# ADVANCED SIGNATURE INDEXING FOR MULTIMEDIA AND WEB APPLICATIONS

The Kluwer International Series on
# ADVANCES IN DATABASE SYSTEMS

Series Editor
## Ahmed K. Elmagarmid

*Purdue University*
*West Lafayette, IN 47907*

***Other books in the Series:***

**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

**FOUNDATIONS OF KNOWLEDGE SYSTEMS: With Applications to Databases and Agents,** *Gerd Wagner* ISBN: 0-7923-8212-9

**DATABASE RECOVERY,** *Vijay Kumar, Sang H. Son* ISBN: 0-7923-8192-0

**PARALLEL, OBJECT-ORIENTED, AND ACTIVE KNOWLEDGE BASE SYSTEMS,** *Ioannis Vlahavas, Nick Bassiliades* ISBN: 0-7923-8117-3

**DATA MANAGEMENT FOR MOBILE COMPUTING,** *Evaggelia Pitoura, George Samaras* ISBN: 0-7923-8053-3

For a complete listing of books in this series, go to http://www.wkap.nl/prod/s/ADBS

# ADVANCED SIGNATURE INDEXING FOR MULTIMEDIA AND WEB APPLICATIONS

*by*

**Yannis Manolopoulos**
*Aristotle University of Thessaloniki*

**Alexandros Nanopoulos**
*Aristotle University of Thessaloniki*

**Eleni Tousidou**
*Aristotle University of Thessaloniki*

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

**Library of Congress Cataloging-in-Publication Data**

*Advanced Signature Indexing for Multimedia and Web Applications*
Yannis Manolopoulos, Alexandros Nanopoulos, Eleni Tousidou
ISBN 978-1-4613-4654-8 ISBN 978-1-4419-8636-8 (eBook)
DOI 10.1007/978-1-4419-8636-8



*Printed on acid-free paper.*

# Contents

# List of Figures

# List of Tables

# About the Authors

*Yannis Manolopoulos* was born in Thessaloniki, Greece in 1957. He has received a B.Eng. (1981) in Electrical Eng. and a Ph.D. (1986) in Computer Eng., both from the Aristotle University of Thessaloniki. He has been with the Department of Electrical Eng. of the Aristotle University, whereas currently he is Professor at the Department of Informatics of the latter university. He has visited on sabbatical leave the Department of Computer Science of the University of Toronto, the Department of Computer Science of the University of Maryland at College Park and the Department of Computer Science of the University of Cyprus. He has published over 120 papers in refereed scientific journals and conference proceedings. He is author of two textbooks on data/file structures, which are/were recommended in the vast majority of the computer science/engineering departments in Greece. He has co-authored a monograph on "Advanced Database Indexing" published by Kluwer (1999). He is/was PC chair/co-chair of the 8th Panhellenic Conference in Informatics (2001), the 6th East-European Conference on Advanced Databases and Information Systems (2002), the 6th Symposium on Spatiotemporal Databases (2003), the 5th Workshop on Distributed Data and Structures (2003) and the 1st Balkan Conference in Informatics (2003). Currently he is Vice-chairman of the Executive Board of the Greek Computer Society (EPY) and member of the Editorial Board of The Computer Journal. His research interests include spatiotemporal databases, Web databases, data mining, data/file structures and algorithms, and performance evaluation of storage subsystems.

*Alexandros Nanopoulos* was born in Craiova, Romania in 1974. He has received his B.Sc. and Ph.D. degrees from the Department of Informatics of Aristotle University of Thessaloniki, in 1996 and 2003, respectively. He has served as external reviewer in several conferences, like SIGMOD, VLDB, ICDE, and EDBT. His research interests include data mining, Web databases, and spatial access methods.

*Eleni Tousidou* was born in Thessaloniki, Greece in 1974. She has received her B.Sc. and Ph.D. degrees from the Department of Informatics of the Aristotle University of Thessaloniki, in 1996 and 2002, respectively. Currently, she is teaching at the Department of Computer Engineering, Telecommunications & Networks of the University of Thessaly. She has been a visitor at the University of Alberta at Edmonton during summer 2001. She has served as external reviewer in several conferences, like VLDB and ICDE. Her research interests include query processing and access methods in object-oriented databases and spatial databases, and complex object handling in multimedia databases.

# Preface

A felicitous wording by Don Knuth (The Art of Computer Programming Vol. 3) stated: "when faced with a mountain of data, people are tempted to use a computer to find the answer to most difficult queries they can dream up". He continues, however, by saying: "The desired calculations are possible, but they're not right for everyone's application." Indeed, novel and emerging application fields incur new requirements for data types and query processing techniques. It is, therefore, a necessity to advance traditional and established methods to meet the new challenges.

In the well honored field of querying large collections of textual information, the two main classes of data structures comprise *inverted* and *signature* indices. Nevertheless, databases have evolved (and will continue to do so). Besides text, data types for multimedia are now supported, such as images, video, and audio. Therefore, new query processing techniques have been devised to facilitate storage and searching of this new "mountain." Moreover, another milestone in the development of today's database systems (and not only) is the Web, which has become the standard means of information dissemination, with applications like e-commerce or e-learning. For such novel applications, recent work has shown that in several cases traditional indices cannot address all the new requirements.

The key concept in this monograph is the development of new indexing and query processing methods for the aforementioned applications. The focus is on the notion of signature. However, what is called signature does not have exactly the same meaning as it did in text databases. The challenges posed by the new applications demand a different designation of hashing information in bit-strings, a.k.a. signatures. For instance, queries like similarity searching, e.g., for similar images or similar

Web-user behavior, are very different from plain searching of text-term containment.

This book may serve as a textbook for graduate students specializing in database and information retrieval systems, or for database professionals that are involved in the development of applications in multimedia databases or the Web. Emphasis has been given on structure description, implementation techniques and clear evaluation of operations performed (from a performance perspective). Technical detail, although increased, does not prevent readers to get accustomed to the discussed issues through explanatory examples. Furthermore, it can serve as a reference on each specific subject, since it also surveys existing techniques.

The book is divided in three parts. The first part consists of four chapters and illustrates the fundamental data structures, which will serve as a basis for the applications given in the next two parts, each consisting of two chapters. Each chapter ends with references for further reading. The final section gives directions for work in other research fields.

In the first part, Chapter 1 defines the fundamental notions for superimposed signatures. Chapter 2 describes signature indexing with tree structures. Emphasis is given on the tree building operations. The next chapter contains indexing methods based on hash structures, along with organization schemes at the data level, which are based on the previous tree methods. Chapter 4 illustrates compression techniques, that manage to reduce I/O overhead and to improve query performance.

The first chapter of the second part explains the representation of image data with thematic layers by using signature based indexing schemes for organizing color with spatial information in image data. Chapter 6 is about the processing of similarity queries over image data.

In the third part, Chapter 7 considers the querying of click-stream data (representing Web-user sessions) and the problem of incorporating similarity measures within signature representations. Chapter 8 describes the storage and querying of large Web-logs, dealing with the important factor of preserving the ordering of Web-user accesses.

Finally, we have included an Addendum that briefly reviews related work on signature based schemes in other paradigms.

YANNIS MANOLOPOULOS

ALEXANDROS NANOPOULOS

ELENI TOUSIDOU

To our parents.
In memorial of
Panagiotis
Manolopoulos.

# Acknowledgments

The content of this monograph is based on research performed during the last years at the Data Engineering Lab of the Department of Informatics of the Aristotle University. This work has been financially supported by national and international funds. However, it would not have been possible to reach these research results without the co-operation of several colleagues. Thus, thanks are due to: Panayiotis Bozanis, Vishal Chitkara, Maria Kontaki, Tadeusz Morzy, Enrico Nardelli, Guido Proietti, and Maciej Zakrzewicz. In particular, we especially thank Mario Nascimento for his contribution in Chapter 6.