

Approximate Matching in XML

Sihem Amer-Yahia

Nick Koudas

Divesh Srivastava

Abstract

The success of XML lies in its ability to easily represent homogeneous data as well as heterogeneous data. In particular, features such as optionality (e.g., a book may or may not have an associated cdrom), repetition (e.g., a chapter may have zero or more sections), alternation (e.g., a book may have either editors or authors), and nesting (e.g., a section may have nested sections) permit considerable variability among XML data conforming to the same schema. When querying or correlating such heterogeneous XML data, exact matching is typically inadequate, resulting in either too few or too many matches. Approximate matching, along with ranking the results of this matching, in the same spirit as Information Retrieval (IR) approaches, is more appropriate. Flexible specification of approximate matching over structure and content, and efficient evaluation of such specifications, create new challenges and exciting opportunities for the database research and development communities.

This tutorial surveys the research in the database and IR communities on this subject, including language proposals for the flexible specification of approximate matching in XML, and optimized evaluation strategies for approximate matching.

Sihem Amer-Yahia (<http://www.research.att.com/~sihem>) is a Senior Technical Staff Member in the Database Research Department at AT&T Labs-Research. She received the Engineer Degree from Institut National D'Informatique (INI), Algiers, Algeria, the MS in computer science from University Paris XI-Dauphine, France and the Ph.D. in computer science from University Paris XI-Orsay and INRIA, France. Her domains of interest are centralized/distributed query optimization and XML storage, loading, publishing and querying.

Nick Koudas (<http://www.research.att.com/~koudas>) is a Principal Technical Staff member in the Database Research Department at AT&T Labs-Research. He received his B.Tech. in Computer Science and Engineering from the University of Patras, Greece and his Ph.D. from the University of Toronto. His current research interests include XML databases, IP network data management and data quality.

Divesh Srivastava (<http://www.research.att.com/~divesh>) is the head of the Database Research Department at AT&T Labs-Research. He received his B.Tech. in Computer Science & Engineering from the Indian Institute of Technology, Bombay, India, and his Ph.D. in Computer Sciences from the University of Wisconsin, Madison, USA. His current research interests include XML databases and IP network data management.