AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Data Management for Multimedia Retrieval

This is a pre print version of the following article:

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/73789 since 2021-04-29T21:11:05Z

Publisher:

Cambridge University Press

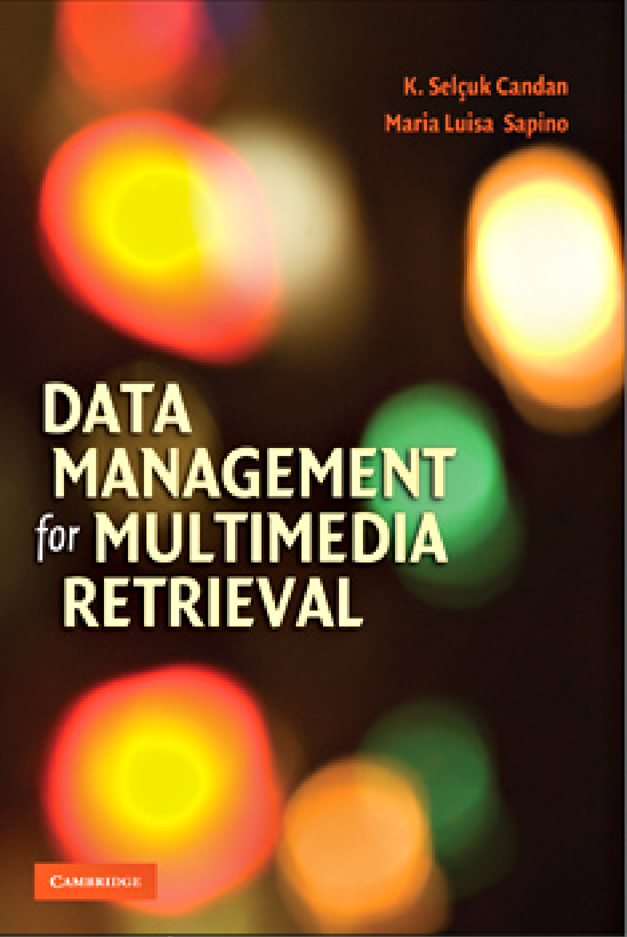
Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

14 July 2024



K. Selçuk Candan

Maria Luisa Sapino

DATA MANAGEMENT for MULTIMEDIA RETRIEVAL

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521887397

This page intentionally left blank

Data Management for Multimedia Retrieval

Multimedia data require specialized management techniques because the representations of color, time, semantic concepts, and other underlying information can be drastically different from one another. The user's subjective judgment can also have significant impact on what data or features are relevant in a given context. These factors affect both the performance of the retrieval algorithms and their effectiveness. This textbook on multimedia data management techniques offers a unified perspective on retrieval efficiency and effectiveness. It provides a comprehensive treatment, from basic to advanced concepts, that will be useful to readers of different levels, from advanced undergraduate and graduate students to researchers and professionals.

After introducing models for multimedia data (images, video, audio, text, and web) and for their features, such as color, texture, shape, and time, the book presents data structures and algorithms that help store, index, cluster, classify, and access common data representations. The authors also introduce techniques, such as relevance feedback and collaborative filtering, for bridging the "semantic gap" and present the applications of these to emerging topics, including web and social networking.

K. Selçuk Candan is a Professor of Computer Science and Engineering at Arizona State University. He received his Ph.D. in 1997 from the University of Maryland at College Park. Candan has authored more than 140 conference and journal articles, 9 patents, and many book chapters and, among his other scientific positions, has served as program chair for ACM Multimedia Conference'08, the International Conference on Image and Video Retrieval (CIVR'10), and as an organizing committee member for ACM SIG Management of Data Conference (SIGMOD'06). In 2011, he will serve as a general chair for the ACM Multimedia Conference. Since 2005, he has also been serving as an associate editor for the *International Journal on Very Large Data Bases* (*VLDB*).

Maria Luisa Sapino is a Professor in the Department of Computer Science at the University of Torino, where she also earned her Ph.D. There she leads the multimedia and heterogeneous data management group. Her scientific contributions include more than 60 conference and journal papers; her services as chair, organizer, and program committee member in major conferences and workshops on multimedia; and her collaborations with industrial research labs, including the RAI-Crit (Center for Research and Technological Innovation) and Telecom Italia Lab, on multimedia technologies.

DATA MANAGEMENT FOR
MULTIMEDIA RETRIEVAL

K. Selçuk Candan
Arizona State University

Maria Luisa Sapino
University of Torino

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521887397

© K. Selcuk Candan and Maria Luisa Sapino 2010

This publication is in copyright. Subject to statutory exception and to the provision of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published in print format 2010

ISBN-13 978-0-511-90188-1 eBook (NetLibrary)

ISBN-13 978-0-521-88739-7 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of urls for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

Preface *page* ix

1 Introduction: Multimedia Applications and Data Management Requirements 1

1.1 Heterogeneity 1
1.2 Imprecision and Subjectivity 8
1.3 Components of a Multimedia Database Management System 12
1.4 Summary 19

2 Models for Multimedia Data 20

2.1 Overview of Traditional Data Models 21
2.2 Multimedia Data Modeling 32
2.3 Models of Media Features 34
2.4 Multimedia Query Languages 92
2.5 Summary 98

3 Common Representations of Multimedia Features 99

3.1 Vector Space Models 99
3.2 Strings and Sequences 109
3.3 Graphs and Trees 111
3.4 Fuzzy Models 115
3.5 Probabilistic Models 123
3.6 Summary 142

4 Feature Quality and Independence: Why and How? 143

4.1 Dimensionality Curse 144
4.2 Feature Selection 145
4.3 Mapping from Distances to a Multidimensional Space 167
4.4 Embedding Data from One Space into Another 172
4.5 Summary 180

5 Indexing, Search, and Retrieval of Sequences 181
5.1 Inverted Files 181
5.2 Signature Files 184
5.3 Signature- and Inverted-File Hybrids 190
5.4 Sequence Matching 191
5.5 Approximate Sequence Matching 195
5.6 Wildcard Symbols and Regular Expressions 202
5.7 Multiple Sequence Matching and Filtering 204
5.8 Summary 206

6 Indexing, Search, and Retrieval of Graphs and Trees 208
6.1 Graph Matching 208
6.2 Tree Matching 212
6.3 Link/Structure Analysis 222
6.4 Summary 233

7 Indexing, Search, and Retrieval of Vectors 235
7.1 Space-Filling Curves 238
7.2 Multidimensional Index Structures 244
7.3 Summary 270

8 Clustering Techniques 271
8.1 Quality of a Clustering Scheme 272
8.2 Graph-Based Clustering 275
8.3 Iterative Methods 280
8.4 Multiconstraint Partitioning 286
8.5 Mixture Model Based Clustering 287
8.6 Online Clustering with Dynamic Evidence 288
8.7 Self-Organizing Maps 290
8.8 Co-clustering 292
8.9 Summary 296

9 Classification 297
9.1 Decision Tree Classification 297
9.2 *k*-Nearest Neighbor Classifiers 301
9.3 Support Vector Machines 301
9.4 Rule-Based Classification 308
9.5 Fuzzy Rule-Based Classification 311
9.6 Bayesian Classifiers 314
9.7 Hidden Markov Models 316
9.8 Model Selection: Overfitting Revisited 322
9.9 Boosting 324
9.10 Summary 326

10 Ranked Retrieval 327
10.1 *k*-Nearest Objects Search 328
10.2 Top-*k* Queries 337

10.3 Skylines 360
10.4 Optimization of Ranking Queries 373
10.5 Summary 379

11 Evaluation of Retrieval 380

11.1 Precision and Recall 381
11.2 Single-Valued Summaries of Precision and Recall 381
11.3 Systems with Ranked Results 383
11.4 Single-Valued Summaries of Precision-Recall Curve 384
11.5 Evaluating Systems Using Ranked and Graded Ground Truths 386
11.6 Novelty and Coverage 390
11.7 Statistical Significance of Assessments 390
11.8 Summary 397

12 User Relevance Feedback and Collaborative Filtering 398

12.1 Challenges in Interpreting the User Feedback 400
12.2 Alternative Ways of Using the Collected Feedback in Query Processing 401
12.3 Query Rewriting in Vector Space Models 404
12.4 Relevance Feedback in Probabilistic Models 404
12.5 Relevance Feedback in Probabilistic Language Modeling 408
12.6 Pseudorelevance Feedback 411
12.7 Feedback Decay 411
12.8 Collaborative Filtering 413
12.9 Summary 425

Bibliography 427

Index 473

Color plates follow page 38