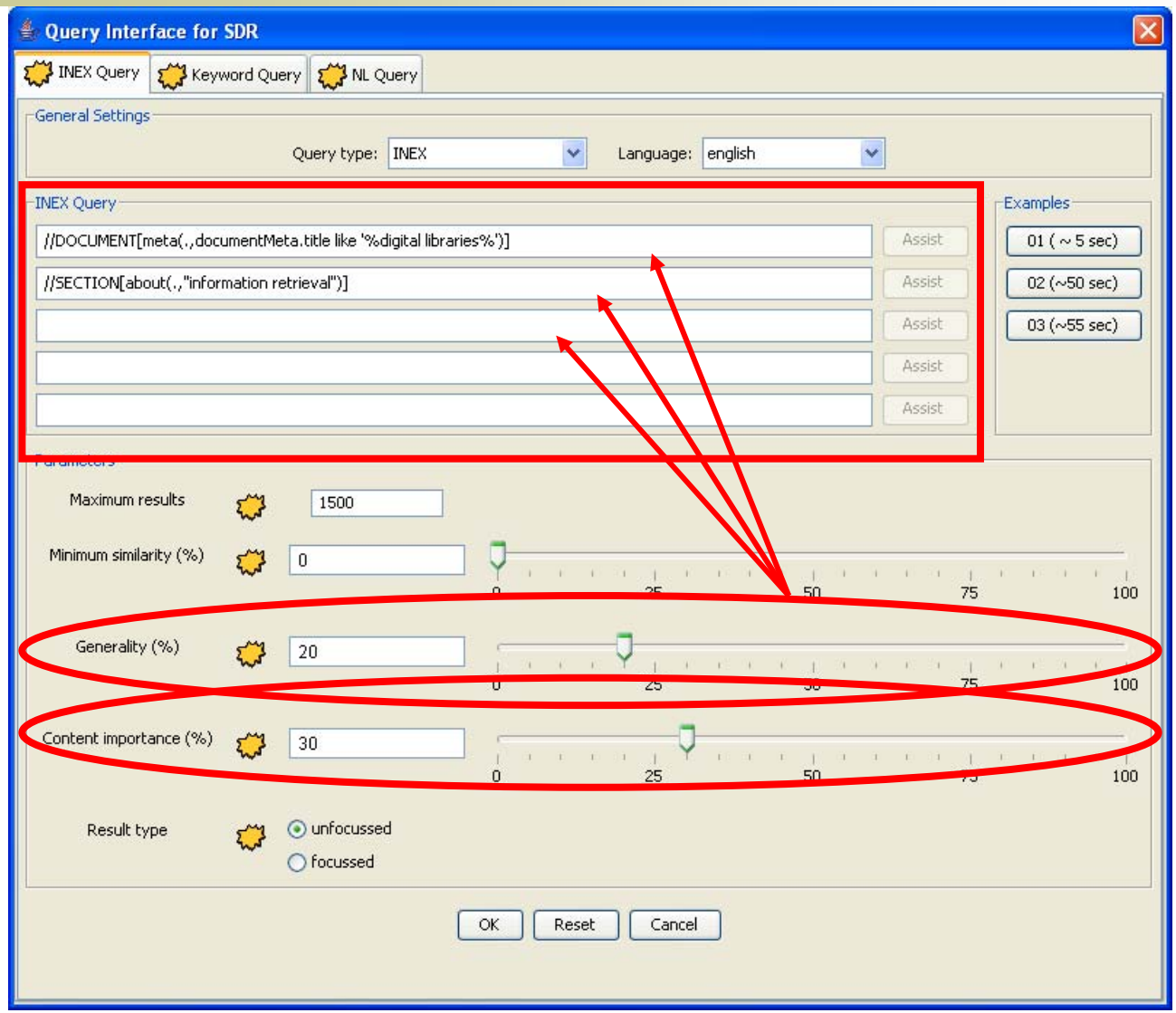


  - Dynamic term space & *idf* calculation

Retrieval

Topic

No. 265



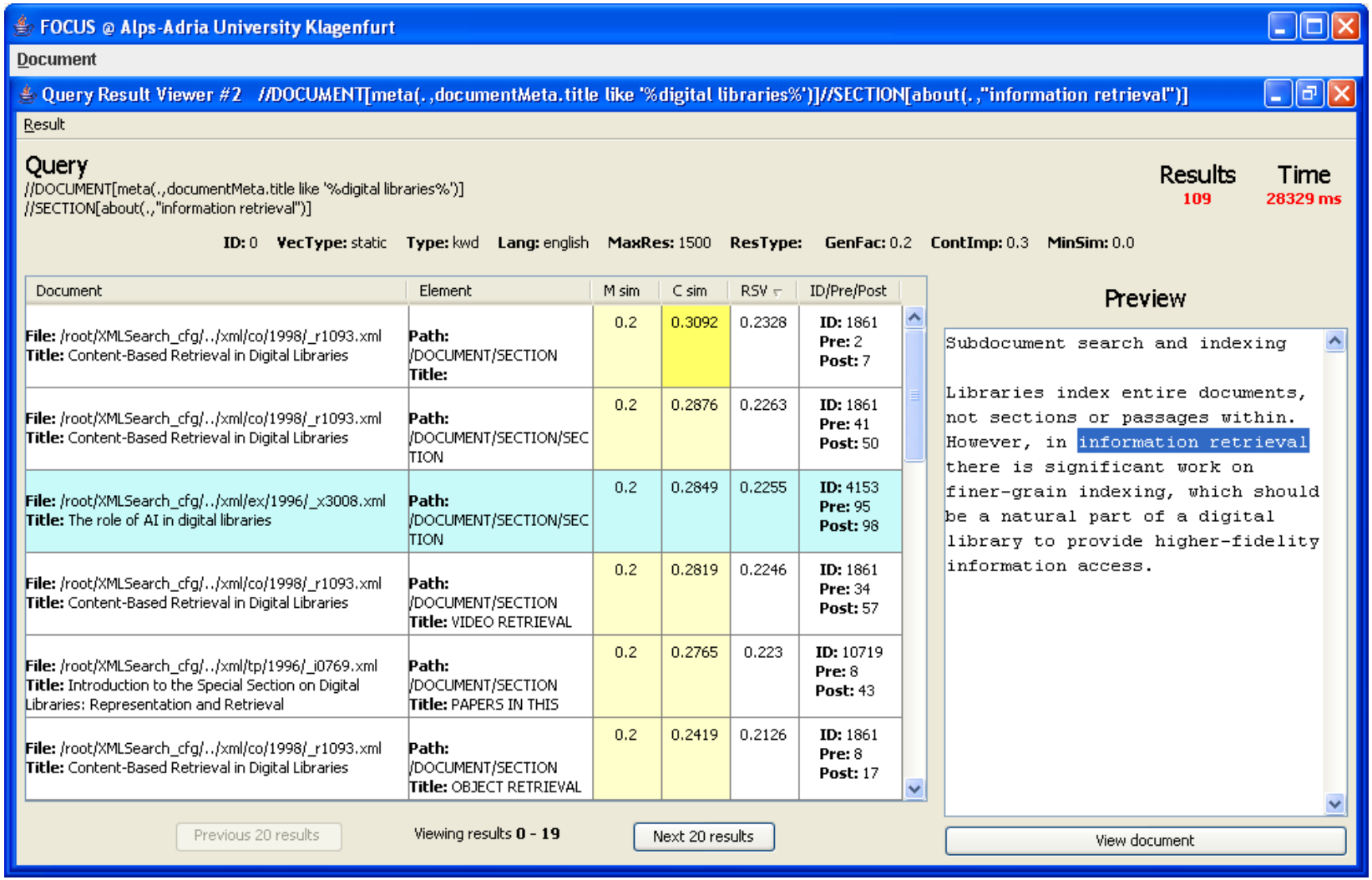
Retrieval

- Result Computation
  - Metadata similarity
  - Content similarity
  - ➔ Combined to calculate RSV

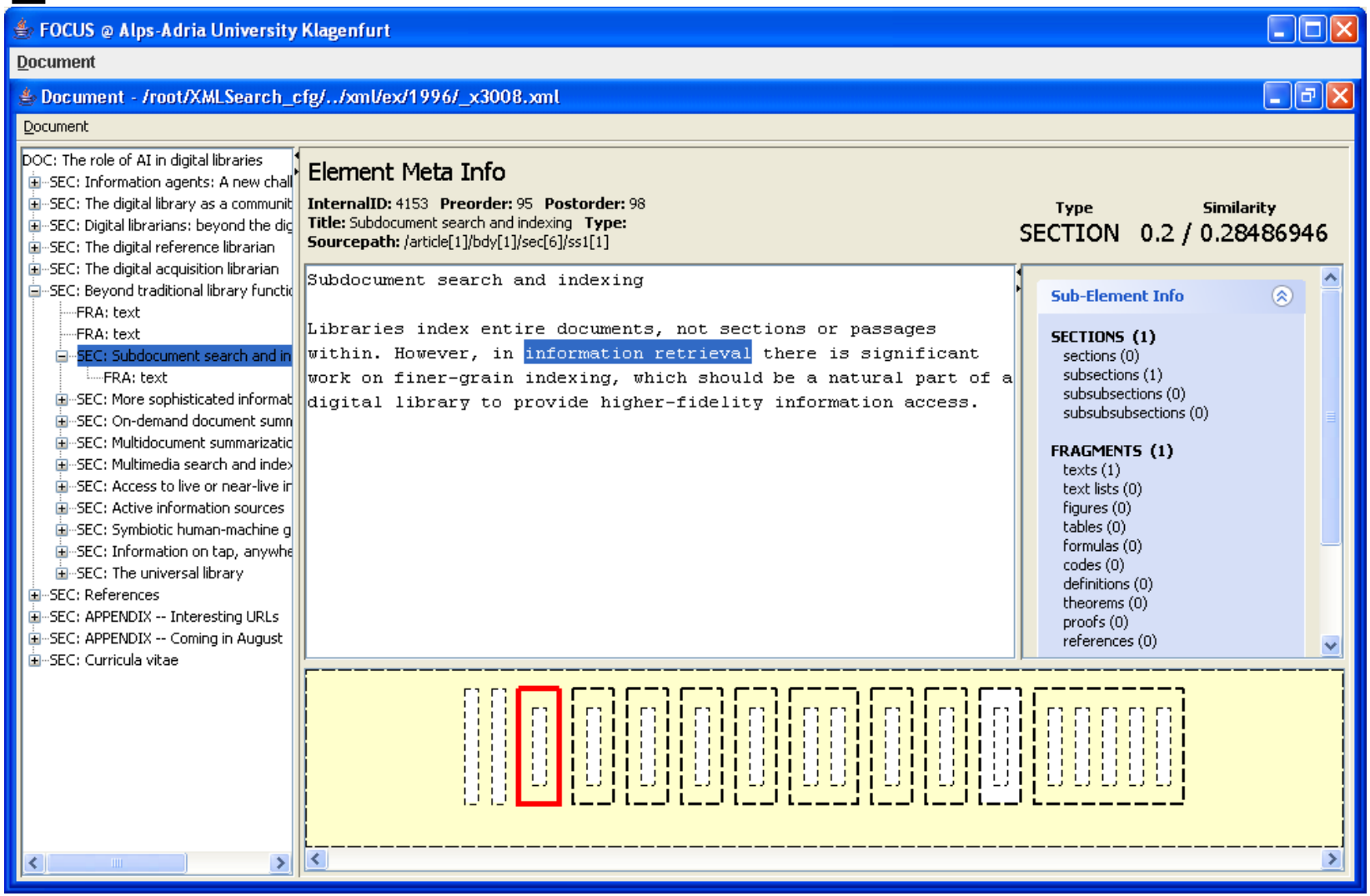
    rsv = contSim * *ci* + metaSim * (1-*ci*)
- User-defined Granularity

  $\mathbf{sim_{new} = sim_{old} * \mathit{gf} + sim_{new} * (1-\mathit{gf})}$
- Result element: **(docID,pre,post,simMeta,simCont)**
- Result filtering (Focused task)
  - exploit preorder and postorder

Retrieval


FOCUS @ Alps-Adria University Klagenfurt
Document
Query Result Viewer #2 //DOCUMENT[meta(.,documentMeta.title like '%digital libraries%')]//SECTION[about(.,"information retrieval")]
Result
Query
//DOCUMENT[meta(.,documentMeta.title like '%digital libraries%')]
//SECTION[about(.,"information retrieval")]
Results
109
Time
28329 ms
ID: 0 VecType: static Type: kwd Lang: english MaxRes: 1500 ResType: GenFac: 0.2 ContImp: 0.3 MinSim: 0.0
Document
Element
M sim
C sim
RSV
ID/Pre/Post
File: /root/XMLSearch_cfg/../xml/co/1998/_r1093.xml
Title: Content-Based Retrieval in Digital Libraries
Path: /DOCUMENT/SECTION
Title:
0.2
0.3092
0.2328
ID: 1861
Pre: 2
Post: 7
File: /root/XMLSearch_cfg/../xml/co/1998/_r1093.xml
Title: Content-Based Retrieval in Digital Libraries
Path: /DOCUMENT/SECTION/SECTION
0.2
0.2876
0.2263
ID: 1861
Pre: 41
Post: 50
File: /root/XMLSearch_cfg/../xml/ex/1996/_x3008.xml
Title: The role of AI in digital libraries
Path: /DOCUMENT/SECTION/SECTION
0.2
0.2849
0.2255
ID: 4153
Pre: 95
Post: 98
File: /root/XMLSearch_cfg/../xml/co/1998/_r1093.xml
Title: Content-Based Retrieval in Digital Libraries
Path: /DOCUMENT/SECTION
Title: VIDEO RETRIEVAL
0.2
0.2819
0.2246
ID: 1861
Pre: 34
Post: 57
File: /root/XMLSearch_cfg/../xml/tp/1996/_i0769.xml
Title: Introduction to the Special Section on Digital Libraries: Representation and Retrieval
Path: /DOCUMENT/SECTION
Title: PAPERS IN THIS
0.2
0.2765
0.223
ID: 10719
Pre: 8
Post: 43
File: /root/XMLSearch_cfg/../xml/co/1998/_r1093.xml
Title: Content-Based Retrieval in Digital Libraries
Path: /DOCUMENT/SECTION
Title: OBJECT RETRIEVAL
0.2
0.2419
0.2126
ID: 1861
Pre: 8
Post: 17
Previous 20 results
Viewing results 0 - 19
Next 20 results
Preview
Subdocument search and indexing
Libraries index entire documents, not sections or passages within. However, in information retrieval there is significant work on finer-grain indexing, which should be a natural part of a digital library to provide higher-fidelity information access.
View document

Retrieval

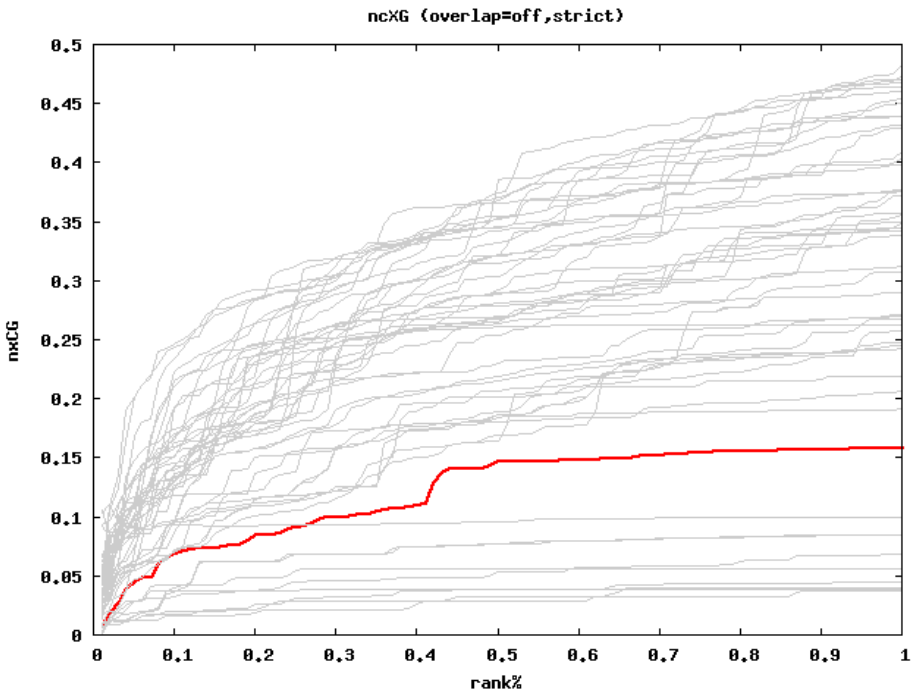

FOCUS @ Alps-Adria University Klagenfurt
Document
Document - /root/XMLSearch_cfg/../xml/ex/1996/_x3008.xml
Document
DOC: The role of AI in digital libraries
SEC: Information agents: A new chall
SEC: The digital library as a communit
SEC: Digital librarians: beyond the dig
SEC: The digital reference librarian
SEC: The digital acquisition librarian
SEC: Beyond traditional library functi
FRA: text
FRA: text
SEC: Subdocument search and in
FRA: text
SEC: More sophisticated informat
SEC: On-demand document summ
SEC: Multidocument summarizatic
SEC: Multimedia search and index
SEC: Access to live or near-live in
SEC: Active information sources
SEC: Symbiotic human-machine g
SEC: Information on tap, anywhe
SEC: The universal library
SEC: References
SEC: APPENDIX -- Interesting URLs
SEC: APPENDIX -- Coming in August
SEC: Curricula vitae
Element Meta Info
InternalID: 4153 Preorder: 95 Postorder: 98
Title: Subdocument search and indexing Type:
Sourcepath: /article[1]/bdy[1]/sec[6]/ss1[1]
Type
SECTION
Similarity
0.2 / 0.28486946
Subdocument search and indexing
Libraries index entire documents, not sections or passages within. However, in information retrieval there is significant work on finer-grain indexing, which should be a natural part of a digital library to provide higher-fidelity information access.
Sub-Element Info
SECTIONS (1)
sections (0)
subsections (1)
subsubsections (0)
subsubsubsections (0)
FRAGMENTS (1)
texts (1)
text lists (0)
figures (0)
tables (0)
formulas (0)
codes (0)
definitions (0)
theorems (0)
proofs (0)
references (0)

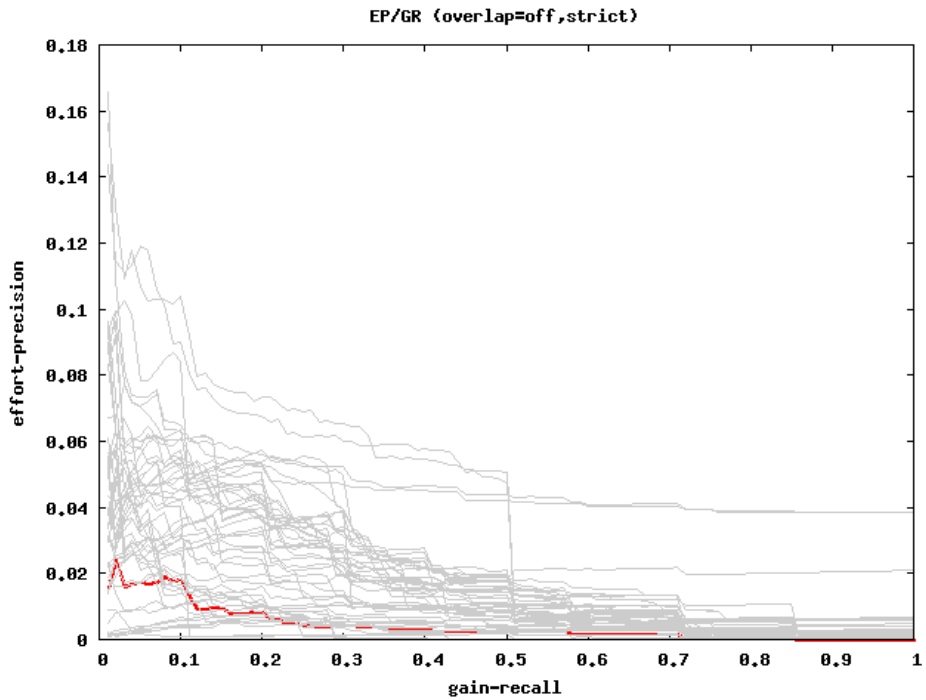
First Results

- Used collection v1.4 (12107 docs)
  - missing 4712 docs from v1.8 (16819 docs)
- XML transformation (XSLT)
  - ~ 0,42 % errors (51 docs) → manually fixed
- Topic transformation
  - exchange element/path names
    e.g., **`/article` → `/DOC`**
  - map conditions for metadata
    e.g., **`about(.//atl,"information retrieval")` → `meta(/DOC,title like '%information retrieval%')`**
- Difficulties:
  - Some topics (215, 231, 241) contain "and", "in", "for", "+", ...
  - Path mapping:
    `/article` → `/DOC`
    `/article/bdy` → `/DOC`

First Results

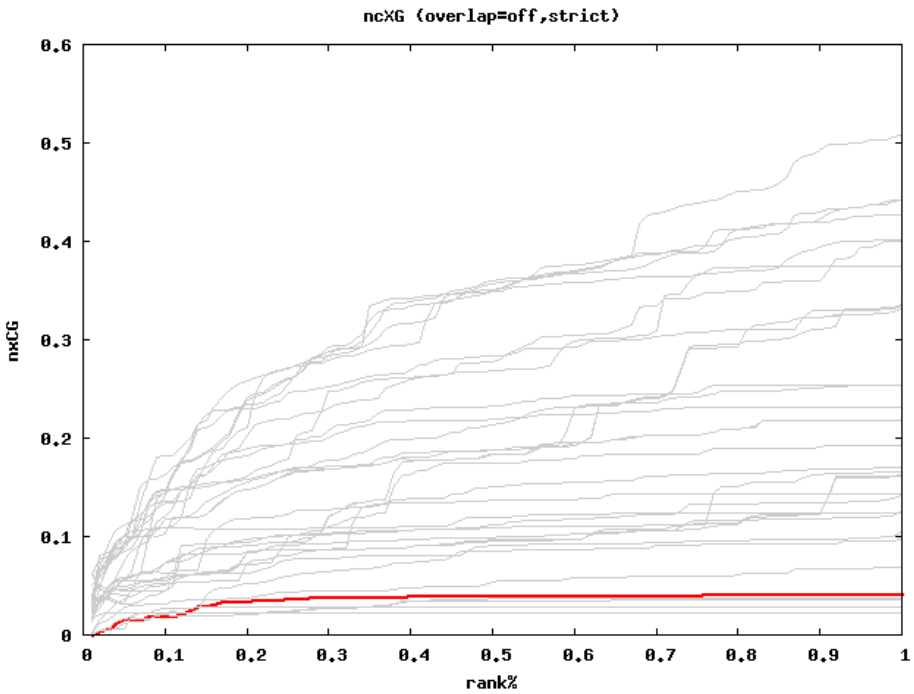
- CO.Thorough

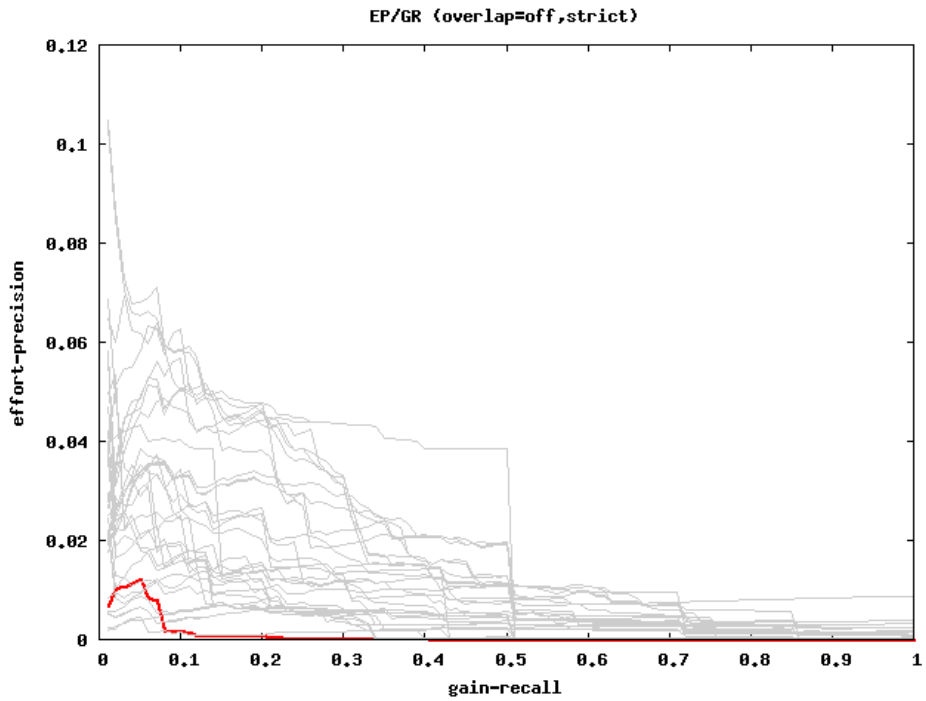




First Results

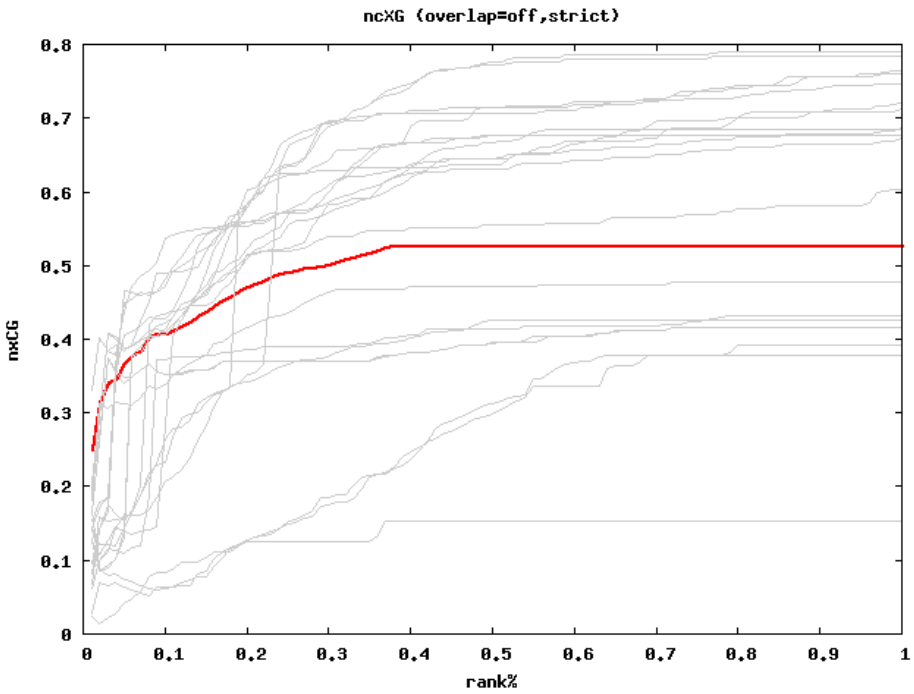
- COS.Thorough

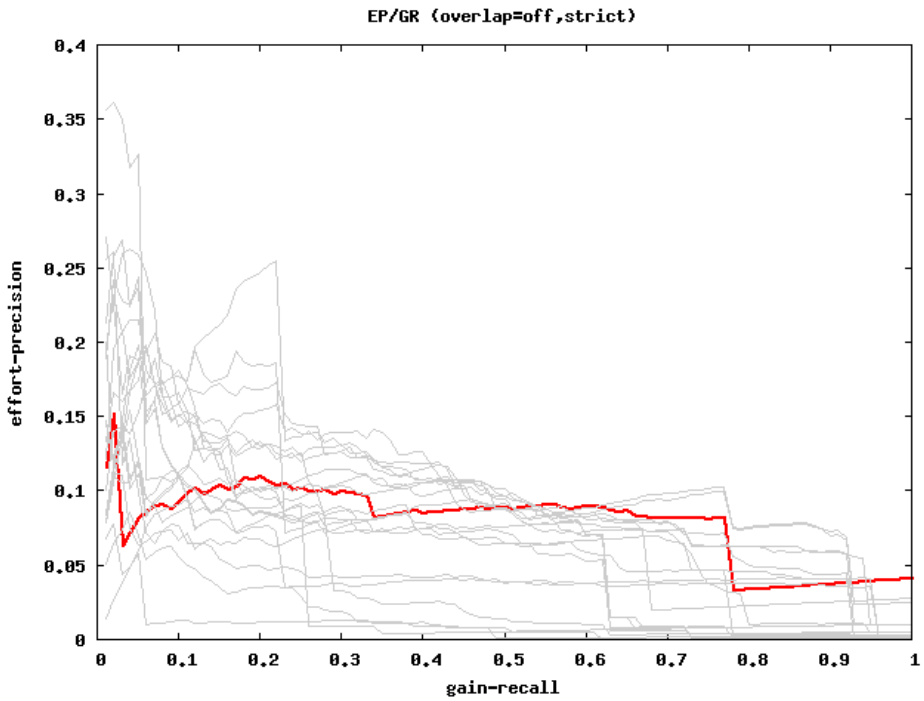




First Results

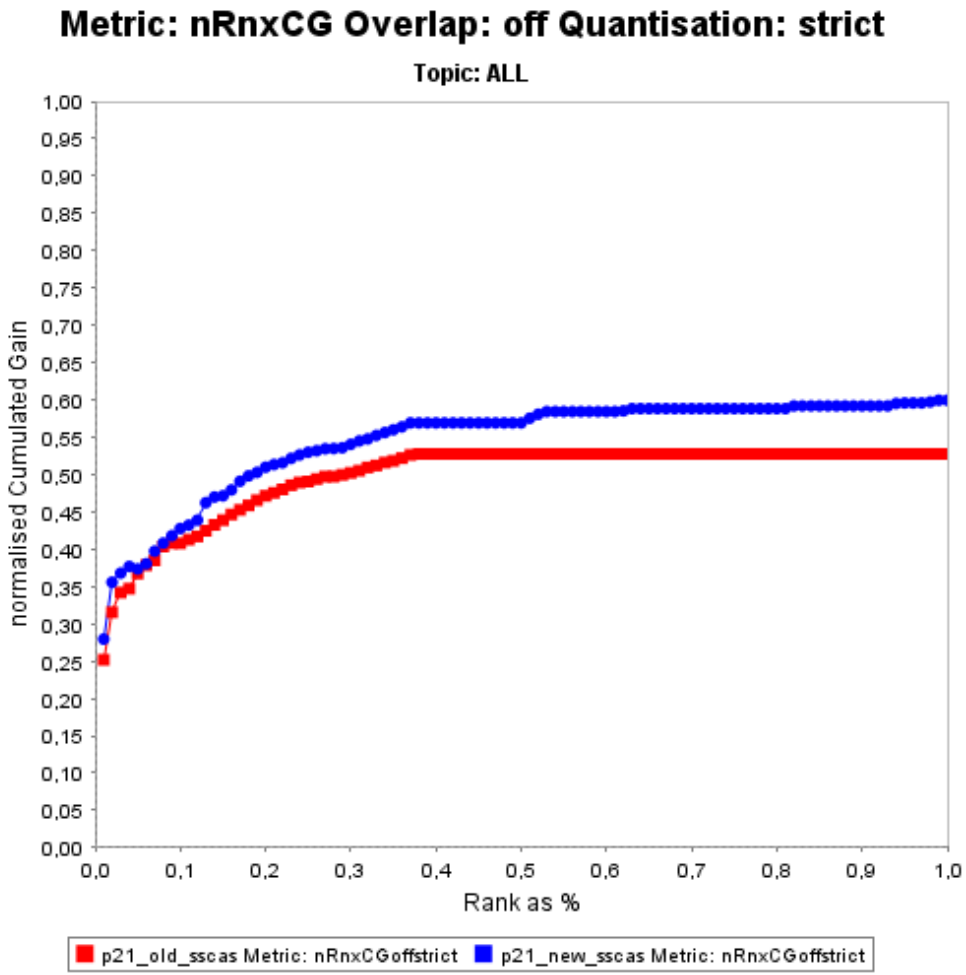
- SSCAS





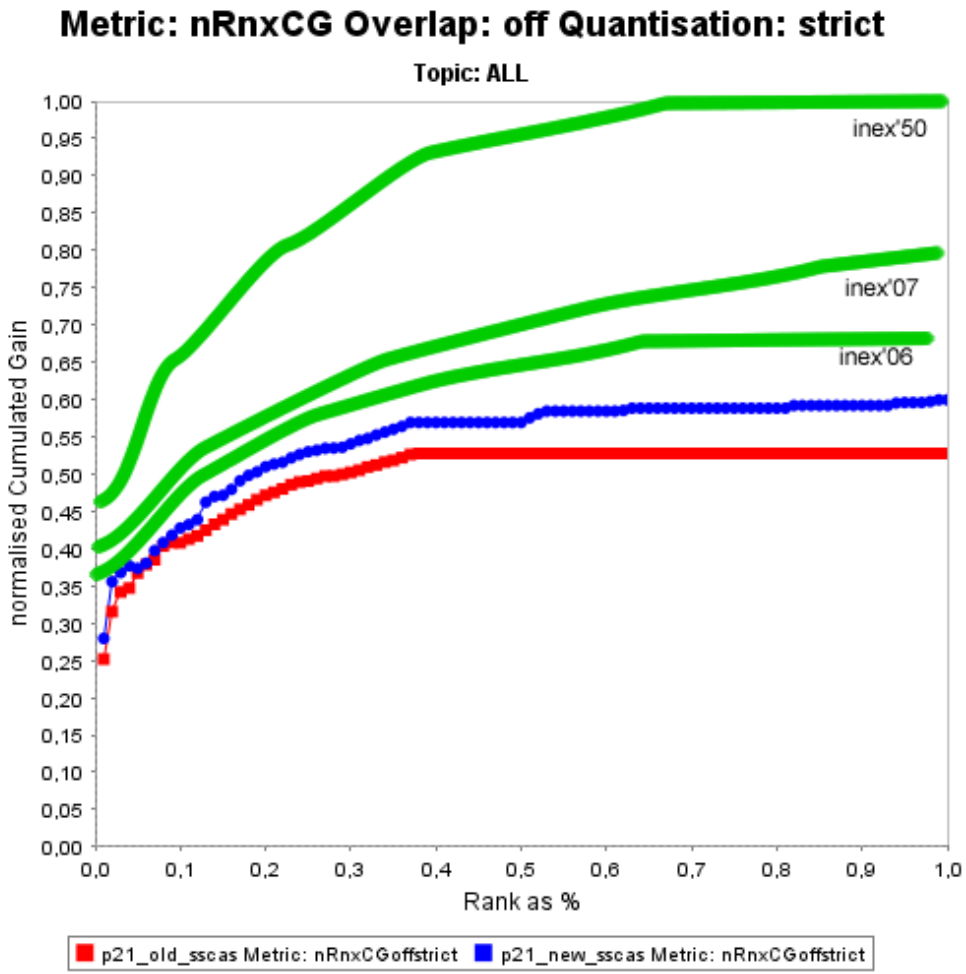
First Progress

- SSCAS



First Progress

- SSCAS



Conclusion

- Independent system architecture
- Dynamic aspects
  - Term frequency propagation
  - Term space
  - *idf* calculation
- Initial experiments
  - Only 3 runs (CO.Thorough, COS.Thorough, SSCAS)
  - Not too bad results (SSCAS)
- Further work
  - Adapt INEX corpus 1.8
  - Improve metadata typing and matching
  - Include multi-term index (e.g., information retrieval)
  - Complete Evaluation