Flexible XML Retrieval using Summaries

Mariano P. Consens
University of Toronto

Outline

- Motivation
- What are Summaries
- Using Summaries in XML Retrieval
- Assessments and Performance
- Conclusions
Scenario: Proteomics Portal

Map the proteins seen in an experiment to the scientific literature.

Cross source queries that integrate text, relational and semi-structured data.

Portal developers must hand-code customized integrated search.

We want to offer search management tools that can take advantage of structure.
Outline

- Motivation
- What are Summaries
- Using Summaries in XML Retrieval
- Assessments and Performance
- Conclusions

Summaries for XML

- Structural summaries are data structures that group together elements that cannot be distinguished w.r.t. some tree pattern (XPath query)
- Summaries locate specific fragments of XML (nodes, paths and subtrees)
- By accessing relevant data directly they help to avoid sequential scans of entire documents during structured query evaluation
- But they can be used to describe XML instances by keeping a synopsis of their structural properties
XSummary Framework

- A framework for describing XML structural summaries using XPath
  - Captures many proposals in the literature: region inclusion graphs (RIGs) [CM94], representative objects (ROs) [NUWC97], dataguides [GW97], reversed dataguides [LS00], 1-index, 2-index and T-index [MS99], ToXin [RM01], A(k)-index [KSBG02], F&B-Index and F+B-Index [KBNK02], HOPI [STW04], etc.
  - Provides a uniform approach for defining new classes of summaries
- Summaries are graphs whose nodes are identified by a summary identifier (SID)
- SIDs are like tags, but they are much more descriptive of the structure around them

Incoming-Outgoing Summaries
Incoming-Outgoing Summaries

Outgoing Summaries
Summaries vs. Schemas

- DTDs and XML Schemas are used for validating instances.
- Both are schemas in the database sense, and thus describe classes of documents and constrain their structure and contents.
- However, they provide only a limited description of the instances that satisfy them and have no mechanism to locate specific instance fragments.
- In contrast, summaries are constructed for a particular instance and consequently provide a tighter description of the data.
- Summaries can be used in broader classes of applications, even when DTDs and XML Schemas are not present or very lax (semi-structured).
Atom Feed (extract)

<table>
<thead>
<tr>
<th>XML Files</th>
<th>Sample1</th>
<th>Sample2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>14113</td>
<td>9282</td>
</tr>
<tr>
<td>ATOM</td>
<td>5660</td>
<td>3093</td>
</tr>
<tr>
<td>RDF</td>
<td>3322</td>
<td>1875</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>23865</strong></td>
<td><strong>14260</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SIDs</th>
<th>Sample1</th>
<th>Sample2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>3503</td>
<td>2905</td>
</tr>
<tr>
<td>ATOM</td>
<td>5724</td>
<td>4212</td>
</tr>
<tr>
<td>RDF</td>
<td>462</td>
<td>382</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>9795</strong></td>
<td><strong>7499</strong></td>
</tr>
</tbody>
</table>

Outline

- Motivation
- What are Summaries
- Using Summaries in XML Retrieval
- Assessments and Performance
- Conclusions
The TReX Search Engine

- Java-based system developed within Eclipse
  - Builds on top of open source components for storage (BerkeleyDB JE), tokenization (Lucene)
  - Extends standard XML parsing API (Stax) to incorporate tokenization and languages (TStax)
  - Builds upon the use of summaries in ToXop for structured XML query optimization and evaluation [Barta, Consens, Mendelzon VLDB’05]

- Participants:
  - S. Ali, X. Gu, Y. Kanza, F. Rizzolo, R. Stasiu

---

NEXI Queries from INEX 2005

<table>
<thead>
<tr>
<th>Topic id</th>
<th>NEXI Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>202</td>
<td>//article[about(., ontologies)]//sec[about(., ontologies case study)]</td>
</tr>
<tr>
<td>203</td>
<td>//sec[about(., code signing verification)]</td>
</tr>
<tr>
<td>219</td>
<td>//sec[about(., learning object granularity)]</td>
</tr>
<tr>
<td>222</td>
<td>//sec[about(., business strategies)]//sec[about(., electronic commerce e-commerce)]</td>
</tr>
<tr>
<td>223</td>
<td>//article[about(., wireless ATM multimedia)]</td>
</tr>
<tr>
<td>233</td>
<td>//article[about(., music, music industry)]and about (.,/bdy, music)</td>
</tr>
<tr>
<td>236</td>
<td>//article[about(., machine translation approaches -programming)]</td>
</tr>
<tr>
<td>260</td>
<td>//bdy//*[about(., model checking state space explosion)]</td>
</tr>
<tr>
<td>270</td>
<td>//article//sec[about(., introduction information retrieval)]</td>
</tr>
<tr>
<td>284</td>
<td>//article[about(.,/bdy, thread implementation)and about (.,/bdy, operating system)]</td>
</tr>
</tbody>
</table>

- Target queries are expressions combining structure and content conditions
  - //article [about(.,/sec, code signing)]
INEX Incoming Summary (extract)

INEX Summaries (fragment/stats)

<table>
<thead>
<tr>
<th>Summary</th>
<th>Incoming</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SIDs with synonyms</td>
<td>7960</td>
<td>148</td>
</tr>
<tr>
<td>Number of SIDs without synonyms</td>
<td>11563</td>
<td>185</td>
</tr>
</tbody>
</table>
### NEXI to SID Translation

<table>
<thead>
<tr>
<th>Topic id</th>
<th>SID</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>202</td>
<td>7, 46, 82, 89, 493, 697, 619, 630,761, 1995, 2239</td>
<td>ontologies, case, study</td>
</tr>
<tr>
<td>203</td>
<td>7, 46, 82, 89, 493, 697, 619, 630,761, 1995, 2239</td>
<td>code, signing, verification</td>
</tr>
<tr>
<td>219</td>
<td>7, 46, 82, 89, 493, 697, 619, 630,761, 1995, 2239</td>
<td>learning, object, granularity</td>
</tr>
<tr>
<td>222</td>
<td>7, 46, 82, 89, 493, 697, 619, 630,761, 1995, 2239</td>
<td>business, strategies, electronic, commerce, e-commerce</td>
</tr>
<tr>
<td>223</td>
<td>7, 46, 82, 89, 493, 697, 619, 630,761, 1995, 2239</td>
<td>wireless, ATM, multimedia</td>
</tr>
<tr>
<td>233</td>
<td>7, 33</td>
<td>synthesizers, music</td>
</tr>
<tr>
<td>236</td>
<td>7</td>
<td>machine, translation, approaches</td>
</tr>
<tr>
<td>260</td>
<td>7, 33</td>
<td>model, checking, state, space, explosion</td>
</tr>
<tr>
<td>270</td>
<td>7, 46, 82, 89, 493, 697, 619, 630,761, 1995, 2239</td>
<td>introduction, information, retrieval</td>
</tr>
<tr>
<td>284</td>
<td>7, 33</td>
<td>thread, implementation, operating, system</td>
</tr>
</tbody>
</table>

- A key aspect of TReX is to translate NEXI expressions into (SID, keyword) pairs.
- Different summaries require different translations, hence different efficiency and effectiveness.

### Evaluation in TReX

- TReX uses two methods for computing NEXI queries.
- Exhaustive Algorithm (EA): queries are computed in a one-pass merge of two types of indexes:
  - Inverted file (can locate keyword positions in elements)
  - Summary-based structural indexes (can answer XPath queries).
- Threshold Algorithm (TA): the top-k answers to a query are computed from relevance-ordered posting lists (which are pre-computed using EA).
Ranking in TReX

- EA can use different scoring functions
- The results presented here use the same variant of BM25 as TopX [Theobald, Schenkel, Weikum 2005], but with Tag replaced by SID (changes according to the summary used, coincides for Label/Tag)

Outline

- Motivation
- What are Summaries
- Using Summaries in XML Retrieval
- Assessments and Performance
- Conclusions
A Glimpse of Experimental Results

- Using the INEX 2005 collection and assessments
- Compare efficiency
- Compare effectiveness

Keep in mind that changing summaries can influence both!
Comparing EA/TA Efficiency (II)

Comparing Summary Effectiveness
Comparing Summary Effectiveness

Outline

- Motivation
- What are Summaries
- Using Summaries in XML Retrieval
- Assessments and Performance
- Conclusions
Conclusions

- We need a better understanding of how structure impacts XML Retrieval
- Our approach is to build search management tools (TReX) with built-in flexibility to exploit retrieval options
- Key: achieve flexibility while retaining competitive performance

- Flexible use of summaries are crucial to achieving TReX objectives

Thanks

Mariano P. Consens
University of Toronto