Queries, INEX 2003 working group report

Holger Flörke, Norbert Fuhr, Kenji Hatano,
Börkur Sigurbjörnsson, Andrew Trotman, Masahiro Watanabe

January 13, 2004

1 CO

There was no discussion on CO topics in the working group. This is to be interpreted a support for leaving the CO topic format uncanged for atleast next year.

2 CAS

The first discussion was about the complexity of the INEX 2003. It seems that people find it difficult to formulate the XPath-like expressions of the topic title. In the initially distributed set of CAS queries, 63% of the queries turned out to be an error [3]. This is in line with research that shows that users have great difficulty with boolean queries, both in databases and information retrieval [2].

In view of the high error rate there was discussion about syntax clarification, expressiveness restrictions and even a new syntax [3].

On top of difficulties with the topic syntax, there was also discussion about the difficulty of expressing natural information need with this collection. It was questioned whether topic authors add structural constraints because they think it is useful of whether they do it only because they need to write a structured query.

In particular, we discussed the topics that put constraints on contents of sections or paragraphs.

- Target elements
  - Natural target elements are scarce for this collection
    - If topic authors ask for a `<p>`, are they honestly asking for a paragraph
    - Wouldn't they be happy with any text bearing element?
    - Would they be happy with for example `//bdy//*`?
  - We found the following natural targets

        - Textual stuff `<sec>`, `<ss1>`, `<p>`...
        - Vitaes `<vt>`
        - Bibliographical entries `<bb>` ...
- Structural conditions
    - Co-occurrences
        - We want certain concepts to be covered in the same unit
            - `//article[about(.//bdy//*,''XML query language'')]`
    - Data-types
        - t.b.a.
    - Roles
        - article author, author affiliation, ...
- Separation of structural constraints and target elements
    - This was tried at INEX 2002 but was abandoned for INEX 2003
    - People wanted to express something like
        - `//bdy//*[about(.,''solar powered robots'') and about(.//fig,''robot on mars'')]`
- What should be used for VCAS/SCO?
    - Extended version of CO?
    - Restricted version of CAS?
- Can we provide query generation tool to help users create CAS topics?

3 The Otago proposal

- Very good idea to restrict XPath usage
- But do we need to change the syntax?
- Can't we just limit the XPath usage but keep the syntax?

4 Query language for INEX 2004

As a query language, the group considered an extension of a subset of XPath. We did not reach an agreement about the actual syntax of the query language. Some wanted an XPath-like syntax, other prefered the syntax proposed by the Otago-group.

The group put together the following list of requirements

- Data types with vague predicates

  - english text: about
  - dates/numeric: ¡,¿,=
  - person names: ≈ Chebychef

- Path specifications
  - element name specifications
  - not instances of tags (sec[3])
  - not child (use only decendant)
  - not element content comparison (au!='kohonen') (use datatypes for this)

- vague interpretation of connectors

We restrict our XPath use to - // (decendant), tagname (role), *, [] (filter)

- Basic building blocs
  - decendant relation (//)
  - search constraints, (filters) ([ ])
  - search predicate (about(.,.))
  - numeric comparison ($<, >, =$)
  - unspecified target element (*)

- Possible extensions
  - attributes
  - specific datatypes (dates, authors, etc.)

- Filter operations
  - How should we interpret boolean combinations of predicates?
    * strict vs. vague
    * $\wedge$ vs. preferable
    * $\vee$ vs. optional
    * $\neg$ vs. prereferably not
  - Is there use for $=, <$ and other strict predicates in a test collection for XML retrieval?
    * There is mixed opinion on this one

- Path operations
  - child vs. descendant axis (/ vs. //)
    * It is not obvious that the child axis is necessary
    * We could probably live without it

- Multiple filters
  - We want to be able to express stuff like ...
  - `//article[about(.,''digital libraries'')]//*[about(., ``security'')]`
  - `//bdy//*[about(.,''solar powered robots'')]//fig[about(.,''robot on mars'')]`
- Equivalent tags
  - The equivalence are usually between textual elements
  - Is there a natural to make distinction between the textual elements?
  - Can we avoid tag equivalences by teaching people how to use the *-axis?

5 Assessing

- Should we submit different runs for each about predicate ?
- Then we can assess the about predicates on their own merit.
- We can then derive assessments for the full query.
- This is related to the notion of support element.

6 Query classification

- We should try to see if we can classify the queries by the IR aspect that they are testing.

7 XML collections

- Can be blame the collection, at least partly, for unnatural structural information needs?
- We should consider other collections if we can get our hands on them
  - Does the Lonely Planet XML collection exist?
  - Can we get more computer science articles, perhaps from somewhere else?.
  - There exists biology collections, but that would be tricky to assess.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley-Longman, 1999.

[2] M. Hearst. *User Interfaces and Visualization*, chapter 10. In [1], 1999.

[3] R. A. O'Keefe and A. Trotman. The simplest query language that could possibly work. In *INEX 2003 Workshop Proceedings*, pages 117–124, 2003.