
**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 Ghaffoor; 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

Copyright © 2003 by Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers in 2003

Softcover reprint of the hardcover 1st edition 2003

All rights reserved. No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without the written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Permission for books published in Europe: permissions@wkap.nl

Permissions for books published in the United States of America: permissions@wkap.com

Printed on acid-free paper.

Contents

List of Figures	ix
List of Tables	xv
Preface	xix
Acknowledgments	xxiii

Part I Advanced Structures for Signatures

1. INTRODUCTION	3
1 Superimposed Signatures	4
2 False-drops	6
3 Signature Construction Methods	6
4 Objective and Organization	8
2. SIGNATURE INDEXING WITH TREE STRUCTURES	13
1 Introduction	13
2 S-trees	14
3 An Analogy to Other Indices Structures	16
4 Split Methods: Complexity vs. Effectiveness	18
5 Analytical Results	23
6 Experimental Evaluation	28
7 Conclusions and Further Reading	32
3. HYBRID STRUCTURES	39
1 Introduction	39
2 Combining Linear Hashing and Signatures	40
3 New Hybrid Structures	44

4	Analytical Study of Hybrid Structures	53
5	Performance Study	57
6	Conclusions and Further Reading	67
4.	COMPRESSION TECHNIQUES	71
1	Introduction	71
2	Compression Scheme for the S-tree	72
3	Querying and the Decompression Scheme	74
4	Performance Study	75
5	Conclusions and Further Reading	78
Part II Signature Indexing for Multimedia Applications		
5.	REPRESENTING THEMATIC LAYERS	83
1	Introduction	83
2	Linear Structures for Thematic Layers	85
3	Introducing the Color Binary String	95
4	Algorithms for Window Querying	97
5	Performance Results	101
6	Conclusions and Further Reading	107
6.	IMAGE INDEXING AND RETRIEVAL WITH SIGNATURE SCHEMES	113
1	Introduction	113
2	Representing Images by Color: the VBA Scheme	118
3	Using the Signature tree in Image Retrieval	122
4	Querying for Nearest-Neighbor Images	123
5	Measuring Effectiveness and Efficiency	126
6	Conclusions and Further Reading	131
Part III Signature Indexing for Web Applications		
7.	RETRIEVING SIMILAR WEB-USER BEHAVIORS	139
1	Introduction	139
2	Background	140
3	Representation of Web-user Transactions	143
4	Processing Similarity Queries	145

<i>Contents</i>	vii
5 Performance Evaluation	151
6 Applications to Recommendation Systems	161
7 Conclusions and Further Reading	163
8. STORAGE AND QUERYING OF LARGE WEB-LOGS	169
1 Introduction	169
2 Related Work	171
3 Equivalent Sets: Considering the Order within Web-Log Access Sequences	172
4 Signature-indexing Schemes for Equivalent Sets	174
5 Comparison of Examined Indices	179
6 Conclusions and Further Reading	181
ADDENDUM: SIGNATURES IN MOBILE COMPUTING AND DATA WAREHOUSING	185
INDEX	193

List of Figures

2.1	An example of an S-tree with $K=4$ and $k=2$.	14
2.2	Comparison of estimation functions of the (a) original and (b) quadratic split methods, as a function of the query weight.	27
2.3	Performance of the two linear algorithms for (a) 100 K entries, (b) 150 K entries, as a function of the query weight.	29
2.4	Performance of the two variations of hierarchical clustering (minimum and mean distance) as a function of the query weight.	30
2.5	Performance of the proposed methods for 512-120 signatures as a function of the query weight.	30
2.6	Performance of the proposed methods for 1024-256 signatures as a function of the query weight.	31
2.7	Performance of the proposed methods for 512-120 signatures as a function of the number of signatures.	31
2.8	Performance of the proposed methods for 1024-256 signatures as a function of the number of signatures.	32
2.9	Performance of the proposed methods for larger page sizes for (a) 2K-512 signatures, (b) 4K-1024 signatures.	32
3.1	An example of PWF with $F=16$ and $h=2$.	41
3.2	An example of LHS structure with $F=16$ and $h=2$.	44
3.3	An example of LHS with $F=12$ and $h=1$.	46
3.4	The LHS structure of Figure 3.3, after the hash table has been expanded.	47

3.5	An example of LOC structure with $F=16$ and $h=2$.	49
3.6	An example of LOG structure with $F=16$ and $h=2$.	51
3.7	Comparison of analytical estimates for (a) the LHS, (b) the LOC and (c) the LOG method, as a function of the query weight.	58
3.8	Comparison of the proposed methods as a function of the weight of inserted signatures.	60
3.9	Performance of the LOG method: (a) Retrieval costs. (b) Storage overhead.	61
3.10	Comparison of the proposed methods in 2K pages for (a) 512-120 signatures, (b) 512-154 signatures, as a function of the query weight.	62
3.11	Comparison of the proposed methods in 4K pages for (a) 1024-256 signatures, (b) 1024-340 signatures, as a function of the query weight.	62
3.12	Time overhead of the proposed methods in 2K and 4K pages for (a) 512-120 signatures, (b) 1024-256 signatures, respectively, as a function of the query weight.	63
3.13	Comparison of the proposed methods for (a) 512-120 signatures, (b) 1024-256 signatures, as a function of the number of inserted signatures.	64
3.14	Comparison of the proposed methods for (a) 512-154 signatures, (b) 1024-340 signatures, as a function of the number of inserted signatures.	65
3.15	Comparison of the proposed methods over a super-set query for 512 signatures.	66
3.16	Storage overhead of the five methods for 512 bits signatures and 2K page size.	66
3.17	Percentage of created overflow pages of the five methods for 512 bits signatures and 2K page size.	67
4.1	An example of step 3 in the S-tree node compression scheme.	73
4.2	An example of node decompression.	75
4.3	Query performance w.r.t. query size: (a) Disk accesses, (b) CPU time for decompression.	77
4.4	Query performance w.r.t. query size: (a) Disk accesses, (b) CPU time for decompression.	77

4.5	Query performance w.r.t. query size: (a) Disk accesses, (b) CPU time for decompression.	78
5.1	(a) An 8×8 image. (b) The feature-Id table.	86
5.2	The quadtree representing the image of Figure 5.1.	86
5.3	(a) A binary image. (b) Its representing binary tree.	89
5.4	Layout of a data page of the S^+ -tree.	90
5.5	The binary tree produced by the image of Figure 5.1.	90
5.6	Left: Page A, Right: Page B.	91
5.7	Left: Page C, Right: Page D.	92
5.8	Left: Page E, Right: Page F.	92
5.9	The prefix B-tree containing the separators of the bintree of Figure 5.4.	93
5.10	Layout of a data page of the S^* -tree.	94
5.11	The B^+ -tree with the separators created by the S^* -tree.	95
5.12	BSL tree representing the image of Figure 5.1.	98
5.13	BHL tree representing the image of Figure 5.1.	98
5.14	BS^* tree representing the image of Figure 5.1.	99
5.15	Space Overhead involved in five different 1024×1024 images containing 64 features.	103
5.16	Exist query where 2 features were queried, image size 1024×1024 , 64 features: (a) Averaged results, (b) Normalized results.	104
5.17	Exist query where 5 features were queried, image size 1024×1024 , 64 features: (a) Averaged results, (b) Normalized results.	104
5.18	Exist query where 10 features were queried, image size 1024×1024 , 64 features: (a) Averaged results, (b) Normalized results.	105
5.19	Exist query for a varying number of queried features, image size 1024×1024 , 64 features, query window 100×100 : (a) Averaged results, (b) Normalized results.	105
5.20	Report query on images of 1024×1024 size containing 64 features: (a) Averaged results, (b) Normalized results.	106
5.21	Select query where 2 features were queried, image size 1024×1024 , 64 features: (a) Averaged results, (b) Normalized results.	107

5.22	Select query where 5 features were queried, image size 1024×1024 , 64 features: (a) Averaged results, (b) Normalized results.	107
5.23	Select query where 10 features were queried, image size 1024×1024 , 64 features: (a) Averaged results, (b) Normalized results.	108
5.24	Select query for a varying number of queried features, image size 1024×1024 , window size 100×100 : (a) averaged results, (b) normalized results.	108
6.1	Sample image set.	120
6.2	Retrieval Effectiveness, in terms of Precision vs. Recall.	128
6.3	Index (or file) sizes.	129
6.4	Query time (# I/Os).	130
7.1	The algorithm for the nearest-neighbor query.	146
7.2	The algorithm for the range query.	147
7.3	The proposed split algorithm.	149
7.4	Example of paging for $Nfr = 4$.	151
7.5	Performance for similarity queries against the query size. Left: k -nearest neighbor query w.r.t. k . Right: range query w.r.t. r (radius) threshold.	153
7.6	Scalability results w.r.t. the database size.	154
7.7	Left: Nearest neighbor query w.r.t. mean size, T , of transactions. Right: Nearest neighbor query w.r.t. total number of items in database.	155
7.8	Sensitivity against the item distribution.	157
7.9	Sensitivity against the buffer size.	158
7.10	Comparison for dynamic data.	159
7.11	Performance results for the real dataset and the k -nearest-neighbor w.r.t. k .	160
7.12	Results on the proposed enhancements.	161
7.13	A recommendation example. Up: The product database. Bottom: The transactions database.	162
8.1	An example of a web access-log and an access sequence.	170
8.2	(a) The relation R of web access sequences. (b) Mapping between URL_ID and URL . (c) An example of a pattern query.	171
8.3	SI method: Search algorithm.	176

8.4	Algorithm for obtaining signatures of approximations of signature sets.	178
8.5	(a) I/O vs. query size for ClarkNet web-log. (b) I/O vs. query size for Synthetic web-log.	180
8.6	(a) Scalability. (b) Tuning of k .	181

List of Tables

2.1	Symbol Table.	24
2.2	Parameters used in experiments and the values tested.	28
3.1	Symbol Table.	54
3.2	Parameters used in experiments and the values tested.	57
6.1	Detailed Signatures of the images in Figure 6.1 using VBA.	121
6.2	Calculation of function (image distance) d .	124
6.3	Calculation of $minDist$ function.	125

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.