Structure and Value Synopses for XML Data Graphs

Neoklis Polyzotis (UW-Madison)
Minos Garofalakis (Bell Labs)
Path Expressions

//author[book/year>2000]/paper
Path Expressions

//author[book/year>2000]/paper
Path Expressions

//author[book/year>2000]/paper
Path Expressions

//author[book/year>2000]/paper
Path Expressions

//author[book/year>2000]/paper

- Efficient evaluation → Path Selectivity
- Need to estimate true selectivities
Contribution

- **XSketch synopses**
  - Structure + Value synopses
  - Graph structured XML Data
  - Branching PEs with value predicates
  - Low estimation error/Low storage
Outline

- Preliminaries
- **XS**ketch Synopses
- Construction
- Experimental Results
- Conclusions
XML Data Model

XML Document \(\iff\) XML Data Graph

- **Graph nodes** \(\Rightarrow\) Elements+Attributes+Values
- **Graph edges** \(\Rightarrow\) Nesting + Reference (ID/IDREF)

A: Author, PB: Publisher
B: Book, N:Name, P:Paper
T: Title, E:Editor
Path Expressions

- **XPath expressions**
  - Simple: A/P/T
  - Branching: A[B]/P/T
  - Values: A/P/T[=v11]

- Result is a set
Path Expressions

- XPath expressions
  - Simple: $A/P/T$
  - Branching: $A[B]/P/T$
  - Values: $A/P/T[=v11]$
- Result is a set
Path Expressions

- XPath expressions
  - Simple: A/P/T
  - Branching: A[B]/P/T
  - Values: A/P/T[=v11]

- Result is a set
Path Expressions

- **XPath expressions**
  - Simple: A/P/T
  - Branching: A[B]/P/T
  - Values: A/P/T[=v11]
- Result is a set
Path Expressions

- XPaths expressions
  - Simple: $A/P/T$
  - Branching: $A[B]/P/T$
  - Values: $A/P/T[=v11]$
- Result is a set
Estimation Problem

- **Size of result set for a PE**
- **Challenges:**
  - **Structural Correlations**
    - **Example:** paper/author  book/author
  - **Value to Value Correlations**
    - **Example:** author[name=v1]/paper[title=v2]
  - **Path to Value Correlations**
    - **Example:** paper/author[=v]  book/author[=v]
Estimation Problem

- Size of result set for a PE
- Challenges:
  - Structural Correlations
    - Example: `paper/author` `book/author`
  - Value to Value Correlations
    - Example: `author[name=v1]/paper[title=v2]`
  - Path to Value Correlations
Estimation Problem

- Size of result set for a PE
- Challenges:
  - Structural Correlations
    - Example: paper/author   book/author
  - Value to Value Correlations
    - Example: author[name=v1]/paper[title=v2]
  - Path to Value Correlations
Estimation Problem

- Size of result set for a PE
- Challenges:
  - Structural Correlations
    - Example: paper/author  book/author
  - Value to Value Correlations
    - Example: author[name=v1]/paper[title=v2]
  - Path to Value Correlations
Outline

- Preliminaries
- XSkech Synopses
  - Synopses Model
  - Estimation
- Construction
- Experimental Results
- Conclusions
Synopsis Model

- Graph Synopsis
- Edge Stability Information
- Value Summaries
Graph Synopsis

- Set of elements (same tag) ↔ Summary Node
- Document Edge ↔ Summary Edge
Backward Edge Stability

- $u \rightarrow_b v$: all elements in $v$ have a parent in $u$
Forward Edge Stability

- \( u \rightarrow_f v \): all elements in \( u \) have a child in \( v \)
Value Summaries

- Summarize values “under” synopsis nodes
- Implementation dependent on values
  - E.g., histograms, pruned suffix trees,...
Path to Value Correlations

- One histogram per summary node
Value to Value Correlations

- Multi-dimensional histograms
- Correlations within stable neighborhood
Value to Value Correlations

- Multi-dimensional histograms
- Correlations within stable neighborhood
Value to Value Correlations

- Multi-dimensional histograms
  - Correlations within stable neighborhood
Outline

- Preliminaries
- **XSKE**tch Synopses
  - Synopses Model
  - Estimation
- Construction
- Experimental Results
- Conclusions
Estimation

Document Graph

\[ \text{?0} \]

\[\text{A}_1 \to \text{A}_2 \to \text{PB}_3 \]

\[\text{N}_4 \to \text{B}_5 \to \text{P}_6 \to \text{P}_7 \to \text{N}_8 \to \text{B}_9 \]

\[\text{V}_4 \to \text{T}_{10} \to \text{T}_{11} \to \text{T}_{12} \to \text{V}_8 \to \text{T}_{13} \to \text{E}_{14} \]

\[\text{V}_{10} \to \text{V}_{11} \to \text{V}_{12} \to \text{V}_{13} \to \text{V}_{14} \]

Synopsis

\[ \text{?}_1 \]

\[\text{N}_2 \to \text{P}_2 \to \text{B}_2 \]

\[\text{V}_N \to \text{T}_2 \to \text{T}_2 \to \text{E}_1 \]

\[\text{b/f} \to \text{b/f} \to \text{b/f} \]

\[\text{f} \to \text{b/f} \to \text{b/f} \]

\[\text{b} \to \text{V}_T \to \text{V}_T \to \text{V}_E \]
Estimation

**Document Graph**

```
?0
→ A1
↑   ↓
|    |   △
N4   B5   P6   P7
|   ↓   ↓   ↓   ↓
V4   T10  T11  T12
↓   ↓   ↓   ↓
V10 V11 V12 V8
```

**Synopsis**

```
?1
→ A(2)
↑   ↓
b/f  b/f
N(2) P(2)
↓   ↓
V_N V_T
```

```
→ PB(1)
↑   ↓
b/f b/f
B(2)
↓
V_T V_E
```

\[ B[E=v1]/T[=v2] \rightarrow 2 \times f(B[E=v1]/T[=v2]) \]
Estimation

Document Graph

Synopsis

\[ B[E=v_1]/T[=v_2] \rightarrow 2 \times f(B[E=v_1]/T[=v_2]) \]
Estimation

\[ B[E=v_1]/T[=v_2] \rightarrow 2 \times f(B[E=v_1]/T[=v_2]) \]
\( B[ E=v1 ] / T[=v2] \rightarrow 2 \times f( B/T[=v2] ) \times f( E=v1 | B ) \times f( B[E] ) \)
Estimation Model

- Break path in stable sub-paths
- Derive correlation scopes
- Apply statistical assumptions
  - Independence
  - Uniformity
Outline

- Preliminaries
- XSketch Synopses
- Construction
- Experimental Results
- Conclusions
Coarsest Synopsis

- All document paths, but also false paths
- High estimation error
- Small size
Perfect Synopsis

- All document paths and no false paths
- Zero estimation error
- Large size
Construction

- **Optimal **X**S**k**e**t**c**h**: NP-Hard
- **Forward Selection Algorithm**
  - Refinements
  - Successively refine coarsest summary
  - Selection criterion: marginal gains
Construction Step
Construction Step

Path Sample $P$

$E = \text{error}(P)$
$S = \text{size}()$

$E' = \text{error}(P)$
$S' = \text{size}()$
Construction Step

Path Sample \( P \)

\[ E = \text{error}(P) \]
\[ S = \text{size}(r) \]

\[ E' = \text{error}(P) \]
\[ S' = \text{size}(r) \]

\[ \text{gain}(r) = \frac{(E - E')}{(S' - S)} \]
Refinements

- Structural Refinements
  - backward-stabilize
  - forward-stabilize
  - backward-split
- Value Refinements
  - value-expand
  - value-remove
  - value-refine
Outline

- Preliminaries
- XSketch Synopses
- Construction
- Experimental Results
- Conclusions
Implementation

- Single-dimensional histograms
  - Integer values
    - Strings hashed to integers
  - Construction: \text{max-diff}(V,A)
## Datasets

<table>
<thead>
<tr>
<th></th>
<th>Elements</th>
<th>Coarsest Summary (KB)</th>
<th>Perfect Summary (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>102,755</td>
<td>7.8</td>
<td>1.9</td>
</tr>
<tr>
<td>XMark</td>
<td>87,480</td>
<td>4.1</td>
<td>3.3</td>
</tr>
</tbody>
</table>
Workload

- 1000 Positive Pes
  - Biased random sample from document
  - Path Length: 2-5
  - 500 contain range predicates
    - Predicates: random, 10% of value domain
- Similar results with negative PEs
Accuracy Metric

- Average Absolute Relative Error

\[
\frac{1}{|W|} \sum_{p \in W} \frac{|\text{count}(p) - \text{estim}(p)|}{\text{count}(p)}
\]
Results - IMDB (Branching)

Branches: 0-2
Avg. Result Count: 478(Predicates)/1901(No Predicates)
Results - IMDB (Simple)

Branches: 0
Avg. Result Count: 483(Predicates)/933(No Predicates)
Conclusions

- Path selectivity estimation is important
- **XS**ketch synopses
  - Branching PEs with predicates
  - Graph-structured data
  - Model: graph synopsis+stability+value summaries
  - Efficient forward selection algorithm
- Experimental Results
  - Accurate synopses with small space requirements
  - Effective construction algorithm
Overflow Slides
Path Expressions

- XPath expressions
  - Simple: A/P/T
  - Branching: A[B]/P/T
  - Values: A[N=v8]/P/T
    - $A/P/T[v11]$ Result is a set
Estimation (a)

\[ A/P/T[=v] \rightarrow 2 \times f(T=v|A/P/T) \]
Estimation (c)

\[
A/B/T[=v] \rightarrow 2 \times f(T=v|B/T) \times f(A/B \mid B/T[=v])
\]
Estimation (c)

\[ A/B/T[=v] \rightarrow 2 \times f(T=v|B/T) \times f(A/B) \]
Value-expand example

\[ A[N=v1]/P/T[=v2] \rightarrow 2 \times f(T=v2) \times f(N=v1) \]
Value-expand example

\[ A[\text{N=v1}] / P / T[\text{=v2}] \rightarrow 2 \times f(\text{T=v2 and N=v1}) \]
Results - XMark (Branching)

Path length: 2-5    Branches: 0-2
Avg. Result Count: 254(Predicates)/1057(No Predicates)

![Graph showing Avg Abs Rel Error (%) vs Summary Size (KB)]
Results - XMark (Simple)

Path length: 2-5   Branches: 0
Avg. Result Count: 302(Predicates)/771(No Predicates)
Synopsis Model

- Graph Synopsis
- Stability Information
- Value Distribution Information
XSketch Structural Synopses

- Previous Work
  - Branching PEs
  - Graph structured XML data
  - Low estimation error/Small size

- Values?
  - Path to value correlations
  - Value to value correlations