

XML Retrieval

- semi-structured information retrieval -

Börkur Sigurbjörnsson

Internet Information 2005/2006

Overview

- Semi-structure vs. XML
- XML Retrieval
 - Queries
 - Retrieval
 - Result presentation
- Demos
 - XML Retrieval
 - Semi-structured Retrieval
- Evaluation

2006-02-22 Internet Information 2005/2006 2

Semi-structure vs. XML

- Semi-structured documents
 - Structure is explicitly marked-up
 - Segments: sections, paragraphs, etc.
 - Layout: emphasize, italics, bold, etc.
 - “Semantic”: author, title, section title, etc.
 - Entities: city, country, person, company, etc.
- XML documents
 - Syntax for marking-up structure
 - Very useful, but not essential

2006-02-22 Internet Information 2005/2006 3

Example XML Document

```
<article>
  <author>Tom Waits</author>
  <section>Champagne for my real friends</section>
  <section>
    <title>Sham friends</title>
    Real pain for my sham friends
  </section>
</article>
```

2006-02-22 Internet Information 2005/2006 4

Example semi-structured corpora

- IEEE journals (XML)
 - Scientific articles (LaTeX) to XML conversion
 - front-matter
 - title, authors, abstract, etc.
 - body
 - sections, paragraphs, emphasis, etc.
 - back matter
 - bibliography, citations, title, author, etc.
- Wikipedia (non XML)
 - Meta-data in XML format
 - Content in Wiki format
 - page title, section titles, categories, etc.

2006-02-22 Internet Information 2005/2006 5

Example (cont'd)

```
<article>
  <front-matter>
    <title>IEEE EXPERT</title>
    <author>IEEE EXPERT</author>
    <section>IEEE EXPERT</section>
  </front-matter>
  <body>
    <h2>The networking utility</h2>
    <p>The networking utility...</p>
  </body>
  <back-matter>
    <h2>IEEE EXPERT</h2>
  </back-matter>
</article>
```

After Partition of India ==

Sadash Hasan Manto arrived in [Lahore] sometime in early [1948]. In [Lahore]] his friends had tried to stop him from migrating to [Pakistan]] because he was... order situation in post-partition India was such that many [Muslims]] felt insecure there. That was the reason that Manto had already sent his family to [Lahore]] and was keen to join them.

Incidentally his friends were right. [Lahore]] turned out to be totally different from [Bombay]]. [Lahore]] was in a state of turmoil due to the influx of hundreds and thousands of refugees in a state of destitution. Those who had survived after wading through the streets of fire and blood were clamoring for food and shelter.

== Life in Lahore ==

Manto had at least one consolation. His nephew "Hamid Jatoi" had already

2006-02-22 Internet Information 2005/2006 6

Example (cont'd)

- IEEE Computer Society Journals
 - 16,819 articles
 - 11,411,134 elements (1,581,030 indexed)
 - 170 different tag-names
- Wikipedia (English version)
 - 2,086,197 pages
 - 4,095,103 “elements”
 - (19 different “tag-names”)

2006-02-22 Internet Information 2005/2006 7

Overview

- Semi-structure vs. XML
- XML Retrieval
 - Queries
 - Retrieval
 - Result presentation
- Demos
 - XML Retrieval
 - Semi-structured Retrieval
- Evaluation

2006-02-22 Internet Information 2005/2006 8

XML Retrieval

- Queries
 - Content only
 - Normal google-like queries
 - Structured queries
 - Content and structure constraints
- Retrieval
 - Retrieve elements
 - Find relevant portions of documents
- Result presentation
 - Use structure to display relevant elements

2006-02-22

Internet Information 2005/2006

9

Queries example

- Content Only
 - aviation systems verification
- Content and Structure
 - “In articles with abstract about aviation systems give sections about verification”
 - `//article[about(//abstract, aviation systems)]//section[about(., verification)]`
- Search forms
 - Restricted version of CaS queries

2006-02-22

Internet Information 2005/2006

10

Indexing

- Content index
 - Element indices
 - Document indices
 - Structure index
 - Pre-post order
- ```

<article>
 <author>Tom Waits</author>
 <section>Champagne for my
 real friends</section>
 <section>
 <title>Sham friends</title>
 Real pain for my sham friends
 </section>
</article>

```

2006-02-22

Internet Information 2005/2006

11

## Structure index

- Calculate
  - Pre-order
  - Post-order
- Store in a RDB

| pre | post | tag     | ... |
|-----|------|---------|-----|
| 1   | 5    | article | ... |
| 2   | 1    | author  | ... |
| 3   | 2    | section | ... |
| 4   | 4    | section | ... |
| 5   | 3    | title   | ... |

```

1<article>
 2<author>Tom Waits</author>1
 3<section>Champagne for my
 real friends</section>2
 4<section>
 5<title>Sham friends</title>3
 Real pain for my sham friends
 </section>4
</article>5

```

2006-02-22

Internet Information 2005/2006

12

## Leaf element index

term (pre,freq)+

```

tom (2,1)
waits (2,1)
champagne (3,1)
real (3,1)(4,1)
friends (3,1)(4,1)(5,1)
sham (4,1)(5,1)
pain (4,1)

```

```

1<article>
 2<author>Tom Waits</author>1
 3<section>Champagne for my
 real friends</section>2
 4<section>
 5<title>Sham friends</title>3
 Real pain for my sham friends
 </section>4
</article>5

```

- Text indexed as part of its parent only
  - Compact indices
  - Good if granularity of structure is coarse
  - Complex retrieval if structure is fine grained
- Used in our Wikipedia demo

2006-02-22

Internet Information 2005/2006

13

## Overlapping element index

term (pre,freq)+

```

tom (2,1)(1,1)
waits (2,1) (1,1)
champagne (3,1) (1,1)
real (3,1)(4,1) (1,2)
friends (3,1)(4,2)(5,1) (1,3)
sham (4,2)(5,1) (1,2)
pain (4,1) (1,1)

```

```

1<article>
 2<author>Tom Waits</author>1
 3<section>Champagne for my
 real friends</section>2
 4<section>
 5<title>Sham friends</title>3
 Real pain for my sham friends
 </section>4
</article>5

```

- Text indexed as part of all ancestors
  - Non compact
  - Simple retrieval can be applied
- Used in our IEEE demo

2006-02-22

Internet Information 2005/2006

14

## Selective indices

- Reduce the overlapping indices
  - Using domain knowledge
    - Choose a set of element types to support
  - Using statistics
    - E.g. choose only element (types) that pass a certain length threshold
  - Using relevance assessments
    - Choose a set of elements that people like
- Gain efficiency without losing effectiveness

2006-02-22

Internet Information 2005/2006

15

## Language Models for XML

- Simple models
  - Standard element retrieval models
- Simple mixture models
  - Smoothing using both article and collection
- More complex mixture models
  - Smoothing at various levels of hierarchy
  - Look at work by Paul Ogilvie

2006-02-22

Internet Information 2005/2006

16

## Length Normalization

- XML elements vary in size
  - The average element is short
  - The average relevant element is longer
- Length normalization is important
  - Explicit length normalization
    - Use  $P(e)$ , e.g.  $\frac{1}{|e|}$
  - Implicit length normalization
    - Smoothing affects length
    - Relevance propagation (from small elements)

2006-02-22

Internet Information 2005/2006

17

## Result presentation

- How to present element results?
  - Ranked list of ...
    - Elements
    - Documents, decorated with element relevance

2006-02-22

Internet Information 2005/2006

18

## Overview

- Semi-structure vs. XML
- XML Retrieval
  - Queries
  - Retrieval
  - Result presentation
- Demos
  - XML Retrieval
  - Semi-structured Retrieval
- Evaluation

2006-02-22

Internet Information 2005/2006

19

## Demos

- System
  - Queries: Content only
  - Retrieval: Lucene language model
  - Result presentation: Decorated documents
- Demos
  - IEEE demo
    - <http://berk.science.uva.nl:8080/xmlfind>
  - Wikipedia demo
    - <http://wikiii.borkur.net>

2006-02-22

Internet Information 2005/2006

20

## Overview

- Semi-structure vs. XML
- XML Retrieval
  - Queries
  - Retrieval
  - Result presentation
- Demos
  - XML Retrieval
  - Semi-structured Retrieval
- Evaluation

2006-02-22

Internet Information 2005/2006

21

## Evaluation

- Ad-hoc element retrieval
  - INEX
    - Initiative for the Evaluation of XML Retrieval
  - TREC HARD
    - Passage retrieval (2003-2004)
- Interactive element retrieval
  - INEX iTrack
  - ProjectIR student project

2006-02-22

Internet Information 2005/2006

22

## INEX: Ad-hoc evaluation

- Goal
  - Evaluating XML element retrieval
- Participants
  - Create topics (content (+structure))
  - Submit runs
  - Assess relevance
- Test collection
  - Topics
  - Assessments
  - Metrics

2006-02-22

Internet Information 2005/2006

23

## Interactive evaluation

- Put people in front of a retrieval system
  - Pre-test questionnaire
  - Simulated work task
    - Find information to satisfy information need
    - Answer a number of (factual) questions
    - ...
  - Post-test questionnaire
- Data analysis
  - Questionnaire data
  - Logged user-system interaction

2006-02-22

Internet Information 2005/2006

24

## Overview (summary)

- Semi-structure vs. XML
- XML Retrieval
  - Queries
  - Retrieval
  - Result presentation
- Demos
  - XML Retrieval
  - Semi-structured Retrieval
- Evaluation