Statistics on The Real XML Data

Kamil Toman
kamil.toman@mff.cuni.cz

Department of Software Engineering
Faculty of Mathematics and Physics
Charles University
Introduction

- XML and related technologies - a leading role among standards for data representation
- Semistructured, selfdescriptive
- Possibility to express the allowed structures
  - DTD, XML Schema, Relax NG, ...
- Different techniques are needed for
  - managing
  - processing
  - querying
  - updating
  - compressing
  - versioning
  - ...

General Processing Techniques

- “As general as possible”
  - correct at first glance
  - unnecessarily complex
  - often inefficient

- With restricted features
  - more down-to-earth
  - more effective
  - restrictions are often “unnatural” (based on particular technique)
  - effectiveness suffers when data do not correspond to expectations
DTD Analysis

- DTDs still dominate among XML schemas
- Most shortcomings have been overcome in XML Schema
  - missing operator for unordered sequences
  - inheritance and modularity
  - types
  - ID <-> IDREF
- Only the simplest features are used
- Very often incorrect (both syntactically and semantically)
DTD Content Models

- Depth less than 6
- ID/IDREF used infrequently
- Unreachable elements are either root elements or useless
  - root element is stated clearly
- General recursivity is used in 58% of all DTDs
- Short simple paths (< 8)
- Cycles are common both
  - small (<100)
  - large (>500)
- Short chain of stars (mode 3)
- Significant number of hubs (elements with large fan-in)
DTD vs. XML Schema

- What extra features of XML Schema not found in DTDs are used in practice?
  - namespaces (22%)
  - extension (27%) and restriction (73%) of simple types
  - extension (37%) and restriction (7%) of complex types
  - final (7%), abstract (12%) and block (2%) attribute of complex type definitions
  - unique (7%), key/keyref (4%) features
  - unordered sequences (4%)
  - redefinitions of types and groups (~0%)
- 85% of XSDs define local tree languages (languages that can be defined by DTDs as well)
- XSD non-determinism
  - not allowed but frequent
Web XML Document Analysis

Distribution of XML documents by zone.
Web XML Document Analysis

Distribution of XML sites by zone.
Distribution of documents by their out-degree. The distribution follows a power law of exponent 1.8.
Web XML Document Analysis

- Web XML document characteristics
  - document size varies from 10B to 4.6kB
  - for documents up to 4kB the number of element nodes is about 50%, the number of attributes about 30%
    - for larger documents the number of elements decreases (~38%) while the number of attributes increases (~50%)
    - 18% of elements have no attributes
  - mixed content found in 72% of documents (5% of contents)
  - 99% of documents shallow (depth < 8)
    - average depth 4
  - only 260 total different recursive elements found
  - in 98% of recursive documents there is only one recursive element
  - 95% of recursive documents do not refer DTD or XSD
Real XML Documents

**Classification**

- *data-centric documents (dat)*
  - database exports, IMDb, list of employees, ...
- *document-centric documents (doc)*
  - Shakespeare's plays, XHTML documents, novels, docbook, ...
- *data exchange documents (ex)*
  - medical information, exchange formats, ...
- *reports (rep)*
  - overviews or summaries
- *research documents (res)*
  - docs with special structures, DNA/RNA, NASA findings, ...
- *semantic web documents (sem)*
  - RDF, OWL, DAML, ...
### Real XML Documents

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td></td>
</tr>
<tr>
<td>Number of XML documents</td>
<td>16,534</td>
</tr>
<tr>
<td>Number of XML collections</td>
<td>133</td>
</tr>
<tr>
<td>Size</td>
<td></td>
</tr>
<tr>
<td>Total size of documents (MB)</td>
<td>20,756</td>
</tr>
<tr>
<td>Minimum size of a document (B)</td>
<td>61</td>
</tr>
<tr>
<td>Maximum size of a document (MB)</td>
<td>1,971</td>
</tr>
<tr>
<td>Average size of a document (MB)</td>
<td>1.3</td>
</tr>
<tr>
<td>Median size of a document (kB)</td>
<td>10</td>
</tr>
<tr>
<td>Schema</td>
<td></td>
</tr>
<tr>
<td>Documents with DTD (%)</td>
<td>74.6</td>
</tr>
<tr>
<td>Documents with XSD (%)</td>
<td>38.2</td>
</tr>
<tr>
<td>Documents without DTD/XSD (%)</td>
<td>7.4</td>
</tr>
</tbody>
</table>

General statistics for XML data
Real XML Documents

Number of Files in Collections

- Document, 40.47%
- Database, 20.64%
- Semantic Web, 4.71%
- Research, 14.82%
- Exchange, 1.32%
- Report, 18.04%
Real XML Documents

Total Sizes of Collections

- Report, 54.78%
- Database, 10.78%
- Semantic Web, 18.75%
- Research, 8.17%
- Document, 5.72%
- Exchange, 1.79%
### Real XML Documents

<table>
<thead>
<tr>
<th>Statistics</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. number of elements</td>
<td>402</td>
<td>4,085</td>
<td>37,502</td>
<td>309,379</td>
<td>427</td>
<td>112,942</td>
</tr>
<tr>
<td>Max. number of attributes</td>
<td>9</td>
<td>1,675</td>
<td>5,182</td>
<td>37,815</td>
<td>129</td>
<td>37,996</td>
</tr>
<tr>
<td>Max. number of empty elements</td>
<td>3</td>
<td>361</td>
<td>123</td>
<td>16,348</td>
<td>6</td>
<td>23,635</td>
</tr>
<tr>
<td>Max. number of mixed elements</td>
<td>0</td>
<td>302</td>
<td>21</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Max. number of distinct el. names</td>
<td>81</td>
<td>48</td>
<td>58</td>
<td>388</td>
<td>44</td>
<td>144</td>
</tr>
<tr>
<td>Max. number of rec. elements</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max. number of distinct paths</td>
<td>79</td>
<td>96</td>
<td>67</td>
<td>312</td>
<td>30</td>
<td>143</td>
</tr>
<tr>
<td>Depth of document</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Max.</td>
<td>5</td>
<td>13</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

Global statistics for 95% XML documents
## Real XML Documents

<table>
<thead>
<tr>
<th>Statistics</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of elements</td>
<td>23,132,565</td>
<td>267,632</td>
<td>2,911,059</td>
<td>1,957,637</td>
<td>21,305,818</td>
<td>25,548,388</td>
</tr>
<tr>
<td>Num. of attributes</td>
<td>33,660,779</td>
<td>102,945</td>
<td>857,691</td>
<td>208,265</td>
<td>2,189,859</td>
<td>10,228,483</td>
</tr>
<tr>
<td>Distinct elem. names</td>
<td>81</td>
<td>134</td>
<td>146</td>
<td>461</td>
<td>210</td>
<td>1,410</td>
</tr>
<tr>
<td>Num. of distinct paths</td>
<td>434</td>
<td>2,086</td>
<td>144</td>
<td>373</td>
<td>426</td>
<td>2,534</td>
</tr>
<tr>
<td>Depth of document</td>
<td>12</td>
<td>459</td>
<td>14</td>
<td>6</td>
<td>19</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sch.</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinct elem. names</td>
<td>76</td>
<td>377</td>
<td>523</td>
<td>3,213</td>
<td>250</td>
<td>-</td>
</tr>
<tr>
<td>Num. of distinct paths</td>
<td>115</td>
<td>11,994</td>
<td>1,665</td>
<td>3,137</td>
<td>568</td>
<td>-</td>
</tr>
<tr>
<td>Depth of schema</td>
<td>12</td>
<td>81</td>
<td>79</td>
<td>5</td>
<td>15</td>
<td>-</td>
</tr>
</tbody>
</table>

Maximum values of global statistics
### Real XML Documents

<table>
<thead>
<tr>
<th>Node type</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>31.7</td>
<td>96.2</td>
<td>92.2</td>
<td>100.0</td>
<td>99.9</td>
<td>99.9</td>
</tr>
<tr>
<td>Empty element</td>
<td>26.8</td>
<td>69.2</td>
<td>89.9</td>
<td>100.0</td>
<td>86.7</td>
<td>92.7</td>
</tr>
<tr>
<td>Mixed element</td>
<td>0.2</td>
<td>76.5</td>
<td>8.7</td>
<td>0.0</td>
<td>10.1</td>
<td>2.4</td>
</tr>
<tr>
<td>Recursive element</td>
<td>0.1</td>
<td>43.3</td>
<td>63.8</td>
<td>0.0</td>
<td>0.7</td>
<td>3.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node type</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>50.0</td>
<td>94.1</td>
<td>52.6</td>
<td>100.0</td>
<td>85.7</td>
<td>-</td>
</tr>
<tr>
<td>Empty element</td>
<td>37.5</td>
<td>94.1</td>
<td>47.4</td>
<td>25.0</td>
<td>71.4</td>
<td>-</td>
</tr>
<tr>
<td>Mixed element</td>
<td>37.5</td>
<td>100.0</td>
<td>50.0</td>
<td>0.0</td>
<td>57.1</td>
<td>-</td>
</tr>
<tr>
<td>Recursive element</td>
<td>12.5</td>
<td>88.2</td>
<td>18.4</td>
<td>0.0</td>
<td>28.6</td>
<td>-</td>
</tr>
</tbody>
</table>

Exploitation rate of global properties (%)
Real XML Documents

Element Frequencies

Attribute Frequencies
Real XML Documents
Real XML Documents

- **New constructs**
  - *trivial element* – content model `a := e | pcdata`
  - *simple element* – *consists only of trivial elements*
  - *complex elements* – otherwise

- **Recursivity**
  - *trivial* - “selfrecursive”, no branching
    - `<a><a><a>...</a></a></a>`
  - *linear* – similar to trivial but can intermix with regular elements, single recursive element
    - `<a><b><a>...</a></b><c/></a>`
  - *pure* – single recursive element, branching possible
    - `<a><b/>><a>...</a><a>...</a><c/>><a>...</a><a><d/></a>`
  - *general* – *more than one recursive element*
### Real XML Documents

<table>
<thead>
<tr>
<th>Sch.</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.06</td>
<td>2.38</td>
<td>3.67</td>
<td>-</td>
<td>0</td>
<td>0.27</td>
</tr>
<tr>
<td>L</td>
<td>0.06</td>
<td>19.92</td>
<td>32.57</td>
<td>-</td>
<td>0.65</td>
<td>2.52</td>
</tr>
<tr>
<td>P</td>
<td>0.03</td>
<td>18.76</td>
<td>22.48</td>
<td>-</td>
<td>0.04</td>
<td>1.46</td>
</tr>
<tr>
<td>G</td>
<td>0.06</td>
<td>16.20</td>
<td>7.80</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sch.</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>14.29</td>
<td>-</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>2.94</td>
<td>7.89</td>
<td>-</td>
<td>28.57</td>
<td>-</td>
</tr>
<tr>
<td>G</td>
<td>12.50</td>
<td>85.29</td>
<td>13.16</td>
<td>-</td>
<td>28.57</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Doc.</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.2</td>
<td>5.0</td>
<td>6.4</td>
<td>-</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>L</td>
<td>0.5</td>
<td>65.3</td>
<td>45.7</td>
<td>-</td>
<td>66.7</td>
<td>92.6</td>
</tr>
<tr>
<td>P</td>
<td>0.7</td>
<td>12.7</td>
<td>26.9</td>
<td>-</td>
<td>0</td>
<td>6.4</td>
</tr>
<tr>
<td>G</td>
<td>98.5</td>
<td>17.0</td>
<td>21.0</td>
<td>-</td>
<td>33.3</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Doc.</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>2.9</td>
<td>-</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>0.1</td>
<td>1.0</td>
<td>-</td>
<td>20.6</td>
<td>-</td>
</tr>
<tr>
<td>G</td>
<td>100.0</td>
<td>99.9</td>
<td>99.0</td>
<td>-</td>
<td>76.5</td>
<td>-</td>
</tr>
</tbody>
</table>

Exploitation rate of types of recursions (%)

Percentage representation of types of recursion (%)
Real XML Documents

- **Shallow Relational Patterns**
  - `<a>`
    - `<b>one</b>` <!-- trivial elements -->
    - `<b>two</b>`
    - `<b>three</b>`
  - `<a>`

- **Relational Patterns**
  - `<x>`
    - `<a>xxx</a>` <!-- trivial elements -->
    - `<b>yyy</b>` <!-- no repetition -->
    - `<c>zzz</c>`
  - `<x>`
    - `<a>111</a>` <!-- trivial elements -->
    - `<c>333</c>` <!-- missing elements allowed -->
  - `<x>`
## Real XML Documents

<table>
<thead>
<tr>
<th>Statistics</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elements involved</td>
<td>29.23%</td>
<td>6.23%</td>
<td>29.53%</td>
<td>94.29%</td>
<td>22.66%</td>
<td>41.56%</td>
</tr>
<tr>
<td>Number of occurrences</td>
<td>170,744</td>
<td>154,133</td>
<td>185,358</td>
<td>40,276</td>
<td>619,272</td>
<td>716,038</td>
</tr>
<tr>
<td>Repetition</td>
<td>Avg.</td>
<td>10.5</td>
<td>3.3</td>
<td>5.8</td>
<td>322.7</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td>Max.</td>
<td>600,572</td>
<td>1,254</td>
<td>615</td>
<td>102,601</td>
<td>15,814</td>
</tr>
<tr>
<td>Fan-out</td>
<td>Avg.</td>
<td>3.6</td>
<td>1.5</td>
<td>2.2</td>
<td>6.2</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>Max.</td>
<td>33</td>
<td>10</td>
<td>18</td>
<td>26</td>
<td>51</td>
</tr>
</tbody>
</table>

Relational pattern statistics for XML documents per category
Real XML Documents

- **Mixed elements**
  - `<text><par>
    Some semistructured text including special formatting
    `<table><tr><td></td>...</tr>...</table>
    and other complex stuff
  </par> ... </par> ...
  </text>

- **Simple mixed elements**
  - `<text>Hello <b>bold</b> world!</text>`

<table>
<thead>
<tr>
<th>Statistics</th>
<th>dat</th>
<th>doc</th>
<th>ex</th>
<th>rep</th>
<th>res</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>1.8</td>
<td>4.1</td>
<td>1.0</td>
<td>-</td>
<td>1.9</td>
<td>1.2</td>
</tr>
<tr>
<td>Max.</td>
<td>6</td>
<td>448</td>
<td>5</td>
<td>-</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Simple mixed contents (%)</td>
<td>55.9</td>
<td>79.4</td>
<td>99.6</td>
<td>-</td>
<td>1.9</td>
<td>78.4</td>
</tr>
</tbody>
</table>

Mixed-content statistics for XML documents per category
Amount of tagging dominates size of document

XML Documents are shallow
- 95% of documents has < 13 max depth,
- average is about 5

Highest amounts of elements, attributes, text nodes and mixed contents are at first levels
- rapid decrease in higher levels (depths)

Data are regular
- data-centric documents can often even described by (fairly simple) relational or shallow relational patterns
- document-centric XML data also contain significant number of patterns

Most documents use some kind of standard schema
Real XML Documents - Conclusions

- **Recursion**
  - occurs quite often (doc ~ 43%, ex ~ 64%)
  - the number of recursive elements is low, though
  - it is simple, depth, branching and ed-pair distance is always less than 10
  - the most common type of recursion is linear and pure recursion
  - schemes specify the most general type of recursion

- **Mixed contents**
  - relatively high usage in document/exchange
  - low usage in data-centric documents
  - mostly simple mixed contents
  - depth is on average less than 10
XML Repositories

- **Native**
  - use some kind of numbering schema
  - the size of indexes is the key problem
  - the length of dynamic identifiers vary
  - usually the structural identifiers are to be changed on certain updates

- **(O)RDBMS**
  - leverage existing technology
  - schema driven vs. generic methods
  - inefficiencies due to large number of joins
  - XPath/XQuery <-> SQL transformation problems

- **Other**: ODBMS, Object managers, filesystem, ...
  - unsuitable for general querying
Hybrid XML Repository

- No existing general technique effective for any input data
  - using general method only if necessary
- Identification of data patterns
  - frequent parts to be processed specifically
  - preserving updatability
  - XML Schema exploitation
- Numbering schema integration
XML Fragments

- **Features of Patterns**
  - Frequent usage in real XML documents
  - Apparent meaning/purpose
  - Existence of effective processing method
  - Apparent typical updates and their possible effective processing
  - Easy recognition

- **Fragment categorization**
  - known and static (path summary schema)
  - known and finite (path summary schema)
  - mapped to relations (bubble node)
  - mapped to XML-aware text (bubble node)
  - unknown or possibly infinite (ORDPATHs like schema)
Adaptability

- Continuous changes should not affect efficiency adversely
- Invocation
  - fragment insertion
  - document insertion
  - query processing
  - automatically maintained background process
- Open issues:
  - similarity function
  - query adaptation
  - transactions
Conclusion

- Hybrid Repository
  - effective pattern recognition possible
  - specific approach for simple fragments
  - seamless numbering schema integration
  - preserving updatability
  - avoids 2+ level object identification
  - leverages existing techniques for querying

  - needs fragment similarity function
  - index building more complex
  - dynamic identifiers of variable length

  - transaction model
  - programming complexity
Thank you

See full text version for references.