Lecture Notes in Computer Science

Edited by G. Goos and J. Hartmanis

471

Beng Chin Ooi

Efficient Query Processing in Geographic Information Systems

Springer-Verlag
Berlin Heidelberg New York London
Paris Tokyo Hong Kong Barcelona

Editorial Board

D. Barstow W. Brauer P. Brinch Hansen D. Gries D. Luckham
C. Moler A. Pnueli G. Seegmüller J. Stoer N. Wirth

Author

Beng Chin Ooi
Computer Science Department, Monash University
Clayton, Victoria, Australia 3168

Current address:
Institute of Systems Science, National University of Singapore
Heng Mui Keng Terrace, Kent Ridge, Singapore 0511

CR Subject Classification (1987): E.1, H.2.3, H.3.1–3

ISBN 3-540-53474-1 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-53474-1 Springer-Verlag New York Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways, and storage in data banks. Duplication of this publication or parts thereof is only permitted under the provisions of the German Copyright Law of September 9, 1965, in its current version, and a copyright fee must always be paid. Violations fall under the prosecution act of the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 1990
Printed in Germany

Printing and binding: Druckhaus Beltz, Hemsbach/Bergstr.
2145/3140-543210 – Printed on acid-free paper

Preface

Geographic Information Systems (GISs) are database systems that allow the manipulation, storage, retrieval, and analysis of geographic data and the display of data in the form of maps. In such a system, the database describes a collection of geographic entities over a two-dimensional map. The information describes entities that have a physical location and extent in some spatial region of interest, and queries involve identification of these entities based on their aspatial and spatial attributes, and the spatial relationships between entities. Conventional query languages must be augmented to contain spatial operators [ChF81] and additional spatial data structures introduced to support efficient geometric operations. In addition, new optimization strategies are required to process the resultant queries efficiently.

One of the important characteristics of relational DBMSs is the optimizer which automatically translates a query expressed in a nonprocedural language into an optimal sequence of access operations to evaluate the query. In such a system, the user need not know the physical structure of the database. Instead the optimizer estimates the cost of possible alternatives for processing the transaction based on the given physical structure of the database and computes the minimum cost sequence. However, existing DBMSs are not designed to support the hybrid queries that involve the selection of data based on spatial relationships.

Efficient indexing structures provide the query optimizer with the means to construct an efficient query execution plan. A new indexing structure called the *spatial kd-tree* (skd-tree) is proposed in this thesis. The new structure supports two types of proximity search, namely the *containment* search and the *intersection* search in order to facilitate the evaluation of queries involving spatial (geometric) operators.

The architecture of the GIS proposed in this thesis utilizes a conventional relational DBMS. The conventional DBMS provides efficient storage and retrieval of aspatial data. A spatial processor is included to process the spatial components of the queries. A global optimizer is proposed to sequence the evaluation of spatial and aspatial components of queries. This involves the decomposition of queries into aspatial and spatial subqueries, the ordering of these subqueries, and the merging of partial results. Among the subquery sequences, the best can be selected. The effectiveness of the method is demonstrated.

This book is the revised and extended version of a Ph.D. dissertation [Ooi88] submitted to the Computer Science Department, Monash University, Victoria, Australia. I am indebted to a number of people who have assisted me in one way or another to

materialize this thesis. In particular, I would like to express my appreciation to my two supervisors, Dr. Ken J. McDonell of Pyramid Australia and Dr. Ron Sacks-Davis of Royal Melbourne Institute of Technology, for their patience, invaluable supervision and guidance throughout the development of this project. I would also like to thank Ron and Ken for reading and checking the correctness of the thesis, and for improving its presentation.

My sincere thanks are due to my departmental supervisor, Dr. Binh Pham, for reading and providing comments on the initial draft of the thesis, Dr. Bala Srinivasan, for useful discussions, Mr. Anthony Maeder for his advice, Dr. David J. Abel of CSIRO in Canberra, Australia for discussions and help, the programmers, Mr. David Hook and Mr. Vincent Trinh, who were supported through a collaborative project funded by CSIRO, for their feedback on the skd-tree algorithms, Dr. Hans Woessner and Mr. J. Andrew Ross of Springer-Verlag for their editorial help, and anonymous referees for their useful suggestions.

To Clement Chan, Nick Ch'ng, Guat Khunn Goh, Mark Goodwin, A.S.M. Sajeev and Thompson Thong, I offer thanks for discussions. I would also like to thank Guat Khunn Goh, Shirley Ho and A.S.M. Sajeev, for reading the initial drafts of the thesis.

Thanks to Monash University for financial support in terms of a Monash Graduate Scholarship, CSIRO, Canberra Australia for funding the project, and Institute of Systems Science, National University of Singapore for supporting me in preparing this manuscript.

To my wife, my mother, brothers and sisters, Po Sim and Kian Fah in particular, and other family members, I give thanks for their love, support and encouragement. This book is dedicated to my late father, Mr. Ooi Arm, whose love will always be remembered.

Singapore, August 1990 Beng Chin Ooi

Table of Contents

1 Introduction 1

1.1 Geographic Information Systems 1

1.2 Motivation 2

1.3 Indexing Structures 4

1.4 Query Optimization 5

1.5 Thesis Synopsis 6

2 Related Work and GEOQL 9

2.1 Geographic Information Systems 9

2.1.1 The Nature of Data 9

2.1.2 On the Design of a GIS 10

2.1.3 Existing Systems 11

2.2 Query Languages 14

2.2.1 QUEL Extensions 14

2.2.2 QBE Extensions 16

2.2.3 SQL Extensions 17

2.2.4 GEOQL — An Augmented SQL 18

2.3 The Need for Efficient Spatial Indexing Mechanisms 22

2.4 Point Structures 23

2.4.1 Introduction 23

2.4.2 Quad-Trees 23

2.4.2.1 Point Quad-Trees 23

2.4.2.2 Region Quad-Trees 25

2.4.3 The kd-Trees 28

2.5 Access Structures for Extended Spatial Objects 32

2.5.1 Bounding Rectangles 32

2.5.2 Search Types 33

2.5.3 Spatial Indexing Methods 33

2.5.4 Corner Stitching 34

2.5.5 Cell Methods based on Dynamic Hashing 35

2.5.5.1 The Grid File 35
2.5.5.2 The EXCELL Method 41
2.5.5.3 PLOP-Hashing 42
2.5.6 Quad-Tree Based Structures 42
2.5.7 Locational Keys 44
2.5.8 Matsuyama's kd-Tree 47
2.5.9 The 4-D-Tree 49
2.5.10 The R-Tree 49
2.5.11 The R^+-Tree 52
2.5.12 The Cell Tree 57
2.5.13 Summary 57
2.6 Optimizations 60
2.6.1 Introduction to Query Optimization 60
2.6.2 Definitions 62
2.6.3 Optimization Strategies 63
2.6.3.1 Query Transformation 64
2.6.3.2 Query Evaluations 71
2.6.3.3 Optimization in System R 73
2.6.3.4 Optimization in INGRES — Query Decomposition 74
2.6.4 Global Optimization Strategies 74
2.6.5 Extensible DBMS and Resident Memory Systems — Current Trends 75
2.6.6 Optimization in GIS 77
2.6.7 Summary 78
2.7 Conclusion 78

3 The Spatial kd-Tree 80
3.1 A New-Structure — The Spatial kd-Tree 80
3.2 Searching 86
3.3 Insertion 88
3.4 Deletion 90
3.5 Directory Paging 96
3.5.1 Page Structure 97
3.5.2 Paging Strategy 98

3.6 Static Tree Construction 100
3.7 Data Storage Structure 105
3.8 Spatial Primitive Operations 107
3.9 Set Operations 111
3.10 Summary 115

4 Performance Analysis and Case Studies 116
4.1 Empirical Analysis 116
4.1.1 Quantities of Interest 116
4.1.2 Parameters 117
4.1.3 Simulation Methodology 117
4.2 A Variant of the skd-Tree 118
4.3 Evaluation with Non-uniform Data 121
4.3.1 The Skew Factor 122
4.3.2 Data Sets 122
4.3.3 Empirical Results 123
4.4 Further Comparisons 129
4.4.1 Object Mapping Structures 130
4.4.2 Object Duplication Structures 132
4.5 Summary 135

5 Query Optimization 136
5.1 Motivation 136
5.2 Extending Techniques 137
5.2.1 Overview of An Extended System Architecture 137
5.2.2 Extended Optimization 138
5.3 Classification of GEOQL Queries 139
5.4 An Introduction to the Extended Optimization 141
5.4.1 Grouping of Predicates 142
5.4.2 Answer Construction 142
5.4.2.1 The Not Operator 143
5.4.2.2 The And Operator 143
5.4.2.3 The Or Operator 143
5.5 Optimization Strategy 144
5.5.1 Parsing 144

5.5.2 Logical Transformation 145
5.5.3 Decomposition 148
5.5.4 Subquery Sequencing 153
5.5.4.1 Results Formation 153
5.5.4.2 SQP Formulation 159
5.5.4.3 Cost Estimation 163
5.6 Conclusions 164
6 Implementation and Experiments 166
6.1 Justification 166
6.2 RTB and ISQL 166
6.3 The GEOQL Database 168
6.4 The Experimental System 169
6.5 Discussion 176
6.5.1 Extensibility 176
6.5.2 Limitations 177
6.6 Conclusion 177

7 Conclusions 178

References 181
Appendix A 196
Appendix B 202
Appendix C 207