IMPLEMENTATION AND OPTIMIZATION OF XML FULL-TEXT SEARCH

EMIRAN CURTMOLA
University of California San Diego
ecurtmola@cs.ucsd.edu

SIHEM AMER-YAHIA
AT&T Research Labs
sihem@research.att.com

ALIN DEUTSCH
University of California San Diego
deutsch@cs.ucsd.edu

Motivation

• Problem:
  – IR studies systems for indexing and searching text
  – All query languages for XML are structure-oriented

• Existing query languages are not powerful enough
  – text search not expressive enough
  – not all XQuery spectrum

• GalaTex is a conformant implementation of XQuery Full-text language, a W3C extension of XQuery and XPath with full-text search primitives such as phrase matching, Boolean connectives, keyword-distance, ordering, stemming that can be combined with navigation over document structure

Queries in XQuery Full-Text

• Context expression:
  – i.e. set of book paragraphs, book chapters as opposed to searching the whole document as in traditional IR

• Return expression:
  – i.e. book titles and paragraphs as opposed to retrieving the whole document as in traditional IR

• Search expression:
  – full-text search primitives: and, or, negation, distance, ordered, window, times etc.

• Score expression:
  – scoring and ranking the results

Full-Text Query Example

• A query example expressed in natural language:
  find the top 10 book paragraphs that contain "users" and "software" at a distance at most 13 words of each other

• The same query example in XQuery Full-Text:
  for $result at $pos in $xmlfile/books/book/paragraph
    score $s as $pos contains "users" & "software" with distance at most 13 words
    order by $s
  return $pos <= 10 return $result

New optimization opportunities

• Define what are “good properties” for a score-aware algebra

• Scoring on both content and structure

• Consistent scoring
  – Equivalent query expressions should result in the same scores for any given document fragment

• Consistent ranking
  – Equivalent query expressions should result in the same topK results for any given document fragment

Optimizations

• Full integration of XQuery Full-Text algebra into XQuery algebra

• Efficient evaluation algorithms for each full-text primitive

• Prune intermediate results as soon as possible
  – Avoid computing cartesian products
  – Push selections (i.e. distance, ordered, window, scope, times primitive) down in the plan
  – Merge multiple selections with the join into a complex join operator
  – Translate the XQuery Full-Text Boolean operators into XQuery Boolean operators

Current Status & Ongoing Work

• Current Status:
  – GalaTex is the first complete implementation of W3C XQuery Full-Text language
  – A web demo including the W3C XQuery Full-Text usecases is available at: http://www.galaxquery.com/galatex

• Ongoing Work:
  – define a good algebra for XQuery Full-Text as a platform for joint optimizations on structure and text that takes scoring information into account