Understanding Content-and-Structure

Jaap Kamps Maarten Marx Maarten de Rijke
Börkur Sigurbjörnsson

Informatics Institute
University of Amsterdam
{kamps,marx,mdr,borkur}@science.uva.nl

INEX 2005 Methodology Workshop
Glasgow, July, 2005

Overview

- Expressive power of query languages
- What do users mean by a target constraints?
- What sort of query trees are popular?

Who is our user?

- Our ideal user
  - Experienced searcher
  - Not necessarily technically literate
  - Example: information professional, librarian
- Ignorant users
  - **Some** knowledge of the **tag-names**
  - **No** knowledge of the **hierarchy** of tag-names
- Semi-ignorant users
  - **Some** knowledge of the **tag-names**
  - **Some** knowledge of the **hierarchy** of tag-names

Expressive power of query languages

- Assign users a query language that matches their knowledge
  - The expressive power of the query language should fit the expressive power of the user
- A query language should be safe and complete
  - **Safety**: If the user cannot distinguish A and B, then no query can
  - **Completeness**: If the user can distinguish A and B, then a query can
- A note on XPath
  - Expressiveness of XPath fragments is a hot research topic
  - It's ideal research purposes for talking about trees
  - It's not meant as an end-user language

Overview

- Expressive power of query languages
- What do users mean by a target constraints?
- What sort of trees are popular?

User wants a specific granularity?

- E3S3 judgements from 2003 and 2004

	article+	sec+	p+	abs	vt
article (11)	**42.89%**	25.62%	18.79%	0.19%	0.76%
sec (20)	10.17%	**38.60%**	25.92%	1.36%	0.20%
p (5)	11.26%	21.13%	**49.30%**	3.52%	–
abs (4)	14.40%	37.29%	22.88%	**11.86%**	–
vt (2)	–	–	42.86%	–	**53.13%**

- Users seem not to take granularity constraint seriously
- Users seem to have a bias toward wanting what they ask
  - Information need controls granularity?
  - Target constraint controls assessments?
- But the assessment guidelines said ... (to be continued)

Can we have structured information needs?

- "But the assessment guidelines said that structural constraints should be ignored"
  - I'd say, rightfully so
- Why should we consider structural constraints strictly?
- What is an abstract?
  - Text inside an `<abs>` tag?
  - "A statement summarizing the important points of a text"? (dictionary.com)
  - Wouldn't an **introduction** section be satisfactory?
  - Wouldn't a **conclusions** section be satisfactory?
- Claim: Structure is never an inherent part of an information need

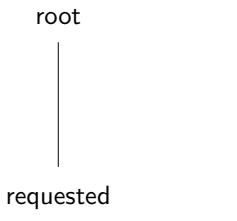
Overview

- Expressive power of query languages
- What do users mean by a target constraints?
- What sort of trees are popular in the INEX collection?

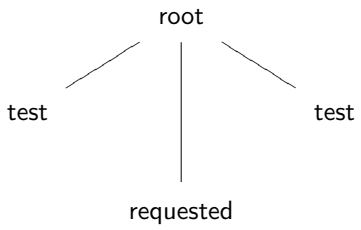
Classification of queries

- We classify the INEX queries based on whether or not they express **hierarchical relationships** between elements
- We do not consider hierarchical relationships with the `<article>` tag since that relationship is trivial
- Examples of non-hierarchical queries:
  - //sec[about(., java thread implementation)]
  - //article[about(.//abs, java)]//sec[about(., thread implementation)]
- Examples of hierarchical queries:
  - //article[about(.//fm//abs, java)]//bdy//sec[about(,. thread implementation)]

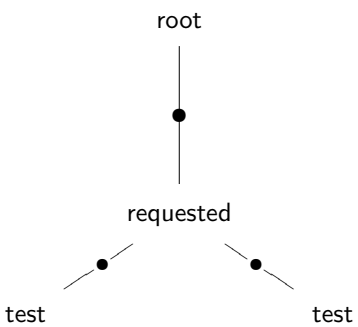
Shapes of the queries



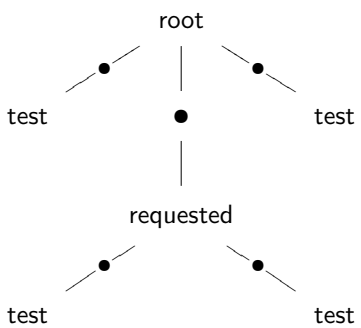
Restricted Search (15%)



Contextual Content Information (47%)

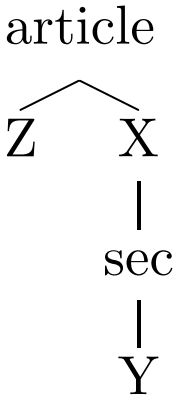


Search Hints (6%)



Search Hints in Context (32%)

Popular trees



- Most popular query template
- Appears 14 times among the 2004 queries
- Popular replacements for X
  - ϵ (10); bdy (4)
- Popular replacements for Z
  - α (8); abs/α (3); atl/α (2); fm/yr/α (1); kwd/α (1)
- Popular replacements for Y
  - α (12); */α (1); st/α (1)

Summary

- Expressive power of query languages
  - Choose a suitably expressive query language for users
  - This will minimize change of "semantic mistakes"
- What do users mean by structural constraints?
  - It's only a hint
  - Structure is never an inherent part of an information need
- What sort of query trees are popular?
  - Non-hierarchical
  - Including constraints on context

Micro average

		article	bdy	sec+	p+	abs	vt	bib+	bb	fig	figc	fm
article	12	**18.5**	**10.5**	**15.2**	**11.8**	0.1	0.5	–	10.5	–	0.5	0.1
sec	20	**5.2**	**5.0**	**38.6**	**25.9**	1.4	0.2	0.9	0.4	0.4	0.5	0.9
p	5	**5.6**	**5.6**	**21.1**	**49.3**	3.5	–	–	–	–	2.1	2.8
'*'	2	**2.6**	**2.6**	**76.9**	**17.9**	–	–	–	–	–	–	–
abs	4	9.3	5.1	37.3	22.9	**11.9**	–	–	–	0.8	1.7	5.1
vt	2	–	–	–	42.9	–	**53.1**	–	–	–	2.9	–
bib	1	–	–	–	–	–	–	–	**50.0**	–	–	–
bb	1	–	–	–	–	–	–	–	**95.7**	–	–	–
fig	2	–	–	72.3	2.1	–	–	–	–	**2.1**	**2.1**	–
fm	1	–	–	–	40.0	–	–	–	–	–	–	–

Macro Average

		article	bdy	sec+	p+	abs	vt	bib+	bb	fig	figc	fm
article	12	**24.7**	**15.0**	**29.3**	**10.5**	0.2	0.5	–	2.6	–	0.5	0.6
sec	20	**9.6**	**9.6**	**39.9**	**28.1**	0.5	0.1	0.2	0.1	0.3	0.2	0.5
p	5	**12.0**	**13.3**	**23.5**	**38.6**	1.1	–	–	–	–	0.7	0.9
'*'	2	**1.7**	**1.7**	**57.8**	**38.9**	–	–	–	–	–	–	–
abs	4	11.3	4.0	15.5	22.3	**11.8**	–	–	–	0.3	6.6	8.5
vt	2	–	–	–	22.0	–	**74.7**	–	–	–	1.5	–
bib	1	–	–	–	–	–	–	–	**50.0**	–	–	–
bb	1	–	–	–	–	–	–	–	**95.7**	–	–	–
fig	2	–	–	66.9	1.2	–	–	–	–	**10.0**	**10.0**	–
fm	1	–	–	–	40.0	–	–	–	–	–	–	–