

# *Semantic* Structure in Structured Document Retrieval

*semantic ≈ meaningful*

Roelof van Zwol, Leen Breure, Tim van Loosbroek

roelof@cs.uu.nl

Utrecht University

The Netherlands

Roelof van Zwol - 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]

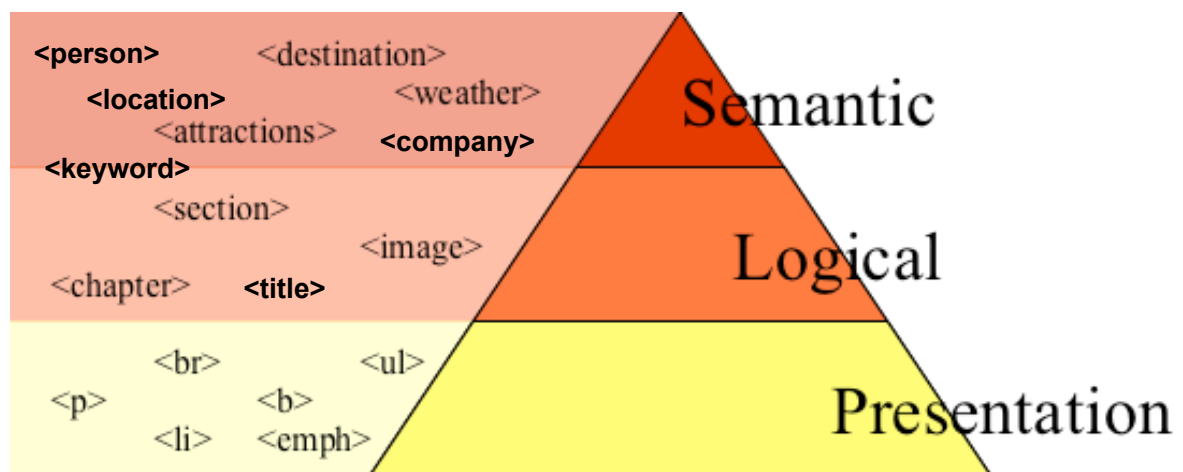
Roelof van Zwol - Utrecht University - The Netherlands

# 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

<paragraph>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. </paragraph>

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

## Reuters XML DC

❑ Based on Reuters21587 collection

	Original	Semantic
# of documents:	20841	
# of unique terms:	53080	
# of nodes:	209135	427854
avg. leaf size:	23.6	10.6
max node depth:	3	4
avg. node depth:	2.6	3.42
# of unique node names	6	11

# Reuters XML DC

Node name	Original	Semantic
article	20841	
content	19043	
paragraph	108682	
title	20840	
dateline	19041	
person	-	24508
company	-	52510
keyword	-	108682
organization	-	1096
location	-	31923

Roelof van Zwol - Utrecht University - The Netherlands

## Experimental Setup

- ❑ 2 Document Collections:
  - Reuters original vs. semantic
- ❑ 15 Topics

Roelof van Zwol - Utrecht University - The Netherlands

# 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
  1. *B<sup>3</sup>-SDR* (extension of GPX model, *INEX-2005*)
  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.

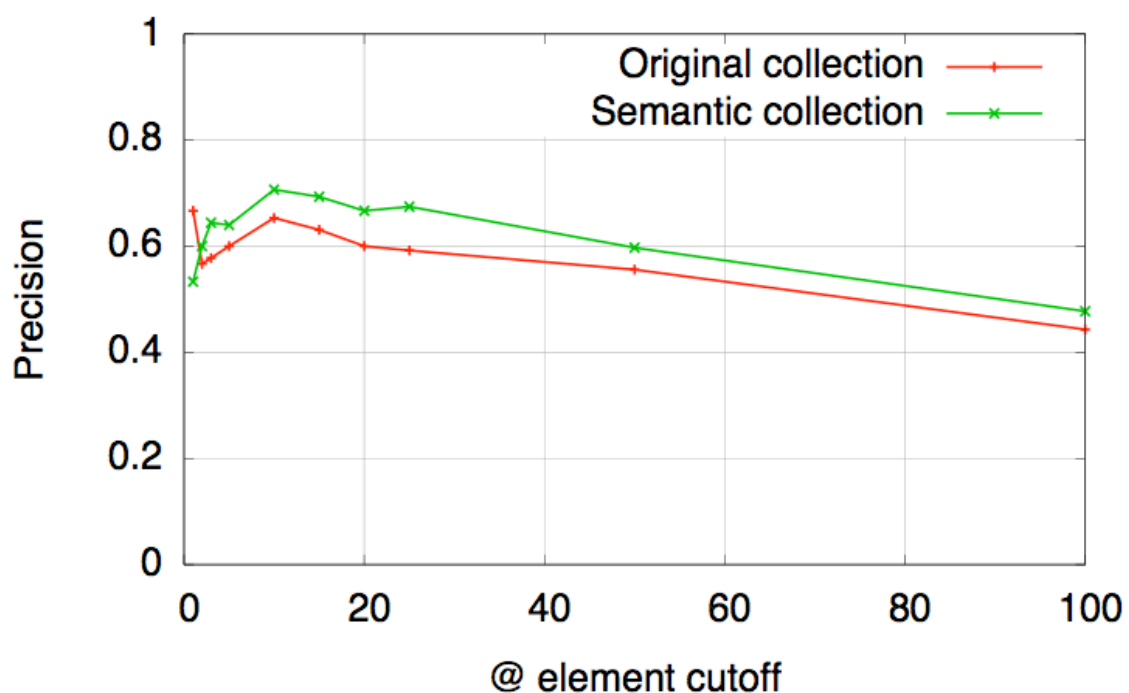
## Results - CO

	Original	Semantic	Semantic <i>No minimum node size</i>
<b>topics</b>	15	15	15
<b>retrieved</b>	1500	1500	1500
<b>relevant</b>	1424	1424	1424
<b>rel_retr</b>	665	<b>716</b>	639
<b>map</b>	0.3418	<b>0.3867</b>	0.3134
<b>bpref</b>	0.5558	<b>0.5846</b>	0.5437

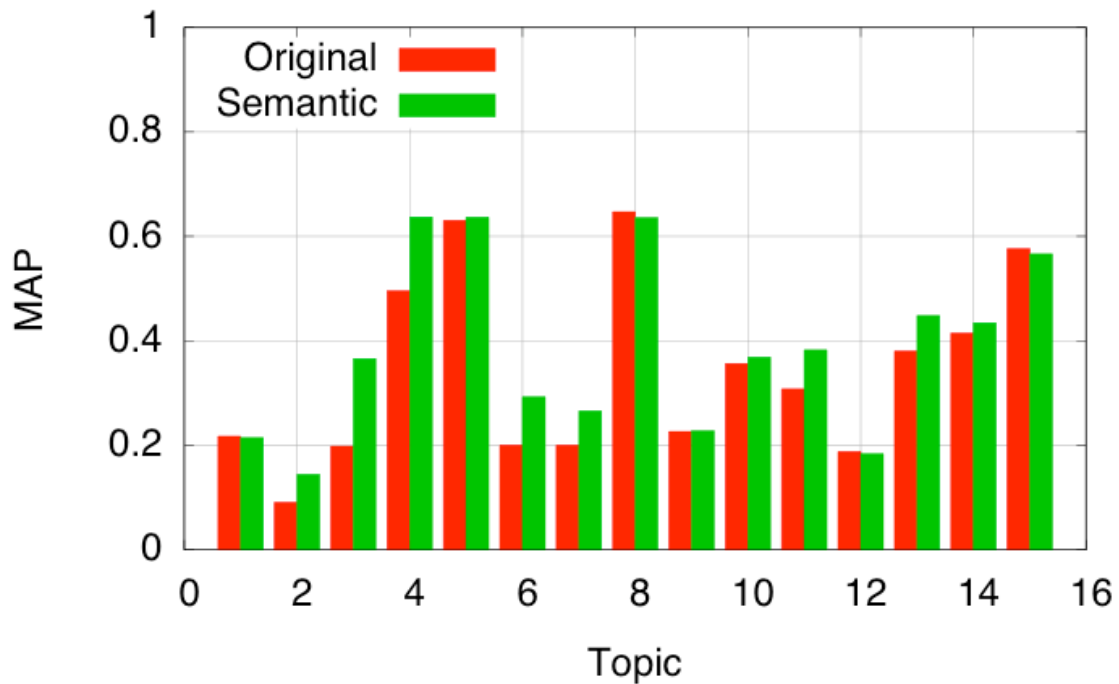
Statistical significant differences based on map (T-test):

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

## Results - CO



## Results - CO



Roelof van Zwol - Utrecht University - The Netherlands

## Results - CAS

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

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

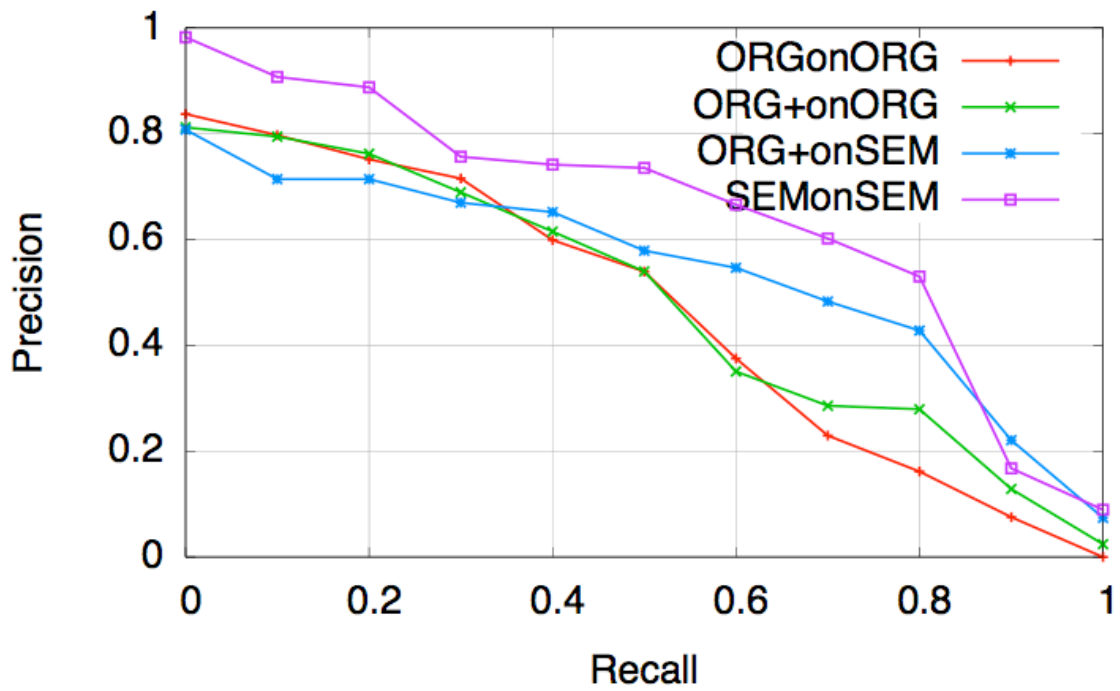
ORGonORG vs SEMonSEM:  $t(15)=-3.015, p<0.010$

ORG+onORG vs SEMonSEM:  $t(15)=-2.817, p<0.015$

ORG+onSEM vs SEMonSEM:  $t(15)=-2.475, p<0.030$

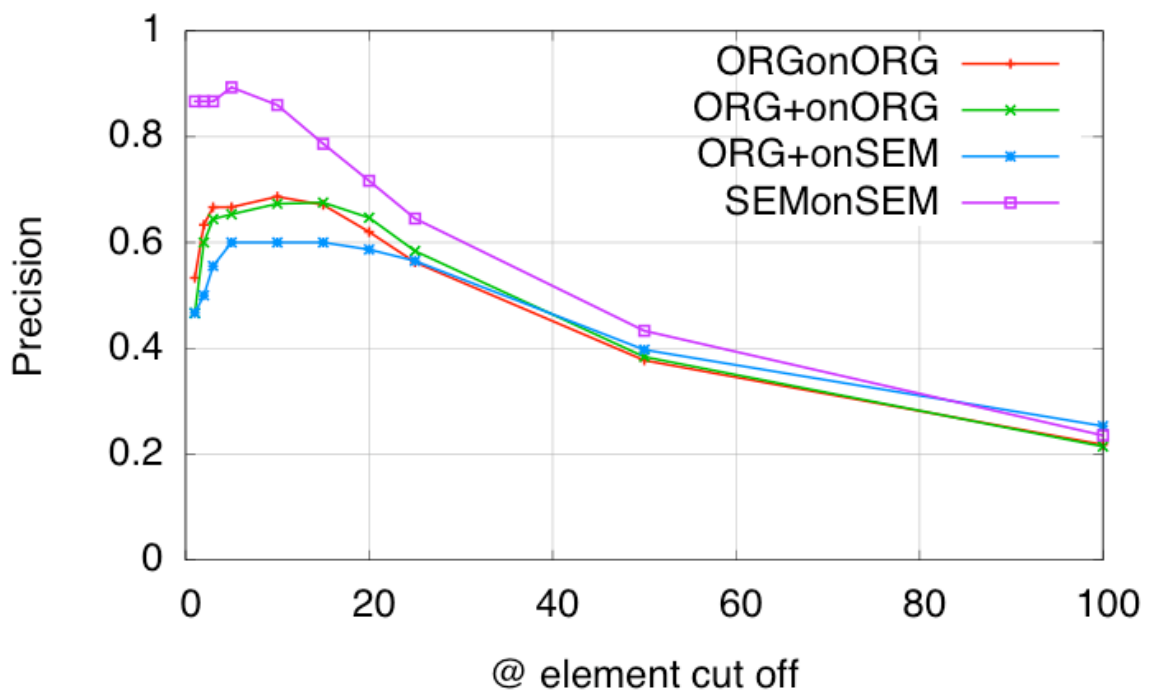
Roelof van Zwol - Utrecht University - The Netherlands

# Results - CAS



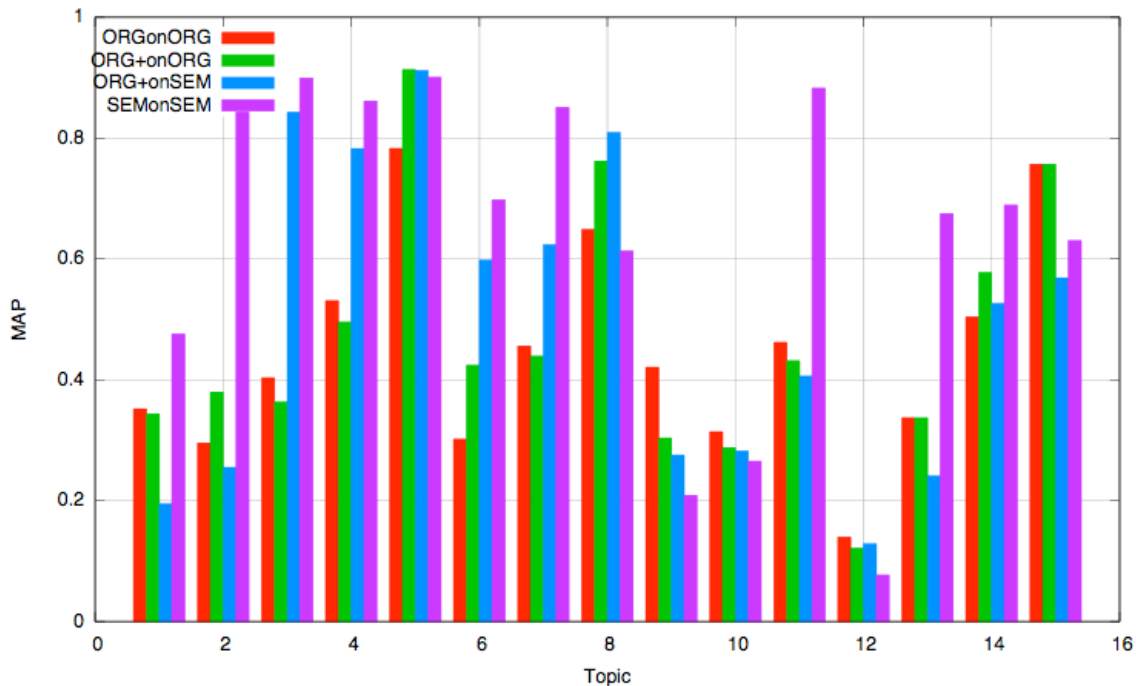
Roelof van Zwol - Utrecht University - The Netherlands

# Results - CAS



Roelof van Zwol - Utrecht University - The Netherlands

# Results - CAS



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.

Roelof van Zwol - Utrecht University - The Netherlands

## Concluding Remarks

- 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.