Colorful XML: One Hierarchy Isn’t Enough

Divesh Srivastava
AT&T Labs–Research
http://www.research.att.com/~divesh/

Joint work with H.V. Jagadish, Laks V.S. Lakshmanan, Monica Scannapieco and Nuwee Wiwatwattana
Thesis

• XML designed for document markup
  ◦ self-describing, meaningful tags
  ◦ elements with attributes, ordered nested sub-elements
• XML now also used for data modeling
  ◦ data model underlies XPath, XQuery, XSchema
• But one hierarchy of XML data model is not “expressive” enough
  ◦ vis a vis XPath-based languages
• Solution: MCT as an elegant, conservative (yet radical) extension
Outline of Talk

• Thesis
• Problem: deep vs shallow XML trees
• Solution: multi-colored trees
• Implementation, experimental results
• Philosophy
Problem: Deep XML Trees

- **Good:** effective use of XPath, with relationships in tree structure

  ```xml
  for $m in document("mdb.xml")//movie-genre
      [name = "Comedy"]//movie[../actor/name = "Bette Davis"]
  where contains($m//movie-award/name, "Oscar")
  return <m-name>
      { $m/name }
  </m-name>
  ```

- **Bad:** redundancy, potential update anomalies
Problem: Shallow XML Trees

• Bad: reliance on value-based joins, not XPath axes

```
for $mg in document("mdb.xml")//movie-genre[name = "Comedy"]//movie-genre,
    $m in document("mdb.xml")//movie,
    $ma in document("mdb.xml")//movie-award,
    $a in document("mdb.xml")//actor[name = "Bette Davis"],
    $r in document("mdb.xml")//movie-role
where contains($ma/name, "Oscar") and $mg/@id = $m/@movieGenreIdRef and
    contains($m/@movieAwardIdRefs, $ma/@id) and
    contains($m/@roleIdRefs, $r/@id) and contains($a/@roleIdRefs, $r/@id)
return <m-name> { $m/name } </m-name>
```

• Good: normal form (XNF), no update anomalies
Solution: Multiple Colored Trees

- **Model:** Finite set of colors, each an (ordered) tree structure
  - Each element has a subset of colors
- **Query:** Effective use of XPath, relationships in multiple trees
- **Update:** Avoids redundancies, no update anomalies
  - No dependency between different colors
Names of comedy movies nominated for Oscar, with Bette Davis

for $m$ in document("mdb.xml")/\{green\}descendant::movie-award
  [contains(\{green\}child::name, "Oscar")]/\{green\}descendant::movie,
$r$ in document("mdb.xml")/\{red\}descendant::movie-genre[\{red\}child::name = "Comedy"]/
  \{red\}descendant::movie[. = $m]/\{red\}child::movie-role,
$r$ in document("mdb.xml")/\{blue\}descendant::actor
  [\{blue\}child::name = "Bette Davis"]/\{blue\}child::movie-role
return createColor(black, <m-name> { $m/\{red\}child::name } </m-name>)

Illustrates colored XPath matching
Return list of Oscar nominated movies, grouped by votes received

```xml
createColor(black, <byvotes> {
  for $v in distinct-values(document("mdb.xml")/green::descendant::votes)
  order by $v return
    <award-byvotes> {
      for $m in document("mdb.xml")/green::movie[child::votes = $v]
      return $m } <votes> { $v } </votes>
    </award-byvotes>
} </byvotes>)
```

Illustrates colored XQuery restructuring

Divesh Srivastava
MCT: What It is Not

- **Colors are not views**
  
<table>
<thead>
<tr>
<th>Views</th>
<th>dependency between data and view</th>
<th>update ambiguities</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCT</td>
<td>colors are independent</td>
<td>no update ambiguities</td>
</tr>
</tbody>
</table>

- **Multiple colored trees are not arbitrary graphs**
  
  - trees are simpler
  
  - key difference with OO and semi-structured models
MCT: Implementation

- Enhance physical data models for storing XML
- Implementation in Timber
  - XML element = content, attributes + structure
  - MCT element = XML element + structure per color
Experimental Evaluation: Goals

- Quantify efficiency in query/update processing
- Demonstrate ease of query specification
Experimental Evaluation: Setup

- **Data sets**
  - TPC-W: from XBench + 5 colors
  - SIGMOD Record: scaled-up (600KB $\rightarrow$ 60MB) + 2 colors

- **Queries**
  - TPC-W: 18 queries, 7 updates
  - SIGMOD Record: 6 queries, 4 updates
  - not in paper, available from web site
Experimental Evaluation: Performance

- MCT performance better than or comparable with shallow tree
  - value join > color crossing > structural join
- Deep tree performance has large variance
  - deep tree suffers with duplicates (retrieval, elimination)

<table>
<thead>
<tr>
<th>Query</th>
<th>Results</th>
<th>MCT</th>
<th>Shallow</th>
<th>Deep</th>
</tr>
</thead>
<tbody>
<tr>
<td>TQ3</td>
<td>4</td>
<td>0.82</td>
<td>0.83</td>
<td>0.16</td>
</tr>
<tr>
<td>TQ7</td>
<td>58</td>
<td>0.02</td>
<td>0.01</td>
<td>112.25</td>
</tr>
<tr>
<td>TQ7D</td>
<td>44929</td>
<td>2.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TQ9</td>
<td>5110</td>
<td>0.55</td>
<td>30.16</td>
<td>0.76</td>
</tr>
<tr>
<td>TQ10</td>
<td>90</td>
<td>6.61</td>
<td>8.96</td>
<td>0.71</td>
</tr>
<tr>
<td>TQ11</td>
<td>63</td>
<td>0.23</td>
<td>9.68</td>
<td>0.25</td>
</tr>
<tr>
<td>TQ13</td>
<td>2893</td>
<td>0.11</td>
<td>2.36</td>
<td>0.23</td>
</tr>
<tr>
<td>TQ14</td>
<td>253</td>
<td>0.09</td>
<td>2.29</td>
<td>0.25</td>
</tr>
<tr>
<td>TQ15</td>
<td>97</td>
<td>0.72</td>
<td>38.11</td>
<td>1.34</td>
</tr>
<tr>
<td>TQ16</td>
<td>92</td>
<td>0.40</td>
<td>20.09</td>
<td>34.61</td>
</tr>
<tr>
<td>TU1</td>
<td>1</td>
<td>0.01</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>TU1D</td>
<td>335</td>
<td>3.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TU3</td>
<td>22</td>
<td>0.36</td>
<td>15.14</td>
<td>0.65</td>
</tr>
<tr>
<td>TU4</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>TU4D</td>
<td>1994</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQ3</td>
<td>20</td>
<td>10.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQ4</td>
<td>6</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SQ4D</td>
<td>1994</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SU2</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SU2D</td>
<td>7</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental Evaluation: Storage

- Deep tree has many more elements, requires more storage
- MCT has same number of elements as shallow, but more storage

<table>
<thead>
<tr>
<th></th>
<th>MCT</th>
<th>Shallow</th>
<th>Deep</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TPC-W</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Num. Elements</td>
<td>1,502,357</td>
<td>1,502,357</td>
<td>3,883,320</td>
</tr>
<tr>
<td>Num. Attrs</td>
<td>153,713</td>
<td>153,713</td>
<td>339,674</td>
</tr>
<tr>
<td>Content Nodes</td>
<td>1,295,818</td>
<td>1,295,818</td>
<td>3,307,589</td>
</tr>
<tr>
<td>Data MBytes</td>
<td>786.27</td>
<td>329.02</td>
<td>893.09</td>
</tr>
<tr>
<td>Index MBytes</td>
<td>520</td>
<td>215</td>
<td>538</td>
</tr>
<tr>
<td><strong>SIGMOD</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Num. Elements</td>
<td>112,408</td>
<td>112,407</td>
<td>125,403</td>
</tr>
<tr>
<td>Num. Attrs</td>
<td>110,086</td>
<td>110,086</td>
<td>111,961</td>
</tr>
<tr>
<td>Content Nodes</td>
<td>108,823</td>
<td>108,823</td>
<td>118,202</td>
</tr>
<tr>
<td>Data MBytes</td>
<td>103.81</td>
<td>88.05</td>
<td>152.95</td>
</tr>
<tr>
<td>Index MBytes</td>
<td>29.7</td>
<td>18.7</td>
<td>20.5</td>
</tr>
</tbody>
</table>
Experimental Evaluation: Query Simplicity

- Quantify simplicity
  - number of path expressions, variable bindings
- MCT comparable with deep, much better than shallow
Philosophy: What are Colors?

- Database = XML document/forest
- Tuple = XML twig
- Relation = MCT color
  - same elements/tuples participate in multiple colors/relations
Philosophy: What Role can XML Play?

- **Physical model:** Timber
  - actually stored, modified

- **Logical model:** MCT
  - query and update abstractions

- **Exchange model:** XML
  - good for serialization
  - cost-based optimal serialization result in paper
Conclusions

- **Technical contributions of paper**
  - MCT model as extension of XML model
  - MCXQuery as extension of XQuery
  - size optimal serialization of MCT as XML
  - implementation in Timber, experiments

- **MCT logical data model**
  - ease of schema design
  - ease of query specification
  - efficiency in query/update processing

- **Ongoing work: schema design, MCT warehouses**