Certain and Possible XPath Answers

Sara Cohen, Hebrew University

Querying of data sources culled from the Web, by naive users, is quite a difficult chore. This paper introduces a novel approach to ease the process of querying XML documents. Instead of specifying a query, the user simply marks positive examples $X^+$ of nodes that fit her information need. She may also mark negative examples $X^-$ of undesirable nodes. A deductive method, to suggest additional nodes that will interest the user, is discussed in this talk.

To be precise, a node $y$ is a certain answer if every query returning all positive examples $X^+$, and not returning any negative example from $X^-$, must also return $y$. Similarly, $y$ is a possible answer if there exists a query returning $X^+$ and $y$, while not returning any node in $X^-$. Thus, $y$ is likely to be of interest to the user if $y$ is a certain answer, and unlikely to be of interest if $y$ is not even a possible answer. The complexity of finding certain and possible answers, with respect to various classes of XPath, is considered. It is shown that for a wide variety of XPath queries (including child and descendant axes, wildcards, branching and attribute constraints), certain and possible answers can be found efficiently, provided that $X^+$ and $X^-$ are of bounded size. To prove this result a novel algorithm is developed.