Designing User Studies for XML Retrieval

10 Aug 2006

Miro Lehtonen

University of Helsinki

Outline

- Background & Introduction
- Two issues under investigation in the current user studies
  - XML elements and XML documents in XML retrieval
  - Queries with structural conditions
- XML retrieval systems and the users
- Relevant issues for future user studies

Three backgrounds & three paths to the research on XML Retrieval

- IR: a new domain/application area for existing methodology
- Databases: full-text queries, results ranked by relevance added to existing implementations
- Document engineering: support for search applications
  - Documentation enriched with metadata
  - “Structural hints” in queries
  - More opportunities for taking advantage of the XML structure
    - Indexing methods
    - Query evaluation
    - Ranking algorithms

Introduction

- Three kind of systems where the user studies on XML retrieval could be conducted
  - (XML) Search engines (most of these are experimental)
  - XML Databases (both experimental and commercial XDBMS's are common)
  - Document management systems (most are operational)
- Problems with INEX-related user studies
  - User studies are conducted on the experimental systems
  - Experimental setting leads to experimental results
    - The research gets sidetracked from the real issues
    - Low impact outside the research community

Issue #1: Documents or elements?

- *"XML retrieval is better than flat document retrieval because we can retrieve relevant elements in addition to whole documents."* ...right?
- Do current user studies compare XML documents with XML elements? ...not really.
  - In practice: sections, subsections, and paragraphs vs. entire articles
- Do the results generalise to content other than articles?
  - Unorganised fragments of documentation? Probably not.
  - XML documents smaller than a section or a paragraph?
    - The relevant answers may consist of several XML documents.
- Just a terminological problem?

Why users could not choose between XML Documents and XML Elements

- The storage units of content depend on the technical implementation
  - Any document collection may be implemented as 1-n XML documents
- The retrieved units always consist of XML elements (at least)
  - The search results are formatted when presented to the user
  - Users are not shown the source format (anything about XML)
- Users only see content: relevant or irrelevant, a lot or a little

It's the size that matters

- Can users judge the size of the answers?
  - They can most likely distinguish between 1) too small, 2) too big, and 3) good enough.
  - Maybe even between 2a) too big (to find the relevant content easily) and 2b) unnecessarily big (to read the entire answer)
- Can users judge that the size is "just right" ?
  - No, unless they are shown the source document (the "context").
  - Only "assessors" who judge the exhaustivity of each XML element can determine which elements are the right size.
- Questions, comments?

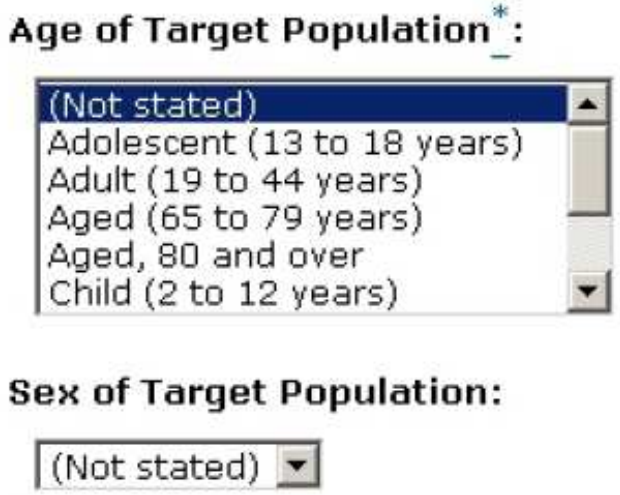
Issue #2: XML in the queries

- *"XML retrieval is better than flat document retrieval because the queries can include conditions on the document structure."* ...right?
  - Out of many XML query languages, XQuery is the most common
- Do current (user) studies compare "Content-only" queries to ones with structural constraints? Yes, but
  - ...only with queries where the structure is irrelevant.
  - ...only with queries where the "structural hints" are not about the content.
  - ...only with queries written by users.
- Do the results generalise?
  - Not to queries on documents where the structure describes the content.
  - Not to queries with structural requirements (not just hints).
  - To a document collection?

The kind of structural structural conditions that help XML Retrieval

- Constraints on any content where the tag names make a difference
  - Data (as in databases)
  - Content that comes with metadata (XML!)
  - For example, the content of the elements `complaint`, `diagnosis`, and `treatment`
  - Content that the tag name disambiguates
- Constraints that are added to the query by the application, e.g. the user interface
  - Even the “bad users” can specify good “structural hints”
- More?

An example of input fields corresponding to the DTD




```
<!ELEMENT target.population (#PCDATA | eligibility | age | sex )*>
<!ELEMENT eligibility (#PCDATA | inclusion.criterion |
          exclusion.criterion | %block; )*>
<!ELEMENT age (#PCDATA | %block; )*>
<!ELEMENT sex (#PCDATA | %block; )*
```

User studies & queries with structure

- Can we study whether and when the structural constraints help?
  - Yes, but don't we know the answer already?
- Can we study how good users are at specifying the queries with structure?
  - Yes, but doesn't it depend on the user interface?
- Can we study which query languages are easy to learn?
  - Yes, unless we want to focus on end-users.
- So what can we study? ..user interface design?
- User studies on Content-And-Structure type queries have a zero impact on anything but research!
- Questions, comments?

The users of XML (Element) Retrieval

- Lots of confusion: lost, found, still wanted... users of... what?
- Features of the systems that qualify:
  - XML documents are indexed for full-text search - required
  - Flexible search granularity (not only whole documents) - optional
  - Support for queries with structure - optional
  - The users are not aware of searching XML documents!
- Typical features of the experimental systems
  - Search results consist of XML elements
  - Document formats other than XML are not supported

Operational systems: the case of Elsevier

- Scopus: Over 12,000 academic journals
  - Tens of thousands of users (how big is a representative sample ?)
  - XML does not show
  - Search filters on subjects, document types, authors, etc.
- MarkLogic XML content server with XQuery-based search
  - Element-level search granularity
  - Support for passage retrieval
  - User interfaces for entering queries are specific to each implementation

Relevant topics for future user studies

- Size (or granularity) of the answers
  - Applies to standalone-type answers only
  - Results generalise to other areas of IR
- User interfaces for XML retrieval
  - How many input fields? What kind of input fields?
  - Presentation of the search results
    - Links to the original source documents
    - Relevant parts reassembled into new documents
- Is XML retrieval superior to other forms of IR?
- Do users prefer searching XML documents to other documents?

Conclusions

- Current user studies do not have representative samples of
  - Different types of XML documents or queries
  - Different systems for XML retrieval
- The results do not generalise!
- Good news: studies are conducted on data that exists in real applications
- Questions, Comments?