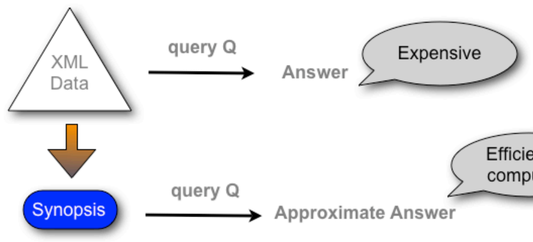
AQAX: Approximate Query Answering for XML

Josh Spiegel, M. Pontikakis, S. Budalakoti, N. Polyzotis
University of California Santa Cruz

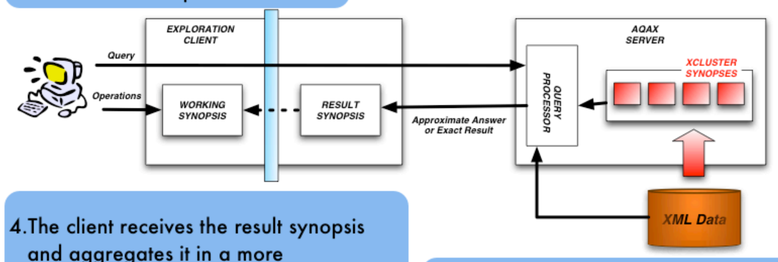
Motivation for AQAX



- Goal: provide feedback for ad-hoc queries
  - The user can see "previews" of results
  - The user may not need the exact answer
- Key for on-line exploration of large XML data sets

System Architecture

1.The user formulates a query and selects a level of precision



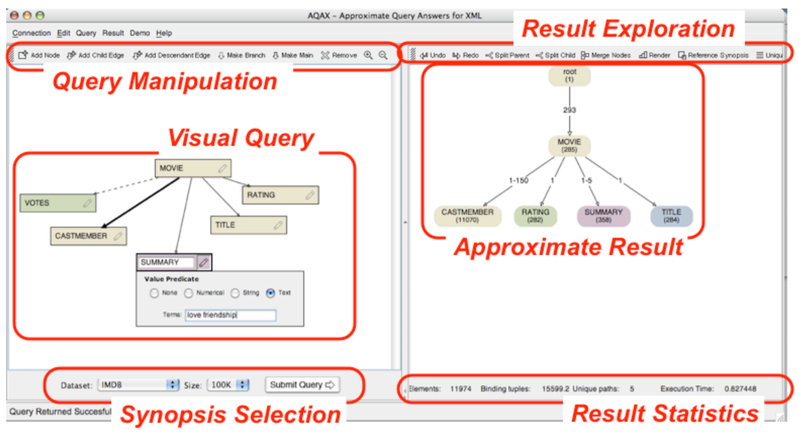
2.The server processes the query and generates an approximate answer
- Summarization model: XCluster

3.The answer is returned to the client in the form of a *result synopsis*

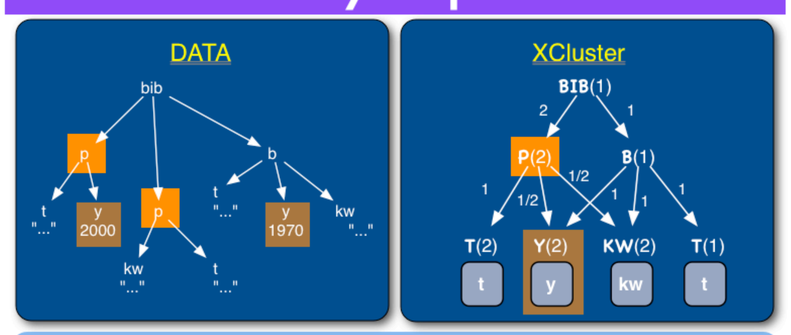
4.The client receives the result synopsis and aggregates it in a more manageable *working synopsis*
- Visual model: XCluster

5.The user explores the answers by manipulating the working synopsis

User Interface



XCluster: Synopsis Model



- Structural information: node- and edge-counts
  - Node-count: number of elements in cluster
  - Edge-count: average number of children
- Value information: value-distribution summaries

Content Heterogeneity

- Elements include values of different types
- Similarly, queries involve predicates of different types

```
<paper>                                   Numerical
  <year>2003</year>                       String
  <title>The history of histograms (abridged)</title>
  <author>Yannis Ioannidis</author>
  <abstract>
    The history of histograms is long and rich, full
    of detailed information in every step. It
    includes...                                   Text
  </abstract>
</paper>
```

```
//paper[year>2000][author contains "Ioannidis"]//
   abstract[ftcontains histograms,history]
```

Range — Substring — Term Containment

Content Summarization

- Types of value summaries:
  - Numerical content: Histograms
  - String Content: Pruned Suffix Tries
  - Text Content: End-biased Term Histograms
- XCluster provides a unified framework for structure and heterogeneous value content

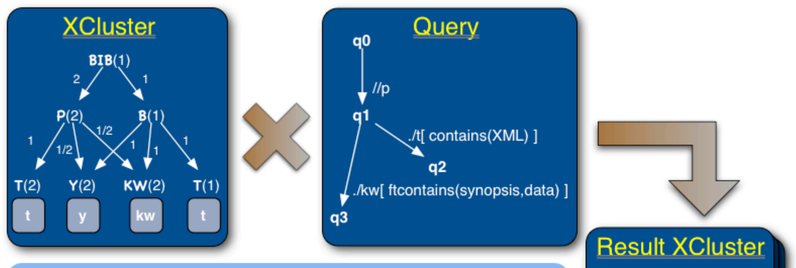
"The history of histograms is long and rich, full of detailed information in every step. It..."

Term	Freq
0 (history)	2
1 (histogram)	7
2 (data)	6
3 (database)	5
4 (information)	3
5 (value)	2

Bucket	Freq
010000	7
001000	6
000100	5
100011	7/3

Text → Term Matrix → Term Histogram

Query Evaluation



- Efficient computation in a single depth-first pass
- Result synopsis has bounded size |Q|*|S|
- Accuracy depends on "tightness" of nodes
  - A node is tight if it contains elements that have similar structure and values

Synopsis Construction

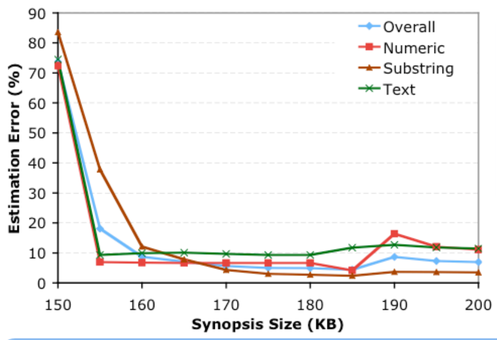
XML Data → Reference Summary (Stage 1) → XCluster with detailed value distributions (Stage 2) → XCluster (Stage 3)

- Goal: cluster elements in nodes according to their structural and value-based similarity
- 3-stage build process:
  - Step 1: Construct reference synopsis (fine grained clustering)
  - Step 2: Cluster based on structure and values
  - Step 3: Build value summaries
- Use of novel heuristics to determine a good clustering of elements

XCLUSTER: Visual Model

- XCluster is used as the visual model for result presentation
- The initial view is the coarsest XCluster summary of the answer
- The user can explore the results by manipulating the synopsis
  - Structural operations: split/merge nodes
  - Value operations: render distributions, compute aggregates

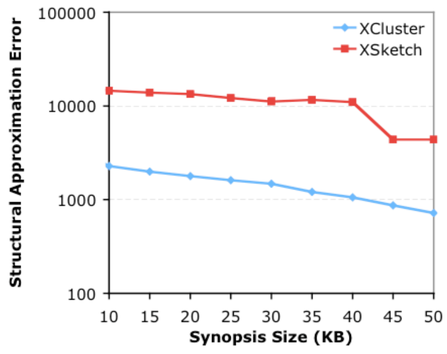
Value Approximation



- Workload: 1000 random twig queries, categorized by type of value predicate
- Estimation error is measured as avg. absolute relative error (lower is better)

- Error is close to 10% with a modest space budget
- Quick convergence to small errors
- XCluster captures well the correlation b/w structure and values

Structure Approximation



- Workload: 1000 random structure-only twig queries
- Error is measured as element simulation distance (lower is better)

- Considerable improvement compared to previous techniques
- XCluster captures well the correlations within the XML structure

References

- **Approximate XML Query Answers**
  N. Polyzotis, M. Garofalakis, Y. Ioannidis
  In Proceedings of SIGMOD, 2004
- **XCluster Synopses for Structured XML Content**
  N. Polyzotis, M. Garofalakis
  In Proceedings of ICDE, 2006