Sherif Sakr
National ICT Australia (NICTA)

**XSelMark: A Micro-Benchmark of Selectivity Estimation of XML Queries.**
XML Benchmarks

- XML benchmarks can be classified into two main categories:
  - Application (Macro) benchmarks: which are used to evaluate the overall performance of an XML management system [XMark, XMach, XOO7, TPOX].
  - Micro-benchmarks: which are designed to assess the performance of specific features of the XML management systems [XPathMark, MemBeR].
Motivation

- Although the XML research community has proposed several benchmarks which are very useful for their intended targets, *none of these benchmarks* fits in the context of being able to evaluate the different selectivity estimation approaches of XML queries.

- Michiels et al. [ISJ 08] have promoted the idea of a crucial need for different micro-benchmarks in order to get a good understanding of the different aspects in implementing efficient query processors in the XML domain.
Motivation

- Estimating the sizes of query results and intermediate results is a crucial part of any effective query optimization process.
- Modern implementations of query processors are heavily reliant on the efficient performance of sophisticated optimizer components to achieve a proper selection of many optimization decisions such as:
  - access paths
  - join orders
  - materialization strategies.
The different alternative execution plans for the same query are represented by different shades in the figure.

Alternative execution plans may differ significantly in their use of system resources or response time.

These differences can be in orders of magnitude of performance between the best and the worst plans which makes selecting the right plan a very important task.

Clearly, accurate selectivity estimation plays a crucial role in making the right decisions.
Motivation

- Due to the lack of a suitable benchmark, it was difficult to assess, evaluate and compare these approaches and in order to get a clear view about the state-of-the-art.
- We faced with this specific problem during our work in handling the selectivity estimation problem of XML queries [KOPS 07, IJWIS 08, VLDB 08].
- XSelMark is considered as a first step to bring a complete and realistic assessment of the state-of-the-art of the available approaches in the domain of selectivity estimation of XML queries along with their strengths and weaknesses.
Merits of Cardinality Estimation

- It is a crucial piece for an effective query optimization process.
- It provides an early feedback about the expected size for the input queries before evaluating the full results.
- Estimated cardinalities can be used as an approximate answers for aggregate queries.
- It gives hints on the possible avenues to optimize the resource allocation of the execution process.
Selectivity Estimation Difficulties in the XML Domain

- the selectivity estimation problem in the XML domain is more complicated than that in the relational domain due to several reasons such as:
  - The absence of a strict schema notion in the XML data.
  - The dualism between structural and value-based querying.
  - The high expressiveness of the XML query.
  - The non-uniform distribution of tags and data.
  - The correlation and dependencies between the occurrences of the elements.
Main Aspects of Selectivity Estimation in the XML Domain

When looking for an efficient, capable and accurate selectivity estimation approach for XML queries, there are several issues that need to be addressed:

- It should support structural and data value queries.
- It must be practical.
- It should be strongly capable.
- It should be composable.
- It must be accurate.
- It should be independent.
The queries of XSelMark benchmark is based on the structure of the XMark document "auction.xml".

XMark is a well-known benchmark for XML data management. The XMark database models an internet auction web site.

The set of queries XSelMark, represents a mix of XML queries which covers a wide set of the major selectivity estimation aspects in the domain of XML queries.

They are designed in a way to facilitate a realistic assessment in terms of the advantages and shortcomings of the proposed XML selectivity estimation approaches and to identify their respective impact.

The set of queries are expressed using the two standard XML query languages XPath and XQuery.

http://xselmark.sourceforge.net/
XSelMark: Query Groups

- Path expressions
- Twig Expressions
- Predicates
- Value-Based Joins (Theta Joins)
- Arithmetic and Comparison operations over Data Value Statistics
- Nested Expressions
- Data Dependent Estimations
XSelMark Queries: Path Expressions

- Path expression with non-recursive axes
- Path expression with recursive axes
- Path expression with wild card
- Branching XPath Expressions
  - `/site//person[profile/age]/name`
- Path expression with ordered-based axes
  - `/site//closed_auction//following::description`
    - Supporting such types of queries requires the selectivity estimation approach to capture specific statistical information about the order of the elements in the XML documents.
Twig expression with element construction

- for $b in //item return

  <Result>

  <name>{$b/name}</name>

  <description>{$b/description}</description>

  </Result>
Most common solutions of this problem rely on histograms for capturing the distribution of data values, and on the use of the uniform distribution when nothing is known about the data involved in the predicate.

Predicate selectivity estimation in the XML domain poses new challenges such as:

- The predicates can be structural-based as well as value-based.
- Positional predicates represents a special form of predicates over the order information of the elements in the XML document.
- XML elements are usually distributed in a non-uniform way.

**Positional Predicates**

/site/open_auctions/open_auction/bidder[3]
Value-based join instances where the values of each operand are constructed by path expression

Value-based join instances where the values of one operand are constructed by path expression and the values of the other operand are constructed by path expression manipulated with arithmetic expression

- for $x$ in /site//increase, $y$ in /site//price where data($x$) > data($y$) * 2
  return <pair>{$x,$y}</pair>

Equi-Joins of data values

- for $x$ in /site//increase, $y$ in /site//price where data($x$) = data($y$)
  return <pair>{$x,$y}</pair>
This group of queries assesses the ability of the selectivity estimation approaches on their ability of not only being able to capture summary information about the data values of the XML elements but also on their ability of applying arithmetic and comparison operations over this summary information in a consistent and accurate way which does not hurt the quality of the selectivity estimation results.

- for $x$ in /site//increase, $y$ in /site//price, $z$ in /site//@income
  
  where data($x$) + data($z$) > data($y$) + data($z$)
  
  return <pair>{$x,$y,$z}</pair>
XQuery, as with many other XML query languages such as SQL/XML, is a free nesting language, where nested queries can be used for many targets such as reshaping elements or computing aggregate values.

Since the result of nested queries may be the input for navigational or filtering operations in the outer query, predicting the size of nested query results will require building on-the-fly statistics about these intermediate results.

```xml
for $b in /site/open_auctions/open_auction
  where sum(data($b/bidder/increase)) > 1000
return <increase>{$b}</increase>
```
This group of queries requires capturing additional specific forms of summary information about the data values of the underlying XML documents.

- **Full Text Search**
  - `/site/item[contains(description,'gold')]`

- **Distinct Operator**

- **Existential Document Order**
  - for $b$ in `/site/open_auctions/open_auction`
    where
      
      some $pr1$ in $b/bidder/personref[@person = "person20"],
      $pr2$ in $b/bidder/personref[@person = "person51"]
    satisfies $pr1 < $pr2
  return <history>{$b}</history>
There is currently a big demand for several micro-Benchmarks which assess specific aspects in the XPath, XQuery and XML management system domains.

Due to the lack of a suitable benchmark, it was difficult to assess, evaluate and compare these approaches and in order to get a clear view about the state-of-the-art.

XSelMark is considered as a first step towards a comprehensive assessment of the available selectivity estimation approaches of XML queries.
Conclusions

- An initial assessment of the features and capabilities of the current approaches [XPathLearner 02, Bloom 04, XSEED 06, OrdXPath 06, XCluster 06] has shown that most of them are limited to supporting the estimation of the structural XPath queries.

- Hence, several avenues for further research and development are still widely open in this domain to provide more accurate, capable and complete frameworks that are aligned with the rich querying capabilities of the standard XML query languages.
References