Similarity Joins in Relational Database Systems

Synthesis Lectures on Data Management

Synthesis Lectures on Data Management is edited by Tamer Özsu of the University of Waterloo. The series publishes 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 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

Similarity Joins in Relational Database Systems

Nikolaus Augsten and Michael H. Böhlen

ISBN: 978-3-031-00723-1 paperback
ISBN: 978-3-031-01851-0 ebook

DOI 10.1007/978-3-031-01851-0

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #38
Series Editor: M. Tamer Özsu, *University of Waterloo*
Series ISSN
Synthesis Lectures on Data Management
Print 2153-5418 Electronic 2153-5426

Similarity Joins in Relational Database Systems

Nikolaus Augsten
University of Salzburg

Michael H. Böhlen
University of Zürich

SYNTHESIS LECTURES ON DATA MANAGEMENT #38

ABSTRACT

State-of-the-art database systems manage and process a variety of complex objects, including strings and trees. For such objects equality comparisons are often not meaningful and must be replaced by similarity comparisons. This book describes the concepts and techniques to incorporate similarity into database systems. We start out by discussing the properties of strings and trees, and identify the edit distance as the de facto standard for comparing complex objects. Since the edit distance is computationally expensive, token-based distances have been introduced to speed up edit distance computations. The basic idea is to decompose complex objects into sets of tokens that can be compared efficiently. Token-based distances are used to compute an approximation of the edit distance and prune expensive edit distance calculations.

A key observation when computing similarity joins is that many of the object pairs, for which the similarity is computed, are very different from each other. Filters exploit this property to improve the performance of similarity joins. A filter preprocesses the input data sets and produces a set of candidate pairs. The distance function is evaluated on the candidate pairs only. We describe the essential query processing techniques for filters based on lower and upper bounds. For token equality joins we describe prefix, size, positional and partitioning filters, which can be used to avoid the computation of small intersections that are not needed since the similarity would be too low.

KEYWORDS

strings, trees, similarity, edit distance, q-grams, pq-grams, token-based distance, lower bound, upper bound, similarity join

To Leni, Magdalena, and Katharina.
Nikolaus

To Franziska, Chantal, and Pascal.
Michael

Contents

Preface

During the last few decades database systems have evolved substantially and today it is common for database systems to manage a large variety of complex objects. At the physical level, there has been significant progress in terms of storing and processing such complex objects. At the logical level, however, complex objects remain a challenge. A key reason is that equality, which is appropriate for simple objects, is often ineffective for complex objects. In this book we describe the essential concepts and techniques toward a principled solution for processing complex objects that must be compared in terms of similarity rather than equality.

An intuitive approach to define the similarity of complex objects is the edit distance, i.e., the number of basic edit operations that are required to transform one object into another. The intuitive nature of the edit distance is the reason why edit distances have become the de facto standard for complex objects. We define the string and tree edit distance, give algorithms to compute these distances, and work out the essential properties that support the effective and efficient processing of complex objects.

Token distances, which decompose complex objects into sets of tokens, have been proposed to deal with the high computational cost of edit distances. The token distance is computed by comparing the token sets that are the result of the decompositions. The more similar two token sets are the smaller is their distance. Token sets are compared by counting the number of identical elements in the sets. Set intersection is an operation that is well supported by database systems and scales to large sets. We survey the different techniques to compute and process token sets, and we discuss in detail three representative decomposition techniques: strings with q-grams, ordered trees with pq-grams, and unordered trees with windowed pq-grams.

Determining the exact distance between complex objects, particularly for joins where all pairs of objects must be compared, is often too expensive. To reduce the costs of such computations, filter and refine approaches have been developed. The goal of the filter step is to cheaply identify candidate pairs for which the exact similarity must be computed. Non-candidate pairs do not have to be considered because their similarity is not sufficient to be included in the result. We describe various filter techniques, and provide lower and upper bounds that can be used to efficiently compute similarity joins.

The book uses strings and trees as representative examples of complex objects. The techniques discussed in this book, however, are general and are also applicable to other types of objects. In particular, graphs are an important data structure that recently have received a lot of attention, and for which edit- and token-based distances have been proposed. At the relevant places we provide references to the vibrant and emerging field of graphs in databases.

The book is intended as a starting point for researchers and students in the database field who would like to learn more about similarity in database systems. Throughout, we offer precisely defined concepts and properties, and we illustrate these with representative and carefully chosen examples. We made an effort to include precise definitions, theorems, and examples, but at the same time kept the description at a level that is understandable to a general audience with an academic background. Much of the material presented in this book has been used in courses taught during the last few years at the Free University of Bozen-Bolzano, Technische Universität München, University of Salzburg, and University of Zürich. Our warm thanks goes to the students who provided constructive feedback and helped to advance the material presented in this book.

Nikolaus Augsten and Michael H. Böhlen
October 2013

Acknowledgments

Several people offered valuable support during the preparation of this book. We warmly thank Tamer Özsu for inviting us to write this lecture, and we thank Diane Cerra for her perfect management of the entire process. We thank both Tamer and Diane for their kind but firm pushes that helped us to make this lecture happen.

Part of the book's material evolved from courses and lectures taught at the Free University of Bozen-Bolzano, Technische Universität München, the University of Salzburg, and the University of Zürich. We thank the many students for their constructive comments and their patience with early versions of the course material.

Tatsuya Akutsu and Robert Elsässer contributed with feedback about the unordered tree edit distance and its complexity, Willi Mann helped with proofreading the book. We would also like to acknowledge the many collaborators and friends who, through discussions and comments have shaped our thinking and understanding of the area: Arturas Mazeika, Chen Li, Johann Gamper, Denilson Barbosa, Themis Palpanas, Curtis Dyreson, Mateusz Pawlik, Jan Finis, Martin Raiber, Theo Härder, Leonardo Andrade Ribeiro, Benjamin Gufler, Gerard Lemson, Walter Costanzi, Franco Barducci, and Roberto Loperfido.

Finally, we would like to acknowledge our funding sources: Michael Böhlen's work is supported by the University of Zürich, the Swiss National Science Foundation, and the Free University of Bozen-Bolzano. Nikolaus Augsten's work is supported by the University of Salzburg, the Free University of Bozen-Bolzano, the Department for Promotion of Educational Policies, University and Research of the Autonomous Province of Bolzano - South Tyrol, and Technische Universität München.

Nikolaus Augsten and Michael H. Böhlen
October 2013