

# Structural Features in XML Retrieval

Georgina Ramírez  
CWI, Amsterdam, The Netherlands  
georgina@cwi.nl

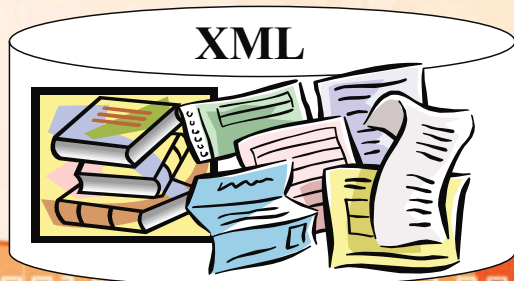


## XML-IR: focused search



**TIJAH!** XML-IR

Barcelona public transport



**Barcelona**  
Barcelona is the capital of Catalonia and is the second largest city in Spain. It is located on the Mediterranean coast ...

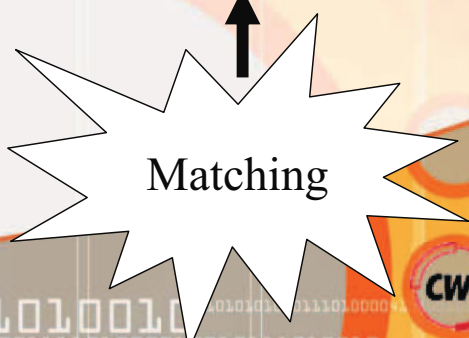
**History**  
The foundation of Barcelona is the subject of two different legends. The first attributes the founding of the city to Hercules 400 years before the building of Rome, ...

**Geography**  
Barcelona is located on the northeast coast of the Iberian Peninsula, facing the Mediterranean sea, on a plateau ...

**Transport**  
Air ... Sea ... Rail ...

**Public transport**  
Barcelona is served by a comprehensive local public transport network that includes a metro, two separate tram networks, a bus network and several funiculars and aerial cable cars. The Barcelona Metro network comprises nine lines, identified by an "L" followed by the line number as well as by individual colours.

**Culture**  
Barcelona's culture is rich, stemming from the city's 2000 years of history. To a greater extent than the rest of Catalonia, ...



# What's new in XML-IR?

```

<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
  <journal> IEEE ... </journal>
</front matter>

<body>
  <section1> <section title> Introduction</section title>
  <par> We present a compression algorithm for ... </par>
  <par> We discuss our compression algorithm in Section 2 </par>

  <section2> <section title> Compression Algorithm </section title>
  <par> Our approach consists of .... </par>
  <par> The compression algorithm is depicted in Figure 1.
    <figure1><img>...</img>
    <figure caption> Compression algorithm </figure caption>
  </par>
</section2>
</body>

<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>
    
```

- User defined markup
  - +Heterogeneity, +semantics
- Different types of markup
  - *Procedural and descriptive*
- Hierarchical structure (XML tree)
  - Explicit relationships between parts of documents.
- Explicitly marked up retrieval units
  - Easier focused retrieval?
- Users can query on structure
  - "I want articles or references to articles on compression algorithms".
  - More freedom?



# What's new in XML-IR?

```

<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
  <journal> IEEE .
</front matter>

<body>
  <section1> <section title> Intro
  <par> We present
  <par> We discuss

  <section2> <section title> Con
  <par> Our approa
  <par> The compre
  <figure1><img>
  <figure cap
  </figure1>

  </section2>
</body>

<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>
    
```

Many types of *Structural Features* providing extra source of information but...

**Are they useful for (focused) retrieval?**

- User defined markup
  - +Heterogeneity, +semantics
- Different types of markup
  - *Procedural and descriptive*
- Hierarchical structure (XML tree)
  - Explicit relationships between parts of documents.
- Explicitly marked up retrieval units
  - Easier focused retrieval?
- Users can query on structure
  - "I want articles or references to articles on compression algorithms".
  - More freedom?

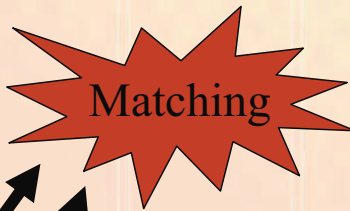


# IR tendencies



**TIJAH!** XML-IR

User query



- Users are different and have different intentions:
  - Query-specific retrieval.
  - Context-aware retrieval.
- Combination of different sources of evidence can improve search effectiveness.

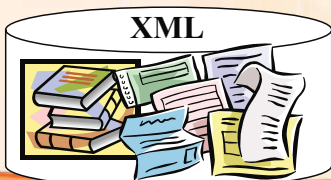


# IR tendencies



**TIJAH!**

User query



**H1** Different search tasks and contextual situations require different use of structural information.

**H2** The more structural evidence we can collect and combine, the better.

- Users are different and have different intentions:
  - Query-specific retrieval.
  - Context-aware retrieval.
- Combination of different sources of evidence can improve search effectiveness.



# Research outline

- Definition of a “pragmatic” retrieval framework where it is possible to combine different structural information and where evidence and parameters can be tuned according to user’s intentions and contextual factors.
- Experimental approaches on three main aspects:
  - New ways of using structure: element’s context information.
  - Use of structural features for relevance feedback.
  - Study of correlations between search tasks and contextual features and structural characteristics of relevant information.
- Evaluation using INEX benchmark.

# This talk

- Definition of a “pragmatic” retrieval framework where it is possible to combine different structural information and where evidence and parameters can be tuned according to user’s intentions and contextual factors.
- Experimental approaches on three main aspects:
  - New ways of using structure: element’s context information.
  - Use of structural features for relevance feedback.
  - Study of correlations between search tasks and contextual features and structural characteristics of relevant information.
- Evaluation using INEX benchmark.

# Evaluation

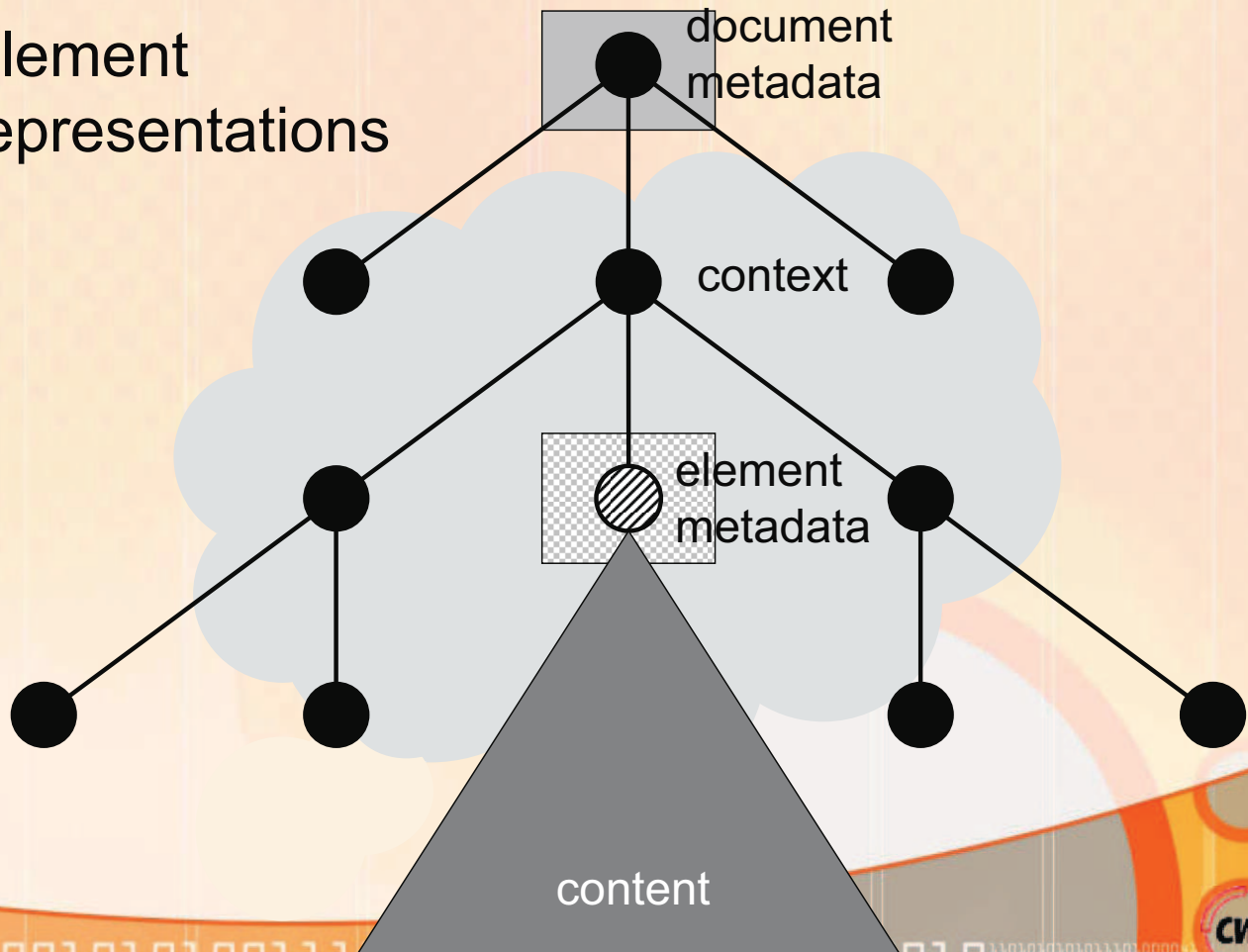
- INEX 2005 collection:
  - 16.819 scientific articles from 24 journals (1995-2004).
- Relevance judgments:
  - Element level.
  - 2 dimensions: Exhaustivity [0,1,2] and Specificity [0...1].
- Tasks:
  - Content-Oriented XML retrieval: keyword based queries.
  - Thorough: “Find *all* relevant elements”.
  - Focused: “Find the *most* exhaustive and specific elements on a *path*.”
- Evaluation Metrics:
  - nxCG and MAep.
  - Two quantizations: Strict / Generalized.

## Element's context Supporting relevance

**SIGIR'06:** G. Ramirez, T. Westerveld, and A.P de Vries, “Using Small XML Elements to Support Relevance”.

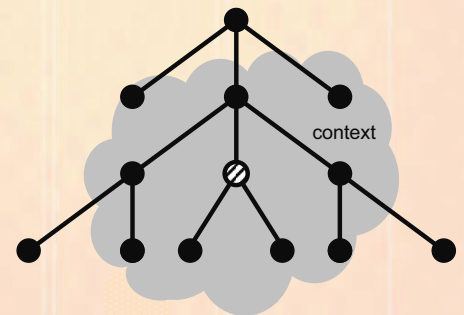
**FQAS'06:** G. Ramirez, T. Westerveld, and A.P de Vries, “Using Structural Relationships for Focused XML Retrieval”.

## Element representations



## Element's context

- Common approaches:
  - 2 main types:
    - Element types: e.g., **articles**, abstracts, titles,...
    - XML Tree Relationships: ancestors, descendants,...
  - Good but too general.
- Element type specific approach:
  - Using “unwanted” elements as context.



# “Unwanted” ?

- XML Retrieval:
  - Undefined retrieval unit.

```

<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
</front matter>
<body>
  <section1><section title> Introduction</section title>
    <paragraph> We present a compression algorithm for ... </paragraph>
    <paragraph> We discuss our compression algorithm in Section 2. </paragraph>
  </section1>
  <section2><section title> Compression Algorithm </section title>
    <paragraph> Our approach consists of .... </paragraph>
    <paragraph> The compression algorithm is depicted in Figure 1.
      <figure1><img>...</img>
      <figure caption> Compression algorithm </figure caption>
    </figure1>
    </paragraph>
  </section2>
  <section3>
    <section title> Conclusions </section title>
    <paragraph> We presented our compression algorithm.</paragraph>
    <paragraph> Our compression algorithm performs very well.</paragraph>
  </section3>
</body>
<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>

```

```

<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
</front matter>
<body>
  <section1><section title> Introduction</section title>
    <paragraph> We present a compression algorithm for ... </paragraph>
    <paragraph> We discuss our compression algorithm in Section 2. </paragraph>
  </section1>
  <section2><section title> Compression Algorithm </section title>
    <paragraph> Our approach consists of .... </paragraph>
    <paragraph> The compression algorithm is depicted in Figure 1.
      <figure1><img>...</img>
      <figure caption> Compression algorithm </figure caption>
    </figure1>
    </paragraph>
  </section2>
  <section3>
    <section title> Conclusions </section title>
    <paragraph> We presented our compression algorithm.</paragraph>
    <paragraph> Our compression algorithm performs very well.</paragraph>
  </section3>
</body>
<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>

```

```

<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
</front matter>
<body>
  <section1><section title> Introduction</section title>
    <paragraph> We present a compression algorithm for ... </paragraph>
    <paragraph> We discuss our compression algorithm in Section 2. </paragraph>
  </section1>
  <section2><section title> Compression Algorithm </section title>
    <paragraph> Our approach consists of .... </paragraph>
    <paragraph> The compression algorithm is depicted in Figure 1.
      <figure1><img>...</img>
      <figure caption> Compression algorithm </figure caption>
    </figure1>
    </paragraph>
  </section2>
  <section3>
    <section title> Conclusions </section title>
    <paragraph> We presented our compression algorithm.</paragraph>
    <paragraph> Our compression algorithm performs very well.</paragraph>
  </section3>
</body>
<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>

```

```

<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
</front matter>
<body>
  <section1><section title> Introduction</section title>
    <paragraph> We present a compression algorithm for ... </paragraph>
    <paragraph> We discuss our compression algorithm in Section 2. </paragraph>
  </section1>
  <section2><section title> Compression Algorithm </section title>
    <paragraph> Our approach consists of .... </paragraph>
    <paragraph> The compression algorithm is depicted in Figure 1.
      <figure1><img>...</img>
      <figure caption> Compression algorithm </figure caption>
    </figure1>
    </paragraph>
  </section2>
  <section3>
    <section title> Conclusions </section title>
    <paragraph> We presented our compression algorithm.</paragraph>
    <paragraph> Our compression algorithm performs very well.</paragraph>
  </section3>
</body>
<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>

```

```

<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
</front matter>
<body>
  <section1><section title> Introduction</section title>
    <paragraph> We present a compression algorithm for ... </paragraph>
    <paragraph> We discuss our compression algorithm in Section 2. </paragraph>
  </section1>
  <section2><section title> Compression Algorithm </section title>
    <paragraph> Our approach consists of .... </paragraph>
    <paragraph> The compression algorithm is depicted in Figure 1.
      <figure1><img>...</img>
      <figure caption> Compression algorithm </figure caption>
    </figure1>
    </paragraph>
  </section2>
  <section3>
    <section title> Conclusions </section title>
    <paragraph> We presented our compression algorithm.</paragraph>
    <paragraph> Our compression algorithm performs very well.</paragraph>
  </section3>
</body>
<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>

```

```
<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
</front matter>
<body>
  <section1><section title> Introduction</section title>
    <paragraph> We present a compression algorithm for ... </paragraph>
    <paragraph> We discuss our compression algorithm in Section 2. </paragraph>
  </section1>
  <section2><section title> Compression Algorithm </section title>
    <paragraph> Our approach consists of .... </paragraph>
    <paragraph> The compression algorithm is depicted in Figure 1.
      <figure1><img>...</img>
      <figure caption> Compression algorithm </figure caption>
    </figure1>
    </paragraph>
  </section2>
  <section3>
    <section title> Conclusions </section title>
    <paragraph> We presented our compression algorithm.</paragraph>
    <paragraph> Our compression algorithm performs very well.</paragraph>
  </section3>
</body>
<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>
```

# Problem

- XML Retrieval:
  - Undefined retrieval unit.
- Many unwanted elements:
  - Too small (*10.3 out of 11.4 Million in the INEX collection*)
  - Unwanted element types
- Standard Retrieval Models:
  - Small elements are ranked high.
  - Result lists contain overlapping elements.

# Length Normalization in XML Retrieval

- **Removing small elements:**
  - Post-filtering result list / Remove at indexing time.
  - Good for high initial precision. Low recall.
- **Defining a subset of retrievable units:**
  - E.g: {article, sections, paragraphs, abstracts}.
  - Requires knowledge of the documents' structure.
- **Length priors:**
  - Prior of relevance according to element's length.
  - Re-ranking of elements -> Low early precision.
  - Diminishes the effect of other XML-IR techniques when combined.
  - Length distribution might be different for each task.

## Do not ignore the small ones! (Use small elements' relevance)

- **Goal:** Reward longer elements in a "content-oriented" manner, not only by its length.
- **Main idea:** Use of small retrieved elements to locate larger, relevant ones:
  - Create relationships/links between small elements and other elements.
  - Use these relationships as element's context information and propagate small elements' score to larger elements before removing them.

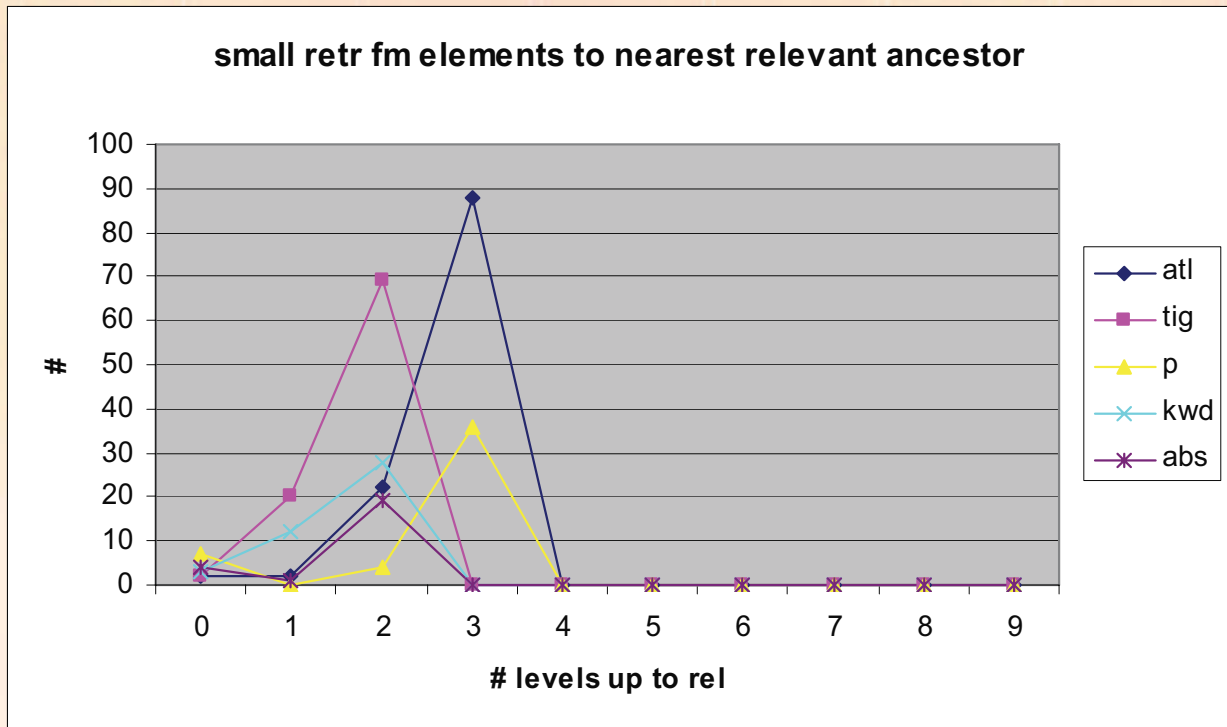
# Links between small and relevant elements

- Some are quite intuitive:
  - section title -> section.
  - Italics -> containing element.
- Author or publisher of documents could give good hints.
- Not necessarily following XML tree structure (e.g. citations).

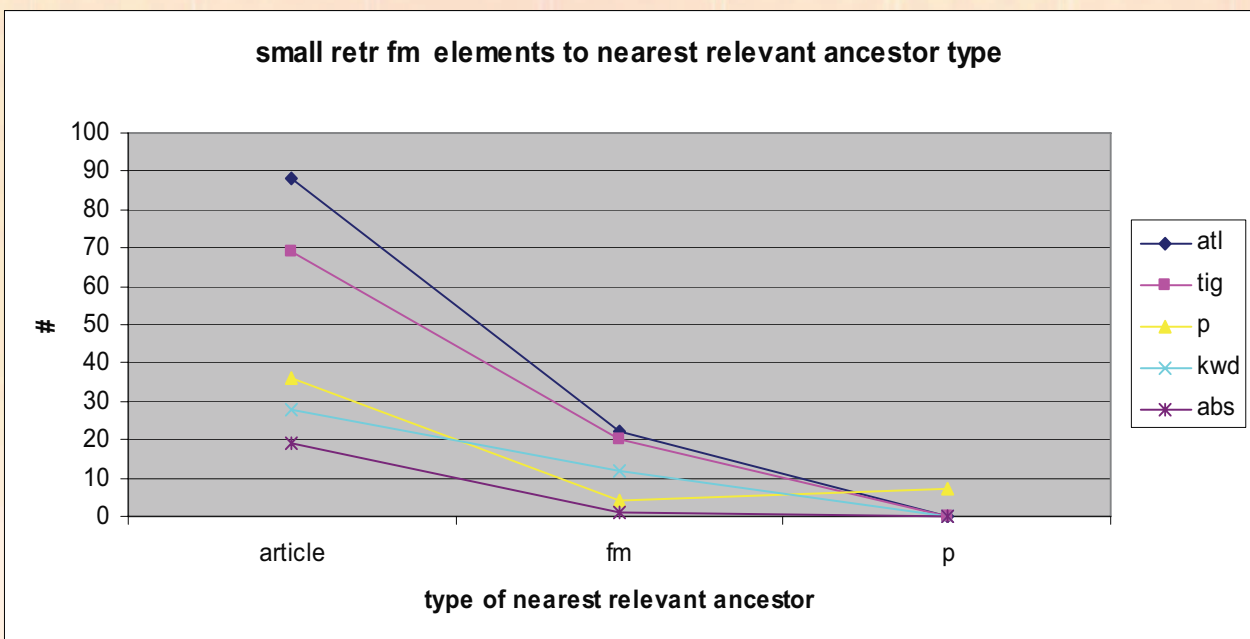
# Analyzing data...

- Relationships between small retrieved (top 1000) and relevant ones:
  - # levels up until reaching a relevant element.
  - Type of the closest relevant element.
- Distinction between FM, BDY and BM.

# FM – number levels up



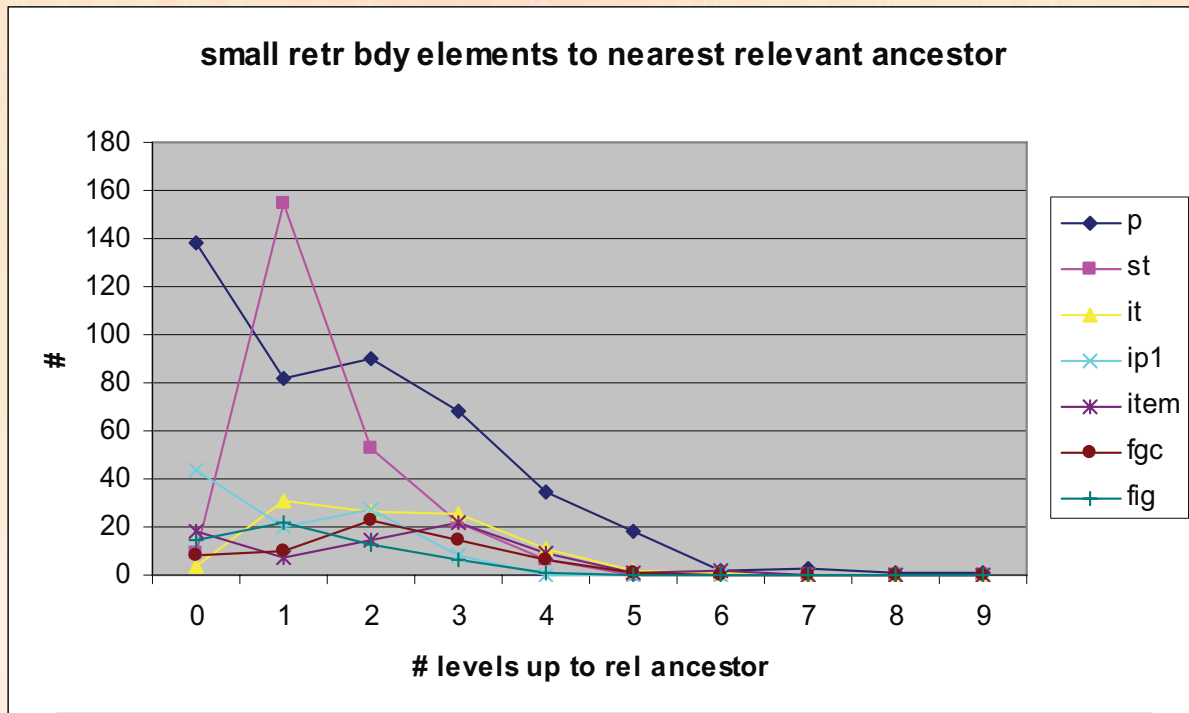
# FM – relevant element types



Article titles, abstracts and other paragraphs often relate to relevant articles.



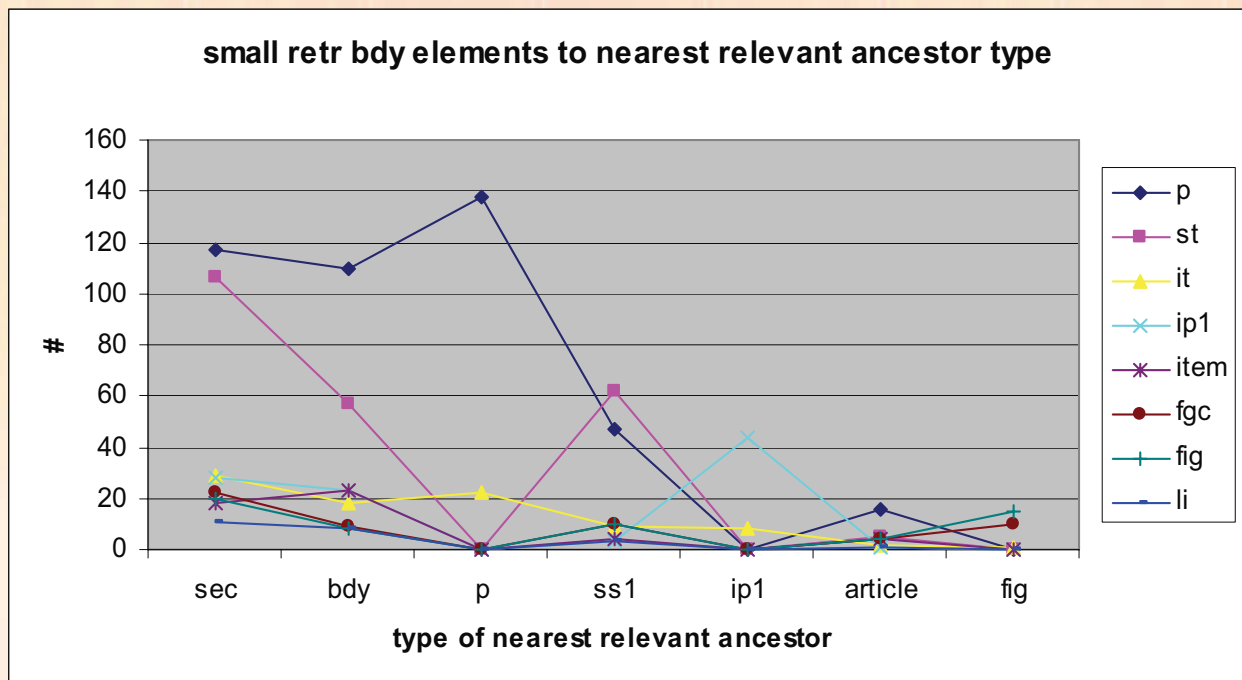
# BDY – number levels up



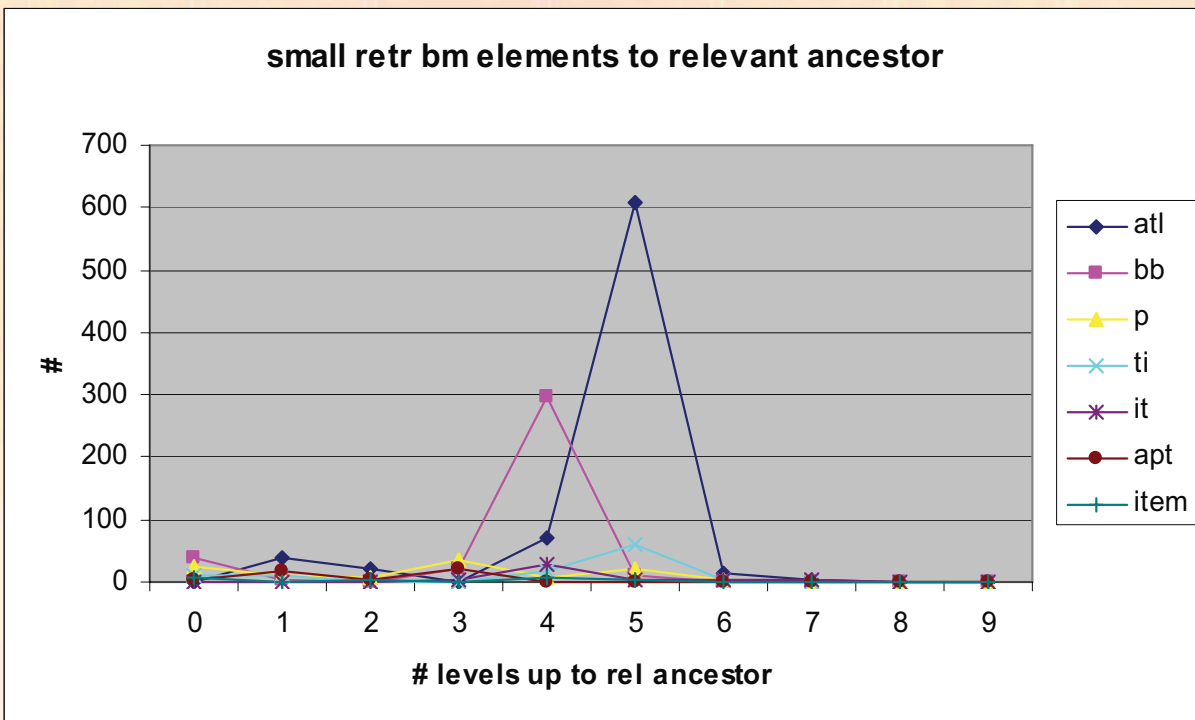
Italics, figures and section titles have often a relevant parent.  
 Figure captions are two levels down from a relevant element.



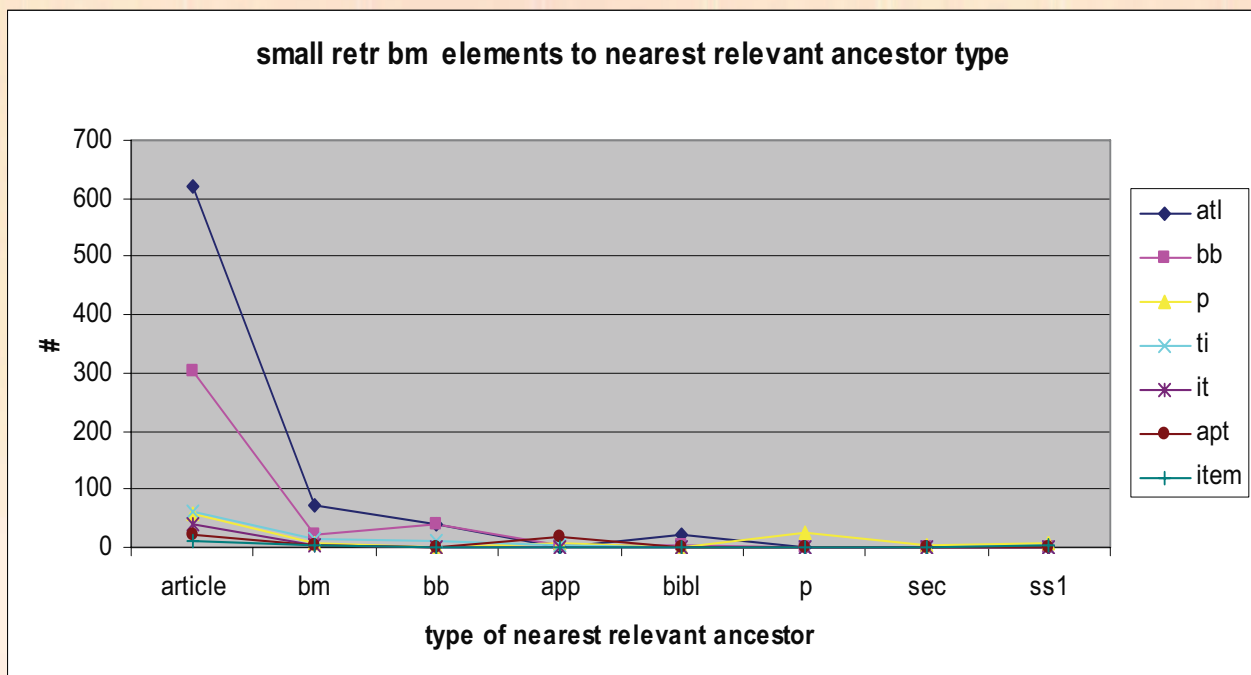
# BDY – relevant element types



# BM – number levels up

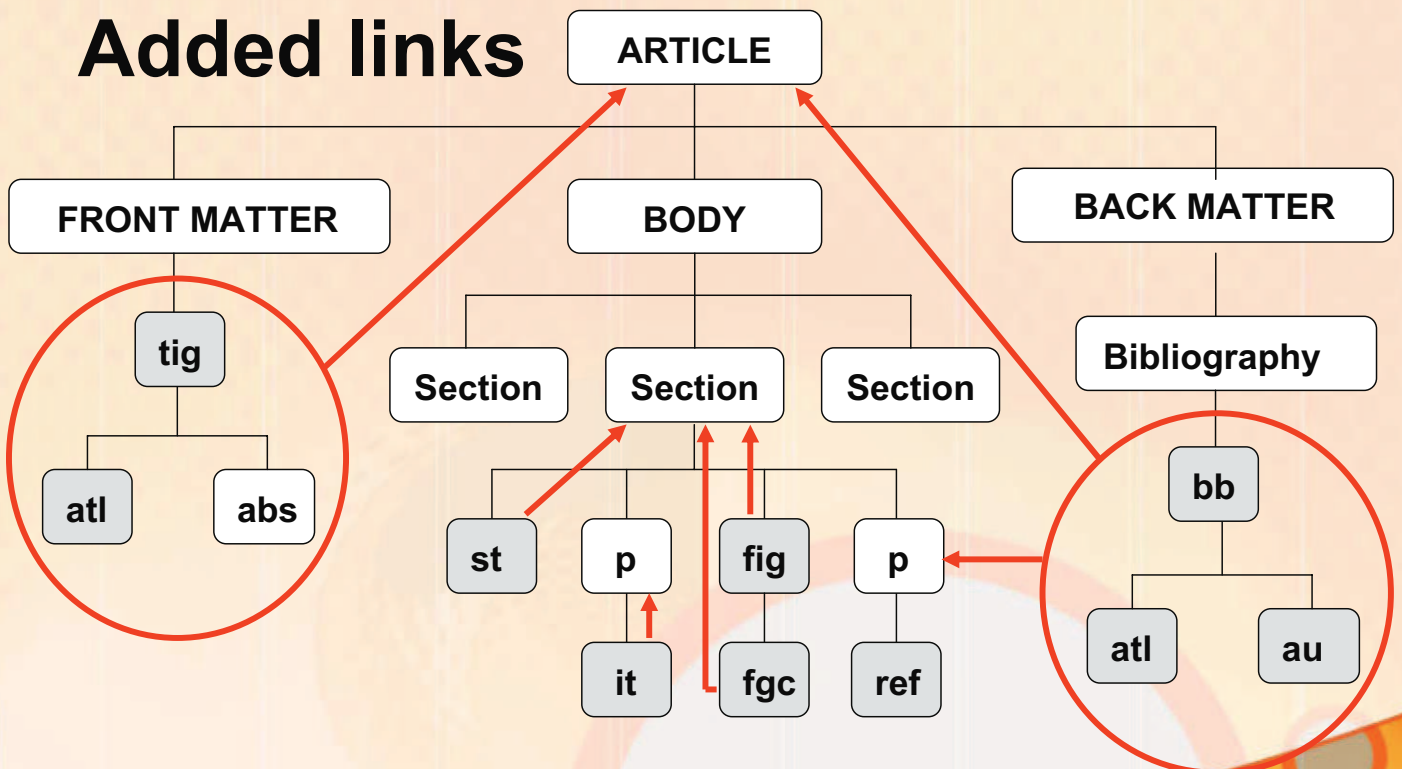


# BM – relevant element types



All small elements have as closest relevant element the article.





$$P(E_j) = cf(P_{RM}(E_j), aggr(i \in inlinks)(P_{RM}(i)))$$

## Experiments

- Individual and global contribution of using different number and types of links.
- Effect of combining the approach with other XML-IR techniques.
- Generalization of the approach when performing a different task.

# Findings

- Precision at several recall levels is improved significantly (over removing small elements baseline, best one for precision).
- Best performing element types are section titles and italics in body part.
- Aggregating scores with MAX better than AVG.
- Main gain obtained when using body links or body and fm links.
- When combined with article weighting recall is increased significantly without losing too much precision. A good tradeoff can be obtained.
- When removing overlapping elements (for a focused task), it performs better than any baseline.

# Conclusions

- Traditional approaches on length normalization in XML Retrieval are useful but they ignore important contextual clues that small elements can bring.
- We presented a novel way of dealing with small unwanted elements in XML Retrieval.
  - The approach performs well across tasks and combined with other XML+IR techniques.
  - The method can be used when the structure of the documents is not known, on any XML collection.
- Element's context is an important source of information. XML elements should not be treated independently.
- Future work:
  - Use the discovered link information in relevance feedback (e.g. query expansion).

# Search tasks and Contextual Features

**HiX'06:** G. Ramirez and A.P. de Vries, "Relevant Contextual Features in XML Retrieval."

**IRIX'05:** G. Ramirez and A.P. de Vries, "XML and Context: Structural Features Relevant to Search Tasks".

Study of the dependencies between different contextual features and the **structural characteristics** of the relevant components.

*Can we identify a measurable dependency between a topic's contextual factors and the structural characteristics of the topic's relevant components?*

```

<front matter>
  <title> A Novel Compression Algorithm </title>
  <author> A. One </author>
</front matter>
<body>
  <section1><section title> Introduction</section title>
    <paragraph> We present a compression algorithm for ... </paragraph>
    <paragraph> We discuss our compression algorithm in Section 2. </paragraph>
  </section1>
  <section2><section title> Compression Algorithm </section title>
    <paragraph> Our approach consists of .... </paragraph>
    <paragraph> The compression algorithm is depicted in Figure 1.
      <figure1><img>...</img>
      <figure caption> Compression algorithm </figure caption>
    </figure1>
    </paragraph>
  </section2>
  <section3>
    <section title> Conclusions </section title>
    <paragraph> We presented our compression algorithm.</paragraph>
    <paragraph> Our compression algorithm performs very well.</paragraph>
  </section3>
</body>
<back matter>
  <bib><bb id="BIBA400761">
    <au> A. Two </au>
    <atl> Compression algorithms </atl>
  </bb></bib>
</back matter>

```

Structural characteristics:  
Element type, Size,  
location, article, journal, ...

- Analysis of data from a user study designed to investigate how persons interact with **components of XML documents**.
- Interactive Track at INEX 2005 (Larsen, Malik and Tombros).

# iTrack Experiments

- Participants:
  - 73 test persons (20 nationalities) - 29 % female - 71 % male.
  - Ages: 19 – 52 (avg. 28). 60% students, 12 % PhD's, 18 % academia, 10 % other.
- Tasks:
  - 2 simulated tasks (1 out of 3 possibilities)
  - 1 information need of their choice :
    - What are you looking for?
    - What is the motivation of the topic? (Why, Problem, Context)
    - What would an ideal answer look like?
- Procedure:
  - 20 minutes to perform each task (task order permuted).
  - 5 types of components: articles, article's metadata (fm), sections (sec), subsections (ss1), and sub-subsections (ss2).
  - Asked (but not forced) to do assessments.
    - Relevant / Partially relevant / Not relevant.

# Analysis

- 68 user formulated tasks.
- Contextual features:
  - Searcher's familiarity with the topic.
  - Type and specificity of the request.
  - User's intention.
- Structural features:
  - Number and type of relevant elements.
  - Number of different articles / journals containing relevant information.

# Searcher's familiarity.

- Recorded in the questionnaires.
- Conversion from 1-5 scale to:
  - No (1-2).                       $\longrightarrow$       8 users
  - Somewhat (3).                 $\longrightarrow$       26 users
  - Yes (4-5).                      $\longrightarrow$       34 users

# Type & specificity of the request

- “What are you looking for (what do you want the system to find)?”
- Two dimensions:
  - Specificity: Broad – Narrow (topically or structurally).
  - Complexity: Simple – Compound (e.g. A and B, technique A in field B).

Complexity	Specificity	Num.	Example
Simple	Broad	12	I search information about web services.
	Narrow	10	I am looking for <b>introductions</b> to Data Mining.
Compound	Broad	20	<b>Papers</b> about 'named entity recognition' <b>and</b> 'clause boundary recognition'.
	Narrow	12	Decidability <b>and</b> complexity <b>results</b> of (bounded/live) Petri Nets.
	Broad + Narrow	14	I want information about web standards <b>and</b> W3Cs role in implementing these in various web browsers.

# User's intention

- Classification according to description.

*“What is the motivation for the topic (why are you looking for this, what problem can be solved with the information, and in what context did the problem arise)?”*

- 5 categories:
  - **Decide** : make decision and/or compare.
  - **Apply** : use in a practical way (design problem).
  - **Explain** : transfer knowledge.
  - **Study** : learning.
  - **Personal interest** : curiosity, general interest, ...

# User's intention

Class	Num.	Example
<b>Apply (A)</b>	9	My computer was upgraded by a friend. I have the state of the art anti virus. Yet the worms keep coming. <b>I want to know what to do.</b>
<b>Decide (CD)</b>	6	The department is <b>trying to decide</b> whether to release a produced Linux-program under a public license such as GNU or GPL.
<b>Explain (E)</b>	10	I am <b>writing an article</b> about the history of information systems and the projections and expectations made by experts when they were introduced.
<b>Study (S)</b>	13	I am <b>taking a course</b> in networks, and <b>want to know more</b> . The literature we used didn't give the right information.
<b>Personal Interest (PI)</b>	6	Just out of <b>general interest</b> . I would like to know, for instance, when spamming was first acknowledged as a problem.

# Structural Features

- Number and type of relevant elements
  - 5 types (articles, fm, sec, ss1, ss2)
  - Only the **relevant** ones.
- Number of different articles / journals containing relevant information.

## Relevance Overview

956 elements assessed.

14,1 elements per task.

Relevance:

31% relevant

34% partially relevant

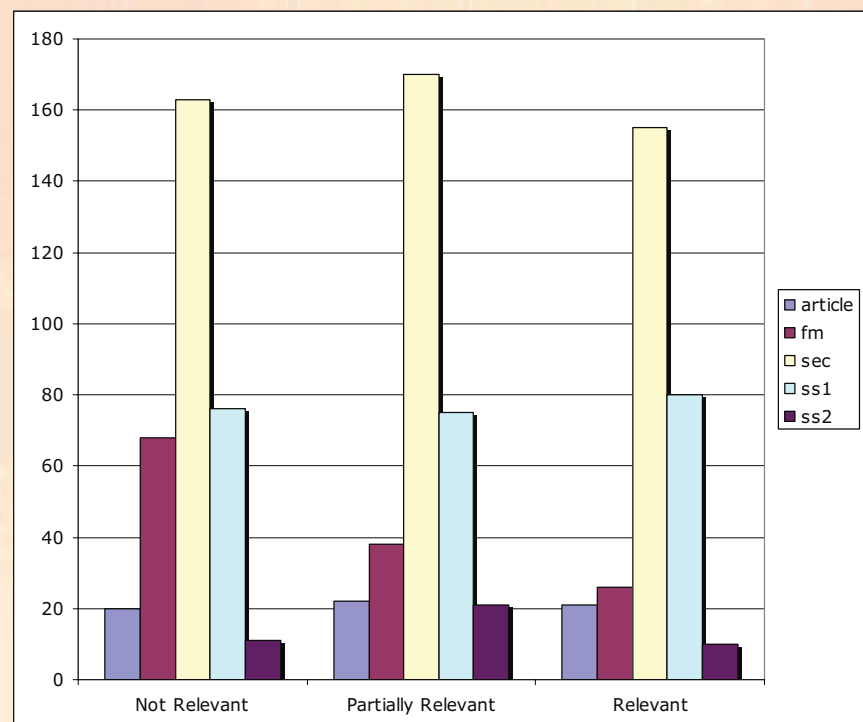
35 % not relevant.

Relative numbers:

Articles: 33%

Sections: 32%

Subsections: 35%



# Findings - Familiarity

- Request type
  - The more familiar the user is with the topic, the more *Compound* tasks are performed.
  - Not knowledgeable users performed mainly *Broad* tasks.
- Intention
  - 65 % of the tasks performed by knowledgeable users were *Study* (S) or *Explain* (E) tasks.
  - The less informed users performed only PI, A and CD tasks.
- Relevant types
  - The more familiar the user is with the topic, the more the metadata is found useful.

# Findings - Request type

- Number relevant elements
  - On average, more relevant elements were found by the users performing *Simple* tasks
- Number relevant journals
  - More relevant journals were found by the users performing *Broad* or *Simple* tasks.
- Relevant types
  - Articles and metadata are more useful for *Broad* tasks.

# Findings - User's intent

- Number relevant elements
  - The users searching for *personal interest* found, on average, more relevant information.
- Number relevant articles / journals
  - The users with *Explain* tasks found less relevant articles and those with *personal interest* tasks the most.
  - The users that had to compare are the ones that found, on average, less relevant journals.
- Relevant types
  - Users searching for *personal interest* or to *apply* the knowledge found the representations of articles (articles and metadata) the less useful.
  - Users of the *Explain* and *Study* categories are the ones that found the metadata information more useful.

# Conclusions

- General behavior:
  - Sections and sub-sections are the most relevant.
  - Very few articles and journals contain relevant information.
- Familiarity with the topic can be a good indicator of type of search and intention.
- Task type and intention give hints of which types of elements users prefer to see.
- Several tendencies can be seen but numbers are small and differences are not statistically different..

# Final conclusions

- Structural features of XML documents can be further exploited to improve retrieval effectiveness.
  - XML elements should not be treated independently.
- Correlations between contextual information and structural features indicate that XML retrieval systems can make use of structural information to adapt to different search task and user's context.



# Thanks

