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
3. The user explores the answers by manipulating the working synopsis
4. The client receives the result synopsis and aggregates it to a more manageable working synopsis
- Visual model: XCluster
- Synoptic model: XCluster
- The answer is returned to the client in the form of a result synopsis

User Interface

- Result Exploration
- Approximate Result
- Synopsis Selection
- Result Statistics

XCluster: Synopsis Model

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

Content Heterogeneity

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

Query Evaluation

- Efficient computation in a single depth-first pass
- Result synopsis has bounded size [Q1*] [5]
- Accuracy depends on “tightness” of nodes
- A node is tight if it contains elements that have similar structure and values

Synopsis Construction

- 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

Value Approximation

- Workload: 1000 random tree queries, categorized by type of value predicate
- Evaluation error is measured as avg. structure 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 tree queries
- Error is measured as element structure 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