A General Framework for Estimating XML Query Cardinality

Carlo Sartiani
Dipartimento di Informatica - Università di Pisa

9th International Workshop on Data Bases and Programming Languages - DBPL03

Topics

- The problem: estimating the result size of XQuery expressions, starting from summarized info about the input data
- The framework: a set of notions and tools implementing them
    - a meta-model

Issues in Result Size Estimation

- Twigs
  - for $y in $x/a, $z in $x/b …
  - branch correlation
- Set cardinality (let ... :=)
  - for $y in $x/a let $z = $x/b …
- Predicates
  - for $y in $x/a, $z in $x/b where $a>200

The Framework

- Model independent
- It offers
    - correlation
    - group cardinality estimation
    - predicate selectivity application

Basics

- Estimation functions compute the distribution of data into query result
- Result distribution is expressed by means of sequences of match occurrences
- Sequence of match occurrences are bound to variables

Match Occurrence

- (l,r,m)
- l: tag of the matching nodes
- r: region of the database
- m: multiplicity of the occurrence

Regions

- Intensional regions: types
- Extensional regions: position intervals, etc
- Mixed regions: intensional + extensional

Tagged Regions

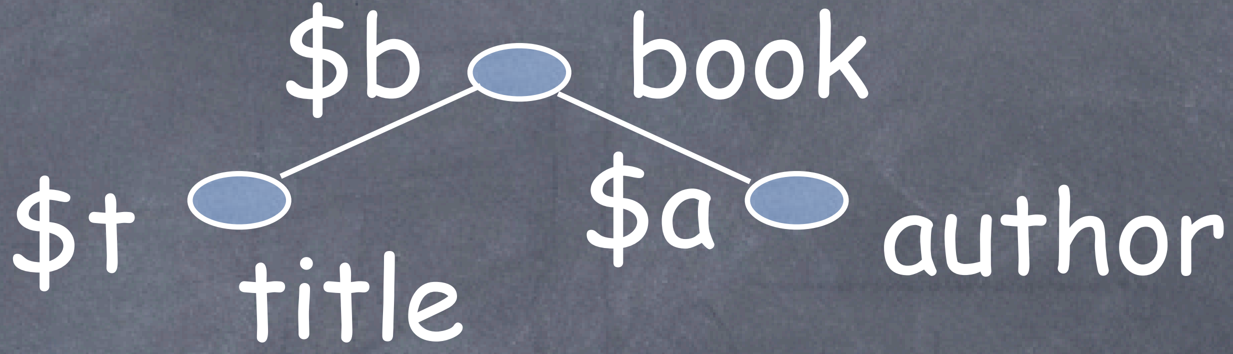
- Regions augmented with tag information
  - (l,r)
- Organized into a graph
  - /-edges, //-edges, etc

Correlation

- (l,r,m) and (l’,r’,m’) are correlated wrt to (l”,r”,m”)
- if (l”,r”) is a common ancestor for (l,r) and (l’,r’) in the tagged region graph

More on Correlation

- (title,r1,m1) correlated to (author,r2,m2) wrt (book,r3,m3) ?
- (title,r1,m1) correlated to (author,r2,m2) wrt (book,r3,m3) ?
- Constrained common ancestor problem
- O(n) time complexity (with proper data structures)



Groups

- Estimating the distribution of data into sets created by the let clause
  - Distributing match occurrences into sets
  - Correlation-based

More on Groups

- Number of groups determined by the cardinality of the root variable
- Performed in $O(n^2)$ time
- Extensible to future groupby constructs

Predicates

- Predicate selectivity depends on
  - the kind of predicates
  - the semantics of the data being filtered
- data($y) > 1994

More on Predicates

- Selectivity factor
  - psf[P]: TaggedRegion -> [0,1]
- Factors propagated to the occurrences of the same twig

Xtasy Model

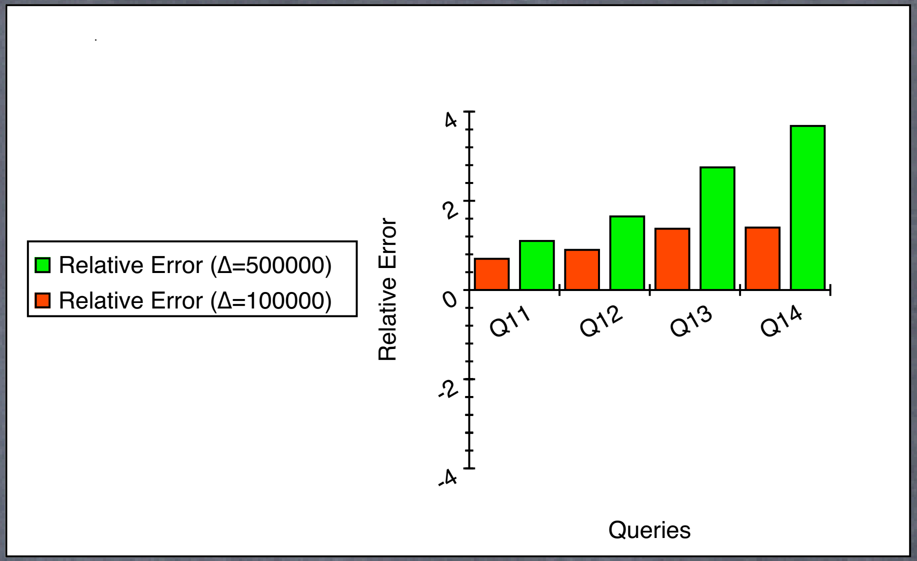
- An instance of the framework
- Extensional regions: (h, [p, p+Δ])
  - h: a level in the tree
  - [p, p+Δ]: a positional interval
- Estimation functions work on physical operators

Benchmark Queries

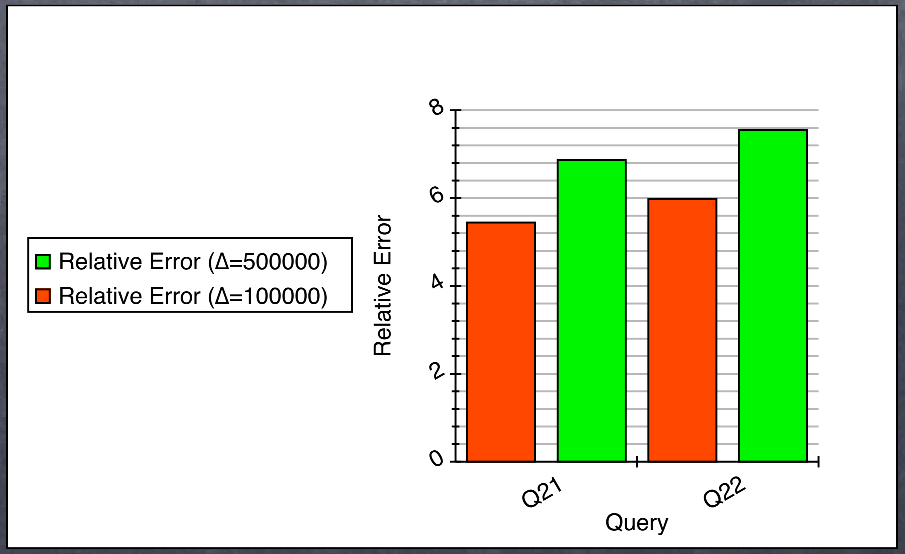
- Six classes of benchmark queries
  - path queries
  - twig queries
  - twig queries with groups
  - queries with predicates
  - nested queries
  - negative queries

Experimental Results (1/3)

- Path queries
- Twig queries



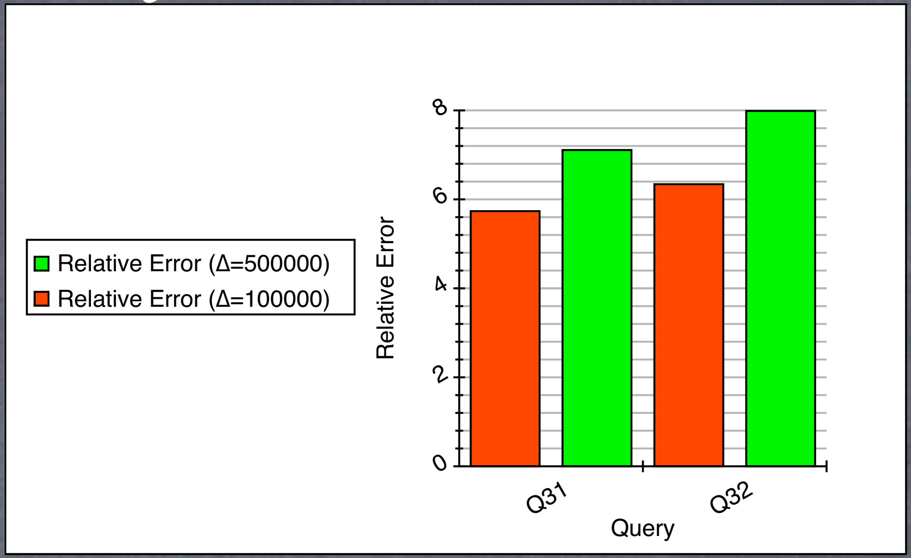
Path queries



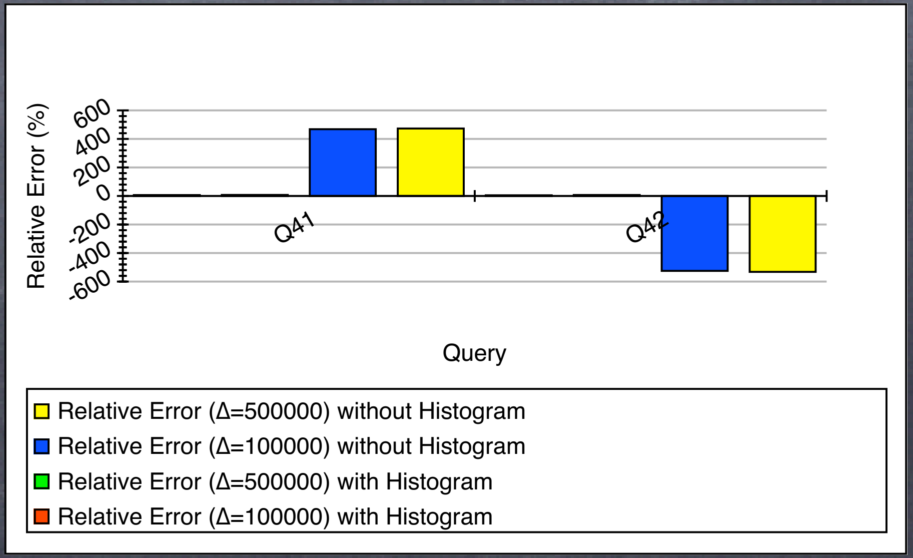
Twig queries

Experimental Results (2/3)

- Twig queries with groups
- Queries with predicates



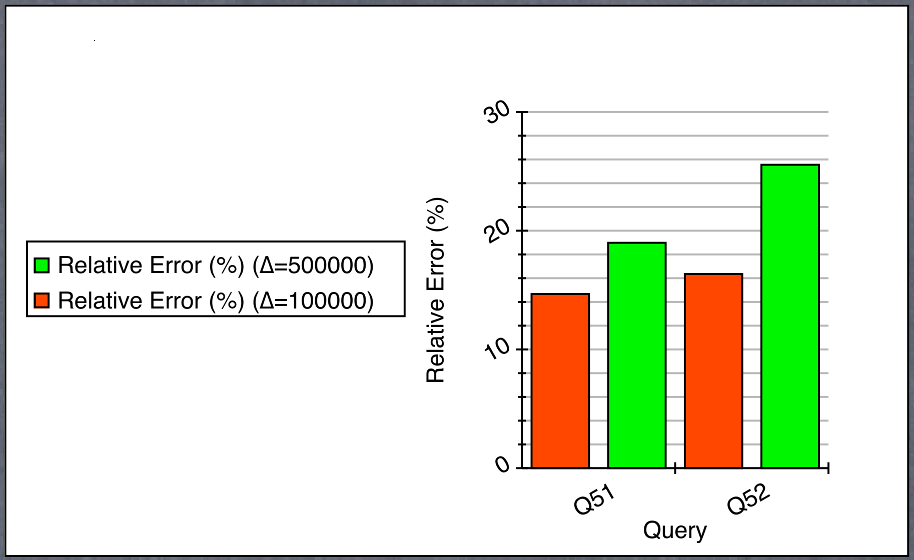
Twig queries with groups



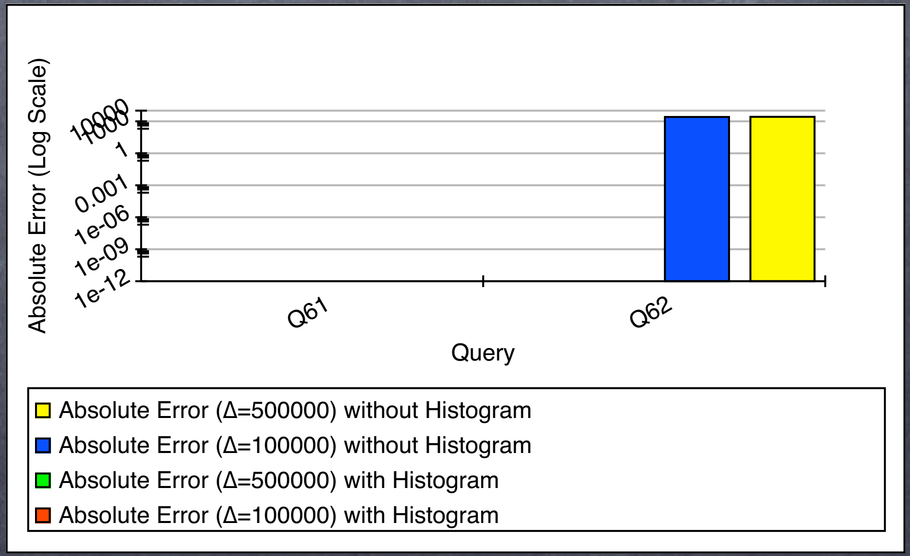
Queries with predicates

Experimental Results (3/3)

- Nested queries
- Negative queries



Nested queries



Negative queries

Conclusions

- An infrastructure for size estimation models
- Future work
  - groupby
  - more tree-oriented vision