Similarity Search
The Metric Space Approach

ADVANCES IN DATABASE SYSTEMS

Series Editor

Ahmed K. Elmagarmid

Purdue University
West Lafayette, IN 47907

Other books in the Series:

STREAM DATA MANAGEMENT, *Nauman Chaudhry, Kevin Shaw, Mahdi Abdelguerfi*, ISBN: 0-387-24393-3

FUZZY DATABASE MODELING WITH XML, *Zongmin Ma,* ISBN: 0-387-24248-1

MINING SEQUENTIAL PATTERNS FROM LARGE DATA SETS, *Wei Wang and Jiong Yang;* ISBN: 0-387-24246-5

ADVANCED SIGNATURE INDEXING FOR MULTIMEDIA AND WEB APPLICATIONS, *Yannis Manolopoulos, Alexandros Nanopoulos, Eleni Tousidou;* ISBN: 1-4020-7425-5

ADVANCES IN DIGITAL GOVERNMENT, *Technology, Human Factors, and Policy,* *edited by William J. McIver, Jr. and Ahmed K. Elmagarmid;* ISBN: 1-4020-7067-5

INFORMATION AND DATABASE QUALITY, *Mario Piattini, Coral Calero and Marcela Genero;* ISBN: 0-7923- 7599-8

DATA QUALITY, *Richard Y. Wang, Mostapha Ziad, Yang W. Lee:* ISBN: 0-7923-7215-8

THE FRACTAL STRUCTURE OF DATA REFERENCE: *Applications to the Memory Hierarchy,* *Bruce McNutt;* ISBN: 0-7923-7945-4

SEMANTIC MODELS FOR MULTIMEDIA DATABASE SEARCHING AND BROWSING, *Shu-Ching Chen, R.L. Kashyap, and Arif Ghafoor*; ISBN: 0-7923-7888-1

INFORMATION BROKERING ACROSS HETEROGENEOUS DIGITAL DATA: A Metadata-based Approach, *Vipul Kashyap, Amit Sheth*; ISBN: 0-7923-7883-0

DATA DISSEMINATION IN WIRELESS COMPUTING ENVIRONMENTS, *Kian-Lee Tan and Beng Chin Ooi*; ISBN: 0-7923-7866-0

MIDDLEWARE NETWORKS: Concept, Design and Deployment of Internet Infrastructure, *Michah Lerner, George Vanecek, Nino Vidovic, Dad Vrsalovic;* ISBN: 0-7923-7840-7

ADVANCED DATABASE INDEXING, *Yannis Manolopoulos, Yannis Theodoridis, Vassilis J. Tsotras;* ISBN: 0-7923-7716-8

MULTILEVEL SECURE TRANSACTION PROCESSING, *Vijay Atluri, Sushil Jajodia, Binto George* ISBN: 0-7923-7702-8

FUZZY LOGIC IN DATA MODELING, *Guoqing Chen* ISBN: 0-7923-8253-6

INTERCONNECTING HETEROGENEOUS INFORMATION SYSTEMS, *Athman Bouguettaya, Boualem Benatallah, Ahmed Elmagarmid* ISBN: 0-7923-8216-1

For a complete listing of books in this series, go to http://www.springeronline.com

Similarity Search
The Metric Space Approach

Pavel Zezula
Masaryk University, Czech Republic

Giuseppe Amato
ISTI-CNR, Italy

Vlastislav Dohnal
Masaryk University, Czech Republic

Michal Batko
Masaryk University, Czech Republic

Springer

Pavel Zezula
Masaryk University, Czech Republic

Giuseppe Amato
ISTI-CNR, Italy

Vlastislav Dohnal
Masaryk University, Czech Republic

Michal Batko
Masaryk University, Czech Republic

Library of Congress Control Number: 2005933400

ISBN-10: 0-387-29146-6 e-ISBN-10: 0-387-29151-2
ISBN-13: 978-0387-29146-8 e-ISBN-13: 978-0387-29151-2

Printed in the United States of America

9 8 7 6 5 4 3 2 1

SPIN 11552659
e-SPIN 11560609

springeronline.com

This book is dedicated to the
10th anniversary of the
Faculty of Informatics,
Masaryk University in Brno

Contents

Foreword

The area of similarity searching is a very hot topic for both research and commercial applications. Current data processing applications use data with considerably less structure and much less precise queries than traditional database systems. Examples are multimedia data like images or videos that offer query-by-example search, product catalogs that provide users with preference-based search, scientific data records from observations or experimental analyses such as biochemical and medical data, or XML documents that come from heterogeneous data sources on the Web or in intranets and thus does not exhibit a global schema. Such data can neither be ordered in a canonical manner nor meaningfully searched by precise database queries that would return exact matches.

This novel situation is what has given rise to similarity searching, also referred to as content-based or similarity retrieval. The most general approach to similarity search, still allowing construction of index structures, is modeled in metric space. In this book, Prof. Zezula and his co-authors provide the first monograph on this topic, describing its theoretical background as well as the practical search tools of this innovative technology.

In Part I, the authors describe ideas and principles, as well as generic partitioning, search and transformation strategies which have been developed for similarity search in metric spaces. Their use is illustrated in an extensive survey of available indexes. Part II concentrates on similarity search techniques for large collections of data. In particular, it starts with the pioneering work on the M-tree, developed by Prof. Zezula as one of the authors, and continues with the description of hash-based techniques for similarity searching, which formed the main topic of Dr. Dohnal's PhD dissertation. The approximate similarity search, representing another important chapter of this book, was mainly developed in the PhD dissertation of Dr. Amato. The final chapter on scalable and distributed index structures for similarity searching reports the latest efforts of the PhD candidate Dr. Batko. All these PhD dissertations have been supervised by Prof. Zezula.

This monograph is a very valuable resource for scientists who are working or want to work on the many aspects of similarity search. The authors are not only leading experts in this field, but also pedagogically first-rate scholars. Their explanations nicely combine mathematical rigor with intuitive examples and illustration. I believe this book will be a great asset for students and researchers alike.

Prof. Gerhard Weikum
Max-Planck Institute of Computer Science
Saarbruecken, Germany

Preface

In the Information Society, information holds the master key to economic influence and success. But the usefulness of information depends critically upon its quality and the speed at which it can be transferred. In domains as diverse as multimedia, molecular biology, computer-aided design and marketing and purchasing assistance, the number of data resources is growing rapidly, both with regard to database size and the variety of forms in which data comes packaged. To cope with the resulting information overkill, it is vital to find tools to search these resources efficiently and effectively. Hence the intense interest in Computer Science in searching digital data repositories.

But traditional retrieval techniques, typically based upon sorting routines and hash tables, are not appropriate for a growing number of newly-emerging data domains. More flexible methods must be found instead which take into account the needs of particular users and particular application domains.

This book is about finding efficient ways to locate user-relevant information in collections of objects which have been quantified using a pairwise distance measure between object instances. It is written in direct response to recent advances in computing, communication and storage which have led to the current flood of digital libraries, data warehouses and the limitless heterogeneity of Internet resources. The scale of the problem can be gauged by noting that almost everything we see, hear, read, write or measure will soon be available to computerized information systems. In such an environment, varied data modalities such as multimedia objects, scientific observations and measurements, statistical analyses and many others, are massively extending more traditional attribute-like data types.

Ordinary retrieval techniques are inadequate in many of these newer data domains because sorting is simply impossible. To illustrate, consider a collection of bit patterns compared using the Hamming distance, i.e., the number of bits by which a given pair of patterns differs. There is no way to sort the patterns linearly so that, selecting any arbitrary member, the other objects can

be ordered in terms of steadily increasing Hamming distance. The same applies to the spectrum of colors. Obviously, we can sort colors according to their similarity with respect to a specific hue, for example pink. But we can't sort the set of all colors in such a way that, for each hue, its immediate neighbor is the hue most similar to it.

This is what has given rise to a novel indexing paradigm based upon distance. From a formal standpoint, the search problem is modelled in metric space. The collection of objects to be searched forms a subset of the metric space domain, and the distance measure applied to pairs of objects is a metric distance function. This approach significantly extends the scope of traditional search approaches and supports execution of similarity queries. By considering exact, partial, and range queries as special cases, the distance search approach is highly extensible. In the last ten years, its attractiveness has prompted major research efforts, resulting in a number of specific theories, techniques, implementation paradigms and analytic tools aimed at making the distance-based approach viable.

This book focuses on the state of the art in developing index structures for searching metric space. It consists of two parts. Part I presents the metric search approach in a nutshell. It defines the problem, describes major theoretical principles, and provides an extensive survey of specific techniques for a large range of applications. This part is self-contained and does not require any specific prerequisites. Part II concentrates on approaches particularly designed for searching in large collections of data. After describing the most popular centralized disk-based metric indexes, approximation techniques are presented as a way to significantly speed up search time at the expense of some imprecision in query results. The final chapter of the book concentrates on scalable and distributed metric structures, which can deal with data collections that for practical purposes are arbitrarily large, provided sufficient computational power is available in the computer network. In order to properly understand Part II, we recommend at a minimum reading Chapter 1 of Part I.

PAVEL ZEZULA, GIUSEPPE AMATO,

VLASTISLAV DOHNAL, AND MICHAL BATKO

Acknowledgments

We wish to acknowledge all the people who have helped us directly or indirectly in completing this book. First of all, we would like to thank Paolo Ciaccia and Marco Patella for their enthusiastic cooperation and important contribution to the development of metric search techniques. We would also like to mention our other collaborators, mainly Fausto Rabitti, Paolo Tiberio, Claudio Gennaro, and Pasquale Savino, who encouraged us to finish the preparation of this book. We are grateful to Melissa Fearon and Valerie Schofield from Springer for their technical support during the preparation phase. And we are indebted to Mark Alexander for his comments, suggestions, and modifications concerning language and style. Many technical details have been corrected due to observations by Matej Lexa, David Novák, Petr Liška and Fabrizio Falchi who have read a preliminary version of the book. We would also like to acknowledge the support of the EU Network of Excellence DELOS - No. 507618, which made this international publication possible by underwriting travel expenses, and which has served as an excellent forum for discussing the book's subject matter. The work was also partially supported by the National Research Program of the Czech Republic Project number 1ET100300419.