

Scalable Processing of Spatial-Keyword Queries

Synthesis Lectures on Data Management

Editor

H.V. Jagadish, *University of Michigan*

Founding Editor

M. Tamer Özsu, *University of Waterloo*

Synthesis Lectures on Data Management is edited by H.V. Jagadish of the University of Michigan. The series publishes 80–150 page publications on topics pertaining to data management. Topics include 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.

Scalable Processing of Spatial-Keyword Queries

Ahmed R. Mahmood and Walid G. Aref
2019

Data Exploration Using Example-Based Methods

Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis
2018

Data Profiling

Ziawasch Abedjan, Lukasz Golab, Felix Naumann, and Thorsten Papenbrock
2018

Querying Graphs

Angela Bonifati, George Fletcher, Hannes Voigt, and Nikolay Yakovets
2018

Query Processing over Incomplete Databases

Yunjun Gao and Xiaoye Miao
2018

Natural Language Data Management and Interfaces

Yun Yao Li and Davood Rafiei
2018

Human Interaction with Graphs: A Visual Querying Perspective

Sourav S. Bhowmick, Byron Choi, and Chengkai Li

2018

On Uncertain Graphs

Arijit Khan, Yuan Ye, and Lei Chen

2018

Answering Queries Using Views

Foto Afrati and Rada Chirkova

2017

Databases on Modern Hardware: How to Stop Underutilization and Love Multicores

Anatasia Ailamaki, Erieta Liarou, Pinar Tözün, Danica Porobic, and Iraklis Psaroudakis

2017

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, Media Restore, and System Failover, Second Edition

Goetz Graefe, Wey Guy, and Caetano Sauer

2016

Generating Plans from Proofs: The Interpolation-based Approach to Query Reformulation

Michael Benedikt, Julien Leblay, Balder ten Cate, and Efthymia Tsamoura

2016

Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics

Laure Berti-Équille and Javier Borge-Holthoefer

2015

Datalog and Logic Databases

Sergio Greco and Cristina Molinaro

2015

Big Data Integration

Xin Luna Dong and Divesh Srivastava

2015

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, and Media Restore

Goetz Graefe, Wey Guy, and Caetano Sauer

2014

Similarity Joins in Relational Database Systems

Nikolaus Augsten and Michael H. Böhlen

2013

Information and Influence Propagation in Social Networks

Wei Chen, Laks V.S. Lakshmanan, and Carlos Castillo
2013

Data Cleaning: A Practical Perspective

Venkatesh Ganti and Anish Das Sarma
2013

Data Processing on FPGAs

Jens Teubner and Louis Woods
2013

Perspectives on Business Intelligence

Raymond T. Ng, Patricia C. Arocena, Denilson Barbosa, Giuseppe Carenini, Luiz Gomes, Jr.,
Stephan Jou, Rock Anthony Leung, Evangelos Milios, Renée J. Miller, John Mylopoulos, Rachel
A. Pottinger, Frank Tompa, and Eric Yu
2013

Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications

Amit Sheth and Krishnaprasad Thirunarayan
2012

Data Management in the Cloud: Challenges and Opportunities

Divyakant Agrawal, Sudipto Das, and Amr El Abbadi
2012

Query Processing over Uncertain Databases

Lei Chen and Xiang Lian
2012

Foundations of Data Quality Management

Wenfei Fan and Floris Geerts
2012

Incomplete Data and Data Dependencies in Relational Databases

Sergio Greco, Cristian Molinaro, and Francesca Spezzano
2012

Business Processes: A Database Perspective

Daniel Deutch and Tova Milo
2012

Data Protection from Insider Threats

Elisa Bertino
2012

Deep Web Query Interface Understanding and Integration

Eduard C. Dragut, Weiyi Meng, and Clement T. Yu

2012

P2P Techniques for Decentralized Applications

Esther Pacitti, Reza Akbarinia, and Manal El-Dick

2012

Query Answer Authentication

HweeHwa Pang and Kian-Lee Tan

2012

Declarative Networking

Boon Thau Loo and Wenchao Zhou

2012

Full-Text (Substring) Indexes in External Memory

Marina Barsky, Ulrike Stege, and Alex Thomo

2011

Spatial Data Management

Nikos Mamoulis

2011

Database Repairing and Consistent Query Answering

Leopoldo Bertossi

2011

Managing Event Information: Modeling, Retrieval, and Applications

Amarnath Gupta and Ramesh Jain

2011

Fundamentals of Physical Design and Query Compilation

David Toman and Grant Weddell

2011

Methods for Mining and Summarizing Text Conversations

Giuseppe Carenini, Gabriel Murray, and Raymond Ng

2011

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

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2019

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Scalable Processing of Spatial-Keyword Queries

Ahmed R. Mahmood and Walid G. Aref

ISBN: 978-1-031-00719-2 paperback

ISBN: 978-3-031-01867-1 ebook

ISBN: 978-1-031-00094-2 hardcover

DOI 10.1007/978-3-031-01867-1

A Publication in the Springer series

SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #56

Series Editor: H.V. Jagadish, *University of Michigan*

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

Series ISSN

Print 2153-5418 Electronic 2153-5426

Scalable Processing of Spatial-Keyword Queries

Ahmed R. Mahmood and Walid G. Aref
Purdue University

SYNTHESIS LECTURES ON DATA MANAGEMENT #56



ABSTRACT

Text data that is associated with location data has become ubiquitous. A tweet is an example of this type of data, where the text in a tweet is associated with the location where the tweet has been issued. We use the term spatial-keyword data to refer to this type of data. Spatial-keyword data is being generated at massive scale. Almost all online transactions have an associated spatial trace. The spatial trace is derived from GPS coordinates, IP addresses, or cell-phone-tower locations. Hundreds of millions or even billions of spatial-keyword objects are being generated daily. Spatial-keyword data has numerous applications that require efficient processing and management of massive amounts of spatial-keyword data.

This book starts by overviewing some important applications of spatial-keyword data, and demonstrates the scale at which spatial-keyword data is being generated. Then, it formalizes and classifies the various types of queries that execute over spatial-keyword data. Next, it discusses important and desirable properties of spatial-keyword query languages that are needed to express queries over spatial-keyword data. As will be illustrated, existing spatial-keyword query languages vary in the types of spatial-keyword queries that they can support.

There are many systems that process spatial-keyword queries. Systems differ from each other in various aspects, e.g., whether the system is batch-oriented or stream-based, and whether the system is centralized or distributed. Moreover, spatial-keyword systems vary in the types of queries that they support. Finally, systems vary in the types of indexing techniques that they adopt. This book provides an overview of the main spatial-keyword data-management systems (SKDMSs), and classifies them according to their features. Moreover, the book describes the main approaches adopted when indexing spatial-keyword data in the centralized and distributed settings. Several case studies of SKDMSs are presented along with the applications and query types that these SKDMSs are targeted for and the indexing techniques they utilize for processing their queries.

Optimizing the performance and the query processing of SKDMSs still has many research challenges and open problems. The book concludes with a discussion about several important and open research-problems in the domain of scalable spatial-keyword processing.

KEYWORDS

spatial-keyword, indexing, systems, big data, query processing

To my wife, my parents, and my children.

Ahmed R. Mahmood

*To my beloved parents, professors Safaa Elhifni and Galal Aref,
and my lovely grandchildren, Hana and Walid.*

Walid G. Aref

Contents

Preface	xv
Acknowledgments	xvii
1 Introduction	1
1.1 Spatial-Keyword Data	1
1.2 Spatial-Keyword Applications	2
2 Querying Spatial-Keyword Data	7
2.1 Spatial-Keyword Query Predicates	7
2.1.1 The Spatial-Keyword Select Predicate	7
2.1.2 The Spatial-Keyword Join Predicates	11
2.1.3 The Spatial-Keyword Group Predicate	12
2.1.4 Continuous Spatial-Keyword Queries	14
2.1.5 Aggregate Spatial-Keyword Predicates	16
2.1.6 Distance Metrics in Spatial-Keyword Predicates	17
2.2 Spatial-Keyword Query Languages	17
2.2.1 GNIP	18
2.2.2 Microblogs Query Language (MQL)	20
2.2.3 Atlas	21
3 Centralized Spatial-Keyword Query Processing	25
3.1 Spatial Indexing	25
3.1.1 Space-Driven Indexes	26
3.1.2 Data-Driven Spatial Indexes	29
3.2 Text Indexes	31
3.3 Spatial-Keyword Indexes	34
3.3.1 Space-First Spatial-Keyword Indexing	35
3.3.2 Text-First Indexing	38
3.3.3 Interleaved Spatial-Keyword Indexing	39
3.3.4 Separate Spatial and Keyword Indexes	42
3.3.5 Spatiotemporal-Keyword Indexing	42
3.4 Case Studies	43

4	Distributed Spatial-Keyword Processing	49
4.1	General-Purpose Big-Data Systems	49
4.1.1	Batch-Oriented Systems	49
4.1.2	Big-Data Streaming Systems	50
4.2	Big Spatial-Keyword Data Management Systems	51
4.2.1	Application on Top of Existing Big-Data System	52
4.2.2	Extension to General-Purpose Big-Data Systems	54
4.2.3	Dedicated Big Spatial-Keyword Data Management Systems	57
4.3	Strategies to Distribute and Index Data in Big SKDMSs	59
4.3.1	Spatial-Only Indexing	59
4.3.2	Hash-Based Indexing	60
4.3.3	Hybrid Indexing	62
4.3.4	Partitioning of Spatial-Keyword Data Streams	65
4.4	Case Studies	68
5	Open Research Problems in Spatial-Keyword Processing	73
5.1	Load Balancing	73
5.2	Spatial-Keyword Benchmarks	73
5.3	Spatial-Keyword Query Optimizer	75
5.4	Big Spatio-Temporal Keyword Query Processing	75
	Bibliography	77
	Authors' Biographies	97

Preface

Scalable processing of spatial-keyword data is an important problem that is being studied actively by database researchers and industrial practitioners. This book targets developers and researchers that are interested in the design of scalable spatial-keyword data management systems (SKDMSs). The book surveys and classifies the main spatial-keyword queries, query languages, indexing techniques, and processing models, and aims to aid developers and researchers improve the scalability, efficiency, and performance of SKDMSs. There are many open research problems in the area of spatial-keyword processing. In this book, we highlight some of these problems to motivate the engagement of database designers and researchers in addressing them. It does not require prior knowledge about spatial-keyword processing or systems. Basic knowledge of database systems and data indexing techniques are all that are needed.

Ahmed R. Mahmood and Walid G. Aref
January 2019

Acknowledgments

Walid G. Aref's research has been partially supported by the National Science Foundation under Grant III-1815796.