Probabilistic Databases

Synthesis Lectures on Data Management

Editor
M. Tamer Özsu, ***University of Waterloo***

Synthesis Lectures on Data Management is edited by Tamer Özsu of the University of Waterloo. The series will publish 50- to 125 page publications on topics pertaining to data management. The scope will largely follow the purview of premier information and computer science conferences, such as ACM SIGMOD, VLDB, ICDE, PODS, ICDT, and ACM KDD. Potential topics include, but not are limited to: query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Probabilistic Databases
Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch
2011

Peer-to-Peer Data Management
Karl Aberer
2011

Probabilistic Ranking Techniques in Relational Databases
Ihab F. Ilyas and Mohamed A. Soliman
2011

Uncertain Schema Matching
Avigdor Gal
2011

Fundamentals of Object Databases: Object-Oriented and Object-Relational Design
Suzanne W. Dietrich and Susan D. Urban
2010

Advanced Metasearch Engine Technology
Weiyi Meng and Clement T. Yu
2010

Web Page Recommendation Models: Theory and Algorithms
Sule Gündüz-Ögüdücü
2010

Multidimensional Databases and Data Warehousing
Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen
2010

Database Replication
Bettina Kemme, Ricardo Jimenez Peris, and Marta Patino-Martinez
2010

Relational and XML Data Exchange
Marcelo Arenas, Pablo Barcelo, Leonid Libkin, and Filip Murlak
2010

User-Centered Data Management
Tiziana Catarci, Alan Dix, Stephen Kimani, and Giuseppe Santucci
2010

Data Stream Management
Lukasz Golab and M. Tamer Özsu
2010

Access Control in Data Management Systems
Elena Ferrari
2010

An Introduction to Duplicate Detection
Felix Naumann and Melanie Herschel
2010

Privacy-Preserving Data Publishing: An Overview
Raymond Chi-Wing Wong and Ada Wai-Chee Fu
2010

Keyword Search in Databases
Jeffrey Xu Yu, Lu Qin, and Lijun Chang
2009

Probabilistic Databases

Dan Suciu
University of Washington

Dan Olteanu
University of Oxford

Christopher Ré
University of Wisconsin-Madison

Christoph Koch
École Polytechnique Fédérale de Lausanne

SYNTHESIS LECTURES ON DATA MANAGEMENT #16

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Probabilistic databases are databases where the value of some attributes or the presence of some records are uncertain and known only with some probability. Applications in many areas such as information extraction, RFID and scientific data management, data cleaning, data integration, and financial risk assessment produce large volumes of uncertain data, which are best modeled and processed by a probabilistic database.

This book presents the state of the art in representation formalisms and query processing techniques for probabilistic data. It starts by discussing the basic principles for representing large probabilistic databases, by decomposing them into tuple-independent tables, block-independent-disjoint tables, or U-databases. Then it discusses two classes of techniques for query evaluation on probabilistic databases. In *extensional query evaluation*, the entire probabilistic inference can be pushed into the database engine and, therefore, processed as effectively as the evaluation of standard SQL queries. The relational queries that can be evaluated this way are called safe queries. In *intensional query evaluation*, the probabilistic inference is performed over a propositional formula called *lineage expression*: every relational query can be evaluated this way, but the data complexity dramatically depends on the query being evaluated, and can be #P-hard. The book also discusses some advanced topics in probabilistic data management such as top-k query processing, sequential probabilistic databases, indexing and materialized views, and Monte Carlo databases.

KEYWORDS

query language, query evaluation, query plan, data complexity, probabilistic database, polynomial time, sharp p, incomplete data, uncertain information

Contents