Fractional XSketch Synopses for XML Databases

Natasha Drukh\textsuperscript{1}, Neoklis Polyzotis\textsuperscript{2}, Minos Garofalakis\textsuperscript{3}, and Yossi Matias\textsuperscript{1}

\textsuperscript{1} School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel, kreimern@post.tau.ac.il, matias@cs.tau.ac.il
\textsuperscript{2} Dept. of Computer Science, Univ. of California - Santa Cruz, Santa Cruz, CA 95064, USA, alkis@cs.ucsc.edu
\textsuperscript{3} Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, USA, minos@research.bell-labs.com

\textbf{Abstract.} A key step in the optimization of declarative queries over XML data is estimating the selectivity of path expressions, i.e., the number of elements reached by a specific navigation pattern through the XML data graph. Recent studies have introduced XSketch structural graph synopses as an effective, space-efficient tool for the compile-time estimation of complex path-expression selectivities over graph-structured, schema-less XML data. Briefly, XSketches exploit localized graph stability and well-founded statistical assumptions to accurately approximate the path and branching distribution in the underlying XML data graph. Empirical results have demonstrated the effectiveness of XSketch summaries over real-life and synthetic data sets, and for a variety of path-expression workloads.

In this paper, we introduce fractional XSketches (fxSketches) a simple, yet intuitive and very effective generalization of the basic XSketch summarization mechanism. In a nutshell, our fxSketch synopsis extends the conventional notion of binary stability (employed in XSketches) with that of fractional stability, essentially recording more detailed path/branching distribution information on individual synopsis edges. As we demonstrate, this natural extension results in several key benefits over conventional XSketches, including (a) a simplified estimation framework, (b) reduced run-time complexity for the synopsis-construction algorithm, and (c) lifting the need for critical uniformity assumptions during estimation (thus resulting in more accurate estimates). Results from an extensive experimental study show that our fxSketch synopses yield significantly better selectivity estimates than conventional XSketches, especially in the context of complex path expressions with branching predicates.

\section{Introduction}

XML has rapidly evolved from a mark-up language to a de-facto standard for data exchange and integration over the web. A testament to this is the increasing volume of published XML data, together with the concurrent development of XML query processors that will allow users to tap into the vast amount of XML data available on the Internet. The successful deployment of such query processors depends crucially on the existence of high-level declarative query languages. There exist numerous proposals that cover a wide range of paradigms, but a common characteristic among all XML-language proposals is the use of path expressions as the basic method to access and retrieve specific
elements from the XML database. A path expression essentially defines a complex navigational path, which can be predicated on the existence of sibling paths or on constraints on the values of visited elements. As a concrete example, in a bibliography database, the path expression //author[book]/paper/sigmod/title (which adheres to the syntax of the standard XPath language [1]) selects the set of all title data elements discovered by the label path //author/paper/sigmod/title, but only for author elements that have at least one book child (a condition specified by the author[book] branch).

Similar to relational optimization, optimizing XML queries with complex path expressions depends crucially on the existence of concise summaries that can provide effective compile-time estimates for the selectivity of these expressions over the underlying (large) graph-structured XML database. This problem has recently attracted the attention of the database research community, and several techniques [2,3,4,5,6,7,8] have been proposed targeting different aspects of the problem. XSketch structural graph synopses [5,6] have recently been introduced as an effective data-reduction tool that enables accurate selectivity estimates for branching path expressions. In a nutshell, XSketch synopses exploit localized graph stability and well-founded statistical assumptions to accurately approximate the path and branching distribution in the underlying XML data graph; furthermore, XSketches can be augmented with summary information on data-value distributions to handle path expressions with value predicates [6]. Compared to previously proposed techniques, the XSketch synopsis mechanism targets the most general version of the estimation problem: XPath expressions with branching and value predicates, over graph-structured, schema-less XML databases. Experimental results with a variety of query workloads on different data sets have demonstrated the effectiveness of XSketches as concise summaries of XML data.

In this paper, we introduce fractional XSketches (fXSketches) a simple, yet intuitive and very effective generalization of the basic XSketch synopses based on the concept of fractional edge stabilities. Briefly, instead of simply recording whether a synopsis edge is stable or not (i.e., the conventional “binary” notion of stability employed in the XSketch model), our fXSketch synopses record the degree of stability for each edge as a fraction between 0 (“no-connection”) and 1 (“fully stable”). As we demonstrate, this natural generalization has a direct positive impact on the underlying estimation framework. First, it simplifies the expressions for query-selectivity estimates, thus allowing for faster estimation. Second, and perhaps most importantly, it lifts the need for certain critical uniformity assumptions during basic XSketch estimation, thus resulting in significantly more robust and accurate estimates. Furthermore, the removal of such uniformity assumptions also reduces the search space (and, therefore, the time complexity) of the synopsis-construction algorithm, since it effectively obviates the need for specialized synopsis-refinement operations to address regions of non-uniformity. These observations are backed up by an extensive experimental study which evaluates the performance of our generalized fXSketch synopses on a variety of XML data sets and query workloads. Our results clearly indicate that fXSketches yield significant improvements in accuracy when compared to original XSketch summaries. These improvements are more apparent in the case of complex path expressions with branching predicates, where the uniformity assumptions of the original XSketch model can introduce large errors;