Synthesis Lectures on Data Management

SYNTHESIS
COLLECTION OF TECHNOLOGY

Sourav S. Bhowmick · Byron Choi

Plug-and-Play Visual Subgraph Query Interfaces

Springer

Synthesis Lectures on Data Management

This series publishes lectures on 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.

Sourav S. Bhowmick · Byron Choi

Plug-and-Play Visual Subgraph Query Interfaces

Springer

Sourav S. Bhowmick
School of Computer Science and Engineering
Nanyang Technological University
Singapore, Singapore

Byron Choi
Hong Kong Baptist University
Hong Kong S.A.R., China

ISSN 2153-5418 ISSN 2153-5426 (electronic)
Synthesis Lectures on Data Management
ISBN 978-3-031-16161-2 ISBN 978-3-031-16162-9 (eBook)
https://doi.org/10.1007/978-3-031-16162-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword by the Series Editor

Graph data has been gaining importance over the years. More and more situations have data best represented in graph form, ranging from knowledge bases to social networks. This has led to considerable interest in querying graph data. Not surprisingly, there is a great deal of work in this area and many clever ways suggested in the literature.

Any query language must have an underlying logic and formalism. The most natural way to express a query is through a textual statement of this underlying logic. A prime example of this is SQL, whose core is a text representation of relational algebra. However, users often find it difficult to express their query needs in the text that correctly states their desire in terms of the logic of the stored data. Hence, there has been a search to define more usable query interfaces. Visual query interfaces have been particularly important in this regard.

With structured data, visual query interfaces are designed with the structure internalized. For example, if we assume that a relational database comprises tables connected by means of primarykey/foreignkey joins, we can define a visual query language that relies on this structure: for instance, a user could specify values for some attributes in a relation and some attributes in a joined relation. However, graph data tends to be much richer. For example, it is meaningful for a user to query for a node with high degree—a query with no obvious parallel in the relational world. This makes it challenging to design a visual query interface for graph data.

This monograph presents what the authors call a "Plug and Play" interface for graph data. The idea is to have a (large) library of templates that cover most common cases, and a simple specification mechanism that appropriately instantiates selected templates in a manner specific to a particular graph database. The end result is a customized, easy-to-use interface.

This Synthesis Lectures series includes many works of importance for the field of databases, with their importance being derived from diverse dimensions. Some books are important because they present an excellent survey of the state of the art in a topic of great interest; others are important because they cover a particularly novel research direction that a research group is pursuing, whose holistic presentation in a book format permits the authors to argue for their research vision in a manner not possible in focused research

papers with tight page limits. This particular book is a great example of the latter: it advances the frontiers of our field and promotes discussion of a new approach. Please enjoy reading it, then discuss, criticize, and praise, as you see fit.

Ann Arbor, MI, USA H. V. Jagadish

Preface

If the user can't use it, it doesn't work.
Susan Dray, President, Dray & Associates, Inc.

Law is not primarily for lawyers or judges—it applies to everyone. However, most people are unable to comprehend legal language on their own. Though technically they can access the law, they are not in any position to vindicate their own rights or defend themselves against legal challenges. One of the reasons that prevent public access to justice is the challenges brought by the usage of legal terminology. For example, the archaic terms "in camera" and "subpoena" are not only difficult to understand by the public, but also may create misunderstandings. For instance, "in camera" may be interpreted by one as appearing in the courthouse through Zoom! Certainly, it is much more intuitive to replace these terms with "in private" and "order to attend court", respectively. Such simplification can potentially make an ordinary person's experience with law more palatable, thereby enabling greater access to justice. The law impacts everyone and hence should be comprehensible by everyone.

Data management tools, like law, are no more primarily for database experts and administrators. It should be accessible to everyone in an increasingly data-democratized and data-driven world. However, query languages—the primary means to access data residing in databases—prevent diverse end users who are not proficient in these languages to take advantage of these tools for their tasks. That is, query languages are like legal terminology that can only be understood and written by database professionals and experts. Since data impacts almost all aspects of life nowadays, it should be easily accessed and searched by end users with diverse skills and backgrounds.

Visual query interfaces are designed to alleviate the access challenge by enabling end users to access and search data through the interactive construction of queries without resorting to any query languages. Given the ubiquity of graphs to model data in a wide variety of domains (e.g., biology, chemistry, ecology, social science, and journalism), this book reports recent work in building visual query interfaces to democratize access to graph data.

Subgraph search query, which is typically represented as a connected graph, is one of the most popular query paradigms for accessing graph data. Since graphs are intuitive to draw, increasingly graph data management tools from academia and industry are exposing visual subgraph query interfaces (VQIs) to enable an end user to draw a

subgraph search query interactively instead of formulating it textually using a graph query language. However, these classical VQIs suffer from several limitations such as high creation and maintenance cost, lack of superior support for visual subgraph query formulation, and poor portability across application domains and data sources that hinder their democratization. This book presents the paradigm of *plug-and-play* VQI, as it stands today, that addresses these limitations. In particular, a broad goal of this book is to draw on well-founded principles of human-computer interaction (HCI) and cognitive psychology to enhance the usability and reach of subgraph query formulation frameworks. Note that it is reasonable to expect this picture to evolve with time.

Our discussion is divided into four parts, moving from "softer" aspects of visual interfaces (e.g., usability, cognitive load) to "harder" aspects of realizing them (e.g., algorithms and data structures) in order to build plug-and-play visual subgraph query interfaces. First, we review, as accurately as possible, a spectrum of classical visual interfaces to enable subgraph query formulation. We discuss their advantages and limitations w.r.t. usability and their impact on the democratization of subgraph querying tools to wider communities.

Second, we introduce the novel paradigm of plug-and-play visual subgraph query interface (i.e., PnP interface). In particular, we describe its architecture, how it can address the limitations of classical VQIs, and the challenges that need to be addressed in order to realize it in practice.

Third, we review frameworks that construct and maintain PnP interfaces. Specifically, we introduce recent visual subgraph query formulation frameworks that depart from the traditional mantra of "manual" VQI construction by exploring a paradigm that automatically generates and maintains a VQI for a given graph data source in a data-driven manner without resorting to any coding. A user can simply plug a PnP interface on top of his or her graph data source and play by formulating subgraph queries visually without resorting to any graph query languages. In particular, a pervasive desire of this review is to emphasize the role of cognitive load-aware "representative objects" in a VQI that facilitates top-down and bottom-up query formulation effortlessly.

The last topic consists of several open problems in this young field. The list presented should by no means be considered exhaustive and is centered around challenges and issues currently in vogue. Nevertheless, readers can benefit by exploring the research directions given in this part.

The book is suitable for use in advanced undergraduate and graduate-level courses on graph data management. It has sufficient material that can be covered as part of a semester-long course, thereby leaving plenty of room for an instructor to choose topics. An undergraduate course in algorithms, graph theory, database technology, and basic HCI should suffice as a prerequisite for most of the chapters. A good knowledge of C++/Java programming language is sufficient to code the algorithms described herein. We have also made the code base of some of the frameworks available through GitHub links. For completeness, we have provided background information on several topics in Chap. 2:

fundamental graph and subgraph query terminology and concepts related to HCI and cognitive psychology. The knowledgeable reader may omit this chapter and perhaps refer back to it while reading later chapters of the book.

We hope that this book will serve as a catalyst in helping this burgeoning area of plug-and-play query interfaces that lie at the intersection of data management, HCI, and cognitive psychology to grow and have a practical impact.

Singapore, Singapore — Sourav S. Bhowmick
Hong Kong S.A.R., China — Byron Choi
July 2022

Acknowledgments

It is a great pleasure for us to acknowledge the assistance and contributions of a large number of individuals to this effort. First, we would like to thank our publishers Morgan & Claypool and Springer Nature for their support. In particular, we would like to acknowledge the efforts, help, and patience of Diane Cerra and Christine Kiilerich, our primary contacts for this edition.

The majority of the work reported in this book grew out of the **DA**ta-driven **V**isual **IN**terface **C**onstruction Eng**I**ne (DAVINCI) project at the Nanyang Technological University (NTU), Singapore. In this project, our broad goal is to explore the paradigm of *data-driven* visual query interface construction to enable effective top-down and bottom-up searches. Specifically, some of the chapters are published in ACM SIGMOD and VLDB, two premium data management venues. Details related to the DAVINCI project can be found at https://personal.ntu.edu.sg/assourav/research/hint/index.html.

Dr. Huey-Eng Chua of NTU, who was a key collaborator for this project, deserves the first thank you. She continuously provided high-quality management of this project by working with all stakeholders. This project would not have been successful without her contributions.

In addition, we would also like to express our gratitude to all the group members and collaborators, past and present. In particular, Kai Huang (NTU & Fudan University), Zifeng Yuan (NTU & Fudan University), Zekun Ye (NTU & Fudan University), Prof. Curtis Dyreson (Utah State University), Prof. Shuigeng Zhou (Fudan University), and Prof. Wook-Shin Han (POSTECH) made substantial contributions to the broader aspect of our research on PnP interfaces.

Quite a few people have helped us with the initial vetting of the text for this book. It is our pleasure to acknowledge them all here. We would like to thank Springer Nature for carefully proofreading the complete book in a short span of time and suggesting the changes which have been incorporated.

We would like to acknowledge our parents and family members who gave us incredible support throughout the years. They were the major force behind our continuous strive for breaking out from the comfort zone of computer science to explore problems that are at the intersection of two or more disparate areas and along the way appreciate the

importance of softer aspects of technology. It has been and continues to be a great learning experience for us. A special thanks go to Professor H. V. Jagadish (UMich, USA) for giving us the opportunity to author this book.

Finally, we would like to thank the MOE Singapore AcFR Tier 1 and Tier 2, for the generous financial support provided for the DAVINCI project. We would also like to thank the School of Computer Science and Engineering at the Nanyang Technological University for allowing the use of their resources to help complete the book. The work done at the Department of Computer Science at HKBU is partially supported by HKRGC GRF 12201119 and 12201518.

July 2022

Sourav S. Bhowmick
Byron Choi

Contents

About the Authors

Sourav S. Bhowmick is an Associate Professor at the School of Computer Science and Engineering (SCSE), Nanyang Technological University, Singapore. His core research expertise is in data management, human–data interaction, and data analytics. His research has appeared in premium venues such as ACM SIGMOD, VLDB, ACM WWW, ACM MM, ACM SIGIR, VLDB Journal, Bioinformatics, and Biophysical Journal. He is a co-recipient of Best Paper Awards in ACM CIKM 2004, ACM BCB 2011, and VLDB 2021 for work on mining structural evolution of tree-structured data, generating functional summaries, and scalable attributed network embedding, respectively. He is a co-recipient of the 2021 ACM SIGMOD Research Highlights Award. Sourav is serving as a member of the SIGMOD Executive Committee, a regular member of the PVLDB Advisory Board, and a co-lead in the committee for Diversity and Inclusion in Database Conference Venues. He is a co-recipient of the VLDB Service Award in 2018 from the VLDB Endowment. He was inducted into Distinguished Members of the ACM in 2020.

Byron Choi is the Associate Head and an Associate Professor at the Department of Computer Science, Hong Kong Baptist University (HKBU). His research interests include graph-structured databases, database usability, database security, and time series analysis. Byron's publications have appeared in premium venues such as TKDE, VLDBJ, SIGMOD, PVLDB/VLDB, and ICDE. He has served as a program committee member or reviewer of premium conferences and journals, including PVLDB, VLDBJ, ICDE, TKDE, and TOIS. He was awarded a distinguished program committee (PC) member from ACM SIGMOD 2021 and the best reviewers award from ACM CIKM 2021. He received the distinguished reviewer award from PVLDB 2019. He has served as the director of a Croucher Foundation Advanced Study Institute (ASI), titled "Frontiers in Big Data Graph Research", in 2015. He was a recipient of the HKBU President's Award for Outstanding Young Researcher in 2016.

The Future is Democratized Graphs

1

Graphs (a.k.a networks) are ubiquitous nowadays in many application domains (e.g., retail and eCommerce, transportation and logistics, healthcare, pharmaceuticals, and life sciences) as they provide powerful abstractions to model complex structures and relationships. Consequently, graph data management tools are expected to play a pivotal role in diverse applications such as customer analytics, fraud detection and prevention, supply chain management, and scientific data analysis. *Markets and Markets* anticipates the global graph database market size is expected to grow from USD 1.9 billion in 2021 to USD 5.1 billion by 2026 (Markets and Markets 2022). Given such growth opportunities, it is paramount for graph data management tools to be user-friendly, efficient, and scalable to support their growing demands from *diverse* end users and applications.

1.1 Querying Graphs

Querying graphs is a key component in any graph data management tool. Although keyword-based search (Wang and Aggarwal 2010) is the simplest paradigm to query graphs, such queries have limited flexibility as they disallow the specification of structural constraints on graphs. Consequently, the most common and important query primitive for graphs is subgraph search (also referred to as subgraph or graph query), where we want to retrieve one or more subgraphs in a graph G that *exactly* or *approximately* match a user-specified query graph Q. Exact subgraph search strictly searches for isomorphic subgraphs in G that matches Q. These queries are typically referred to as *subgraph matching* (Sun and Luo 2020) or *subgraph enumeration* (Afrati et al. 2013) query based on whether Q is a labeled or unlabeled query graph, respectively. On the other hand, a similar or approximate search allows the topology of the query graph to be mismatched to a certain degree. These approaches utilize edit distance (Bunke and Kim 1998), common connected subgraphs (Shang et al. 2010), or graph homomorphism (Fan et al. 2010; Song et al. 2018) to retrieve *similar* query results.

S. S. Bhowmick and B. Choi, *Plug-and-Play Visual Subgraph Query Interfaces*, Synthesis Lectures on Data Management, https://doi.org/10.1007/978-3-031-16162-9_1