Query Answer Authentication

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.

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

Query Answer Authentication

HweeHwa Pang and Kian-Lee Tan

ISBN: 978-3-031-00759-0 paperback
ISBN: 978-3-031-01887-9 ebook

DOI 10.1007/978-3-031-01887-9

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #24
Series Editor: M. Tamer Özsu, *University of Waterloo*
Series ISSN
Synthesis Lectures on Data Management
Print 2153-5418 Electronic 2153-5426

Query Answer Authentication

HweeHwa Pang
Singapore Management University

Kian-Lee Tan
National University of Singapore

SYNTHESIS LECTURES ON DATA MANAGEMENT #24

ABSTRACT

In data publishing, the owner delegates the role of satisfying user queries to a third-party publisher. As the servers of the publisher may be untrusted or susceptible to attacks, we cannot assume that they would always process queries correctly, hence there is a need for users to authenticate their query answers.

This book introduces various notions that the research community has studied for defining the correctness of a query answer. In particular, it is important to guarantee the completeness, authenticity and minimality of the answer, as well as its freshness. We present authentication mechanisms for a wide variety of queries in the context of relational and spatial databases, text retrieval, and data streams. We also explain the cryptographic protocols from which the authentication mechanisms derive their security properties.

KEYWORDS

correctness of query answer, data integrity, database security, outsourced database

Contents

Preface xi

Acknowledgments xiii

1 Introduction 1

1.1 System and Threat Models 2
1.2 Cost Factors 3
1.3 Organization 4

2 Cryptography Foundation 5

2.1 Cryptographic Protocols 5
2.2 Cost of Cryptographic Protocols 8

3 Relational Queries 11

3.1 Background 11
3.2 Selection Query 12
3.2.1 Overview of Authentication Techniques for Range Selection 12
3.2.2 Merkle Hash Tree Technique 13
3.2.3 Signature Aggregation Technique 18
3.3 Projection Query 22
3.4 Join Query 23
3.5 Aggregation Query 26
3.5.1 Partial Sum Hierarchy 27
3.5.2 Certified Partial Sum Hierarchy 29
3.5.3 Query Processing using the Partial Sum Hierarchy 31
3.5.4 Data Organization 31
3.5.5 Extension to Other Aggregation Functions 32
3.5.6 Other Techniques Related To Aggregation Query 33
3.6 Summary 34

4 Spatial Queries **37**
4.1 Background 37
4.2 Authenticating Window Query 39
4.2.1 Merkle Hash-based Schemes 40
4.2.2 Signature-Chain based Schemes 42
4.3 Authenticating kNN/Range Query 46
4.3.1 A Computational Geometry Approach 47
4.3.2 A Signature-based Scheme 50
4.4 Authenticating Reverse Nearest Neighbor Queries 52
4.5 Summary 55

5 Text Search Queries **57**
5.1 Background on Text Search 57
5.2 Problem Formulation 59
5.3 Choice of Authentication Approach 61
5.4 Threshold with Random Access 62
5.4.1 Authentication with Merkle Hash Trees 65
5.4.2 Authentication with Chain Merkle Hash Trees 67
5.5 Threshold with No Random Access 68
5.6 Summary 71

6 Data Streams **73**
6.1 Data Stream Model 73
6.2 Tumbling Merkle Tree 74
6.2.1 TM-Tree Construction 74
6.2.2 One-Shot Selection Queries 75
6.2.3 Sliding Window Selection Queries 75
6.2.4 Aggregation Queries 76
6.3 Continuous Monitoring of Tuples in a Selection Range 76
6.4 Summary 78

7 Conclusion **79**

Bibliography **81**

Authors' Biographies **89**

Preface

In the data publishing model, also known as database-as-a-service, a data owner outsources the database management functionalities to a third-party publisher. Users who need to access the database of the owner will then submit their queries to the publisher. The model is gaining popularity commercially as it reduces the total cost of ownership (in terms of manpower, hardware and software costs) to the data owner and offers quality service to the consumers of the data.

As the publisher may be malicious or its systems may be vulnerable to security breaches, one key challenge in data publishing is to ensure query answer authenticity: Given that the publisher may not be trusted, it is critical for end users to have an assurance that the query answers returned by the publisher are indeed the same answers that the owner would have given. This book brings together a collection of authentication mechanisms that have been investigated for various application domains to address the challenge.

In Chapter 1, we begin by introducing the data publishing model. We discuss its benefits, and formulate the associated system and threat models. We then bring out the need for users to check the authenticity of their query answers, and summarize the cost factors to take into account in developing a query authentication mechanism.

In Chapter 2, we lay the foundation and background knowledge for the rest of the book. In particular, we review cryptographic protocols that form the building blocks for the query answer authentication schemes that follow. We also examine the costs of various cryptographic primitives.

Chapter 3 presents query authentication schemes for relational databases. We identify the key requirements of an effective authentication scheme. We examine schemes built upon both Merkle Hash Tree and signature aggregation for a wide variety of relational queries, including selection, projection, join as well as aggregation. We also look at incorporating authentication information into index structures in order to facilitate efficient processing.

In Chapter 4, we focus on authentication mechanisms for spatial databases. In particular, we present authentication methods for window, range, kNN and RNN queries. These methods are based on Merkle Hash Tree, signature chain as well as geometry, and employ spatial data structures like R-tree and KD-tree.

Unlike traditional relational and spatial databases, text search offers a different set of challenges.

Even if the search engine returns all the relevant documents, it may alter their ranking within the result. In Chapter 5, we investigate methods for guaranteeing the correctness of query answers for text search. These methods require pre-certification of the inverted lists that store the frequency of every combination of document and term, from which document scores are computed. Adaptations of the threshold algorithm to support query answer authentication are presented.

In Chapter 6, we describe techniques for authenticating streaming data. Here, authentication mechanisms must additionally ensure that the relative order of each datum is preserved. We present methods based on Merkle Hash Tree for sliding window queries as well as aggregation sliding window queries.

Finally, Chapter 7 concludes the book. Query answer authentication is a relatively young field that is becoming increasingly important. There are many outstanding challenges and issues that merit further research before the field matures. We highlight some of the more interesting ones in this chapter.

This book can be used as a reference for a variety of audience. It can serve as a reference text in graduate level database security courses that cover query authentication. It also provides a good survey to graduate students working on securing databases under the database-as-a-service model (e.g., database outsourcing and cloud computing). Researchers, technologists and developers will also find this book a good source for learning more about assuring users of the authenticity of their query answers

HweeHwa Pang and Kian-Lee Tan
February 2012

Acknowledgments

We are thankful to several people who have made this book possible. In particular, M. Tamer Özsu, Editor of the Synthesis Lectures on Data Management, offered us the opportunity to write this book. Tamer also read an earlier draft of the book and provided valuable comments that improve its literary style. Throughout this project, Diane D. Cerra, the Executive Editor of Morgan & Claypool, provided us with excellent editorial support necessary for the completion of this book.

HweeHwa Pang and Kian-Lee Tan
February 2012