**Semantic Structure in Structured Document Retrieval**

*semantic \(\approx\) meaningful*

Roelof van Zwol, Leen Breure, Tim van Loosbroek
roelof@cs.uu.nl
Utrecht University
The Netherlands

Objectives:

- **Focus in structure document retrieval:**
  (≈XML Retrieval)

  - exploiting the available **structural** information in documents to implement a more **focused** retrieval strategy and return **document components**, the so-called XML elements - instead of complete documents - in response to a **user query**.

  [INEX06]
Motivation

- Can we effectively use the structure of XML documents to enhance retrieval performance?
  - Impact on all facets: query formulation, retrieval strategy, result presentation
  - Information need:
    - Content-only, Content and Structure (NEXI query language)
    - Interpretation of structure: vague (hints) vs. strict

Motivation

- Types of structure:
Hypothesis

- Will (automatically derived) semantic structure lead to higher retrieval performance,
  1. If only keyword-based search is used?
  2. If the user is aware of the semantic structure, and uses it in his/her request?

Reuters - towards a meaningful structure

- Automated detection and annotation of named entities based on:
  - A set of regular expressions
  - (Con)textual clues:
    - Ltd, Corp, Minister, President, ... of State, Organization, county.
  - Dictionaries, gazetteers, etc.
  - Negative lookups
- Named entities: person, company, organization, location, keyword...
- Alternatives:
  - Use the available categories?
  - Existing tools, such as Lingpipe, Gate, DiasDem
Example

A leading member of Britain's opposition Labour Party said there was strong evidence Prime Minister Margaret Thatcher approved the sale of anti-aircraft missiles to Nicaraguan Contra rebels during talks last year with U.S. Officials involved in the Iran arms scandal. 

Roelof van Zwol - Utrecht University - The Netherlands

Reuters XML DC

Based on Reuters21587 collection

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td># of documents:</td>
<td>20841</td>
<td></td>
</tr>
<tr>
<td># of unique terms:</td>
<td>53080</td>
<td></td>
</tr>
<tr>
<td># of nodes:</td>
<td>209135</td>
<td>427854</td>
</tr>
<tr>
<td>avg. leaf size:</td>
<td>23.6</td>
<td>10.6</td>
</tr>
<tr>
<td>max node depth:</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>avg. node depth:</td>
<td>2.6</td>
<td>3.42</td>
</tr>
<tr>
<td># of unique node names</td>
<td>6</td>
<td>11</td>
</tr>
</tbody>
</table>

Roelof van Zwol - Utrecht University - The Netherlands
### Reuters XML DC

<table>
<thead>
<tr>
<th>Node name</th>
<th>Original</th>
<th>Semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td>article</td>
<td>20841</td>
<td></td>
</tr>
<tr>
<td>content</td>
<td>19043</td>
<td></td>
</tr>
<tr>
<td>paragraph</td>
<td>108682</td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>20840</td>
<td></td>
</tr>
<tr>
<td>dateline</td>
<td></td>
<td>19041</td>
</tr>
<tr>
<td>person</td>
<td>-</td>
<td>24508</td>
</tr>
<tr>
<td>company</td>
<td>-</td>
<td>52510</td>
</tr>
<tr>
<td>keyword</td>
<td>-</td>
<td>108682</td>
</tr>
<tr>
<td>organization</td>
<td>-</td>
<td>1096</td>
</tr>
<tr>
<td>location</td>
<td>-</td>
<td>31923</td>
</tr>
</tbody>
</table>

### Experimental Setup

- **2 Document Collections:**
  - Reuters original vs. semantic
- **15 Topics**
**Experimental Setup**

**Topic: 14**

**Description:** Find out what the connection is between IBM and Intel

**Narrative:** IBM decided to use Intel’s processor chips inside their PCs.

**CO:**
- IBM Intel

**CAS - original:**
- //paragraph[about(._, IBM Intel)]

**CAS - original+:**
- //paragraph[about(._, IBM) and about(._, Intel)]

**CAS - semantic:**
- //paragraph[about(../company, IBM) and about(../company, Intel)]

---

**Experimental Setup**

- 2 Document Collections:
  - Reuters original vs. semantic
- 15 Topics
- 2 Systems
  2. [no_name_yet] (*INEX-2006*)
- 2 sets of relevance judgments
  - *General* - any element marked relevant
  - *Strict* - subset of general, consisting of the target elements of the CAS variant.
Results - CO

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Semantic</th>
<th>Semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>No minimum node size</td>
</tr>
<tr>
<td>topics</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>retrieved</td>
<td>1500</td>
<td>1500</td>
<td>1500</td>
</tr>
<tr>
<td>relevant</td>
<td>1424</td>
<td>1424</td>
<td>1424</td>
</tr>
<tr>
<td>rel_retr</td>
<td>665</td>
<td>716</td>
<td>639</td>
</tr>
<tr>
<td>map</td>
<td>0.3418</td>
<td>0.3867</td>
<td>0.3134</td>
</tr>
<tr>
<td>bpref</td>
<td>0.5558</td>
<td>0.5846</td>
<td>0.5437</td>
</tr>
</tbody>
</table>

Statistical significant differences based on map (T-test):

Original vs Semantic: $t(15) = -3.090, p<0.01$
Results - CO

![Graph showing MAP vs Topic for Original and Semantic conditions]

Results - CAS

<table>
<thead>
<tr>
<th></th>
<th>ORGORG</th>
<th>ORG+onORG</th>
<th>ORG+onSEM</th>
<th>SEMonSEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>topics</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>retr.</td>
<td>1500</td>
<td>943</td>
<td>1010</td>
<td>655</td>
</tr>
<tr>
<td>relevant</td>
<td>515</td>
<td>515</td>
<td>515</td>
<td>515</td>
</tr>
<tr>
<td>rel_retr</td>
<td>327</td>
<td>322</td>
<td>380</td>
<td>353</td>
</tr>
<tr>
<td>map</td>
<td>0.4468</td>
<td>0.4624</td>
<td>0.4963</td>
<td>0.6378</td>
</tr>
<tr>
<td>bpref</td>
<td>0.6807</td>
<td>0.687</td>
<td>0.7702</td>
<td>0.7182</td>
</tr>
</tbody>
</table>

Statistical significant differences based on map (T-test):

- ORGonORG vs SEMonSEM: $t(15)=-3.015$, $p<0.010$
- ORG+onORG vs SEMonSEM: $t(15)=-2.817$, $p<0.015$
- ORG+onSEM vs SEMonSEM: $t(15)=-2.475$, $p<0.030$
Results - CAS

![Graph showing precision and recall for different methods]

Roelof van Zwol - Utrecht University - The Netherlands

Results - CAS

![Graph showing precision and recall at element cut off]

Roelof van Zwol - Utrecht University - The Netherlands
Roelof van Zwol - Utrecht University - The Netherlands

Results - CAS

![Graph showing MAP scores for different categories across topics]

Roelof van Zwol - Utrecht University - The Netherlands

Concluding Remarks

- Replicate the experiment on a larger scale:
  - Reuters RCV1 collection
  - Larger topic set
  - More systems
  - INEX 2006 - Wikipedia collection?

- Differentiate in types of semantic tagging:
  - Small elements - named entities
  - Large elements - appear in top of the XML tree structure

- Alternative tagging tools: Lingpipe, Gate, etc.
But, based on the results presented here, we can conclude that:

- Enrichment of XML structure (with semantics) is beneficial for both keyword-based and content-and-structure queries.

- **Semantic** structure makes SDR **meaningful**.

**Usage:** Oil companies are *exploring* scientific documents, to find new information about potentially interesting places.