

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

	Preface: A Great Promise	xi
	Acknowledgments	xv
1	Overview	1
1.1	Two Examples	1
1.2	Key Concepts	5
1.2.1	Probabilities and their Meaning in Databases	5
1.2.2	Possible Worlds Semantics	5
1.2.3	Types of Uncertainty	6
1.2.4	Types of Probabilistic Databases	6
1.2.5	Query Semantics	6
1.2.6	Lineage	7
1.2.7	Probabilistic Databases v.s. Graphical Models	8
1.2.8	Safe Queries, Safe Query Plans, and the Dichotomy	9
1.3	Applications of Probabilistic Databases	10
1.4	Bibliographic and Historical Notes	13
2	Data and Query Model	17
2.1	Background of the Relational Data Model	17
2.2	The Probabilistic Data Model	19
2.3	Query Semantics	21
2.3.1	Views: Possible Answer Sets Semantics	22
2.3.2	Queries: Possible Answers Semantics	22
2.4	C-Tables and PC-Tables	23
2.5	Lineage	27
2.6	Properties of a Representation System	29
2.7	Simple Probabilistic Database Design	30
2.7.1	Tuple-independent Databases	31
2.7.2	BID Databases	35
2.7.3	U-Databases	37
2.8	Bibliographic and Historical Notes	41

3	The Query Evaluation Problem	45
3.1	The Complexity of $P(\Phi)$	45
3.2	The Complexity of $P(Q)$	48
3.3	Bibliographic and Historical Notes	51
4	Extensional Query Evaluation	53
4.1	Query Evaluation Using Rules	55
4.1.1	Query Independence	55
4.1.2	Six Simple Rules for $P(Q)$	56
4.1.3	Examples of Unsafe (Intractable) Queries	61
4.1.4	Examples of Safe (Tractable) Queries	62
4.1.5	The Möbius Function	65
4.1.6	Completeness	69
4.2	Query Evaluation using Extensional Plans	75
4.2.1	Extensional Operators	75
4.2.2	An Algorithm for Safe Plans	80
4.2.3	Extensional Plans for Unsafe Queries	81
4.3	Extensions	84
4.3.1	BID Tables	84
4.3.2	Deterministic Tables	86
4.3.3	Keys in the Representation	87
4.4	Bibliographic and Historical Notes	87
5	Intensional Query Evaluation	91
5.1	Probability Computation using Rules	92
5.1.1	Five Simple Rules for $P(\Phi)$	92
5.1.2	An Algorithm for $P(\Phi)$	96
5.1.3	Read-Once Formulas	98
5.2	Compiling $P(\Phi)$	99
5.2.1	d-DNNF ⁻	100
5.2.2	FBDD	101
5.2.3	OBDD	101
5.2.4	Read-Once Formulas	102
5.3	Approximating $P(\Phi)$	102
5.3.1	A deterministic approximation algorithm	102
5.3.2	Monte Carlo Approximation	104
5.4	Query Compilation	108

5.4.1	Conjunctive Queries without Self-Joins	109
5.4.2	Unions of Conjunctive Queries	110
5.5	Discussion	119
5.6	Bibliographic and Historical Notes	120
6	Advanced Techniques	123
6.1	Top- k Query Answering	123
6.1.1	Computing the Set Top_k	124
6.1.2	Ranking the Set Top_k	129
6.2	Sequential Probabilistic Databases	129
6.3	Monte Carlo Databases	134
6.3.1	The MCDB Data Model	134
6.3.2	Query Evaluation in MCDB	135
6.4	Indexes and Materialized Views	137
6.4.1	Indexes for Probabilistic data	137
6.4.2	Materialized Views for Relational Probabilistic Databases	140
	Conclusion	143
	Bibliography	145
	Authors' Biographies	163