

Springer

Berlin

Heidelberg

New York

Barcelona

Hong Kong

London

Milan

Paris

Tokyo

Marcus Jürgens

Index Structures for Data Warehouses



Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany
Juris Hartmanis, Cornell University, NY, USA
Jan van Leeuwen, Utrecht University, The Netherlands

Author

Marcus Jürgens
Freie Universität Berlin
Fachbereich für Mathematik und Informatik, Institut für Informatik
Takustraße 9, 14195 Berlin, Germany
E-mail: marcus.juergens@gmx.de

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Jürgens, Marcus:
Index structures for data warehouses / Marcus Jürgens. - Berlin ; Heidelberg ;
New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ;
Tokyo : Springer, 2002
(Lecture notes in computer science ; 1859)
Zugl.: Berlin, Freie Univ., Diss., 2000
ISBN 3-540-43368-6

CR Subject Classification (1998): H.3.1, H.2.7, H.3, H.2

ISSN 0302-9743

ISBN 3-540-43368-6 Springer-Verlag Berlin Heidelberg New York

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 any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2002
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10846408 06/3142 5 4 3 2 1 0

Preface

This thesis investigates which index structures support query processing in typical data warehouse environments most efficiently. Data warehouse applications differ significantly from traditional transaction-oriented operational applications. Therefore, the techniques applied in transaction-oriented systems cannot be used in the context of data warehouses and new techniques must be developed.

The thesis shows that the time complexity for the computation of *optimal* tree-based index structures prohibits its use in real world applications. Therefore, we *improve* heuristic techniques (*e. g.* R^* -tree) to process range queries on aggregated data more efficiently. Experiments show the benefits of this approach for different kinds of typical data warehouse queries. Performance models *estimate* the behavior of standard index structures and the behavior of the extended index structures. We introduce a new model that considers the distribution of data. We show experimentally that the new model is more precise than other models known from literature. Two techniques *compare* two tree-based index structures with two bitmap indexing techniques. The performance of these index structures depends on a set of different parameters. Our results show which index structure performs most efficiently depending on the parameters.

Acknowledgements

I am very grateful to have had the opportunity to write my Ph.D. Thesis under the supervision of Professor Hans-Joachim Lenz who brought the area of data warehouses to my attention. In countless meetings he gave me helpful feedback. I would like to thank Professor Heinz Schweppe for his constructive suggestions and the invitation to cooperate with the database group at the Freie Universität Berlin. Professor Freytag supported me with beneficial ideas and outstanding comments.

The graduate school in Distributed Information Systems would not be possible in this efficient form without its speaker Professor Oliver Günther. His commitment gives this school a constructive and pleasant environment.

I would like to express my thanks to all members of database groups participating in this graduate school for their interesting and encouraging talks and discussions. In particular, I am grateful for the constructive discussions with Agnès Voisard and Annika Hinze. The *Deutsche Forschungsgemeinschaft (DFG)* supported me as a fellowship recipient. Professor Joseph Bronstad and Leslie Hazelwood did not give up trying to correct my English.

Contents

1	Introduction	1
1.1	Goals	2
1.2	Outline	3
2	State of the Art of Data Warehouse Research	5
2.1	Introduction	5
2.2	Traditional Transaction-Oriented Systems	5
2.3	Data Warehouses for Decision Support	7
2.4	OLAP Vs. OLTP	9
2.5	Accelerating Query Speed	10
2.5.1	Denormalized Schemas	10
2.5.2	Materialized Views	11
2.5.3	No Locking	13
2.5.4	On-line Aggregation	13
2.5.5	Index Structures	14
2.6	Summary	14
3	Data Storage and Index Structures	15
3.1	Introduction	15
3.2	Memory Hierarchy	15
3.3	Mechanics of Disks	16
3.4	Data Space and Queries	18
3.4.1	Data Space	18
3.4.2	Queries	18
3.5	Tree-Based Indexing	19
3.5.1	Top-Down, Bottom-Up, and Bulk Loading	20
3.5.2	Point Quadtrees	21
3.5.3	kd -tree	22
3.5.4	kdb -tree	22
3.5.5	R -tree	23
3.5.6	R* -tree	23
3.5.7	Other Relatives of the R -tree Family and Other Tree Structures	24
3.5.8	Generic Tree Structures	27

3.6	Bitmap Indexing	27
3.6.1	Standard Bitmap Indexing	27
3.6.2	Multi-component Equality Encoded Bitmap Index ...	29
3.6.3	Range-Based Encoding	31
3.6.4	Multi-component Range-Based Encoding	32
3.6.5	Other Compression Techniques / Combination of Bitmaps and Trees	33
3.7	Arrays	34
3.8	Summary	34
4	Mixed Integer Problems for Finding Optimal Tree-Based Index Structures	35
4.1	Introduction	35
4.2	Optimization Problem Parameters	35
4.3	Mapping into a Mixed Integer Problem	36
4.4	Problem Complexity	38
4.5	Model Evaluation	39
4.6	Summary	41
5	Aggregated Data in Tree-Based Index Structures	43
5.1	Introduction	43
5.2	□Fit for Aggregation□ Access Method	47
5.3	Materialization of Data	48
5.4	Modified Operations	50
5.4.1	Insert Operation	51
5.4.2	Delete Operation	51
5.4.3	Update Operation	51
5.4.4	Creating Index Structures, Bottom-Up Index Structures	51
5.4.5	Point Query Algorithm	52
5.4.6	Range Query Algorithm	52
5.5	Storage Cost	52
5.6	Height of Tree	55
5.7	Overlaps of Regions	55
5.8	Experiments	56
5.8.1	Cost Model	57
5.8.2	Physical Index Structure	57
5.8.3	Implementation	58
5.8.4	Generation of Test Data	58
5.8.5	Query Profile	59
5.8.6	Results of Experiments	60
5.9	Summary	62

6	Performance Models for Tree-Based Index Structures	63
6.1	Introduction	63
6.2	Fit for Modeling	63
6.3	Performance Models for Access Leaf Nodes	64
6.3.1	GRID Model	64
6.3.2	SUM Model	66
6.3.3	Equivalence of GRID Model and SUM Model	68
6.3.4	FRACTAL Model	69
6.3.5	Equivalence between FRACTAL Model, SUM Model, and GRID Model	71
6.4	PISA Model	72
6.5	Computational Efficiency of SUM Model and PISA Model . . .	75
6.6	Adapting PISA Model to Different Distributions	77
6.6.1	Uniformly Distributed Data	77
6.6.2	Skewed Data	78
6.6.3	Normally Distributed Data	80
6.7	Model Evaluation	81
6.7.1	Uniformly Distributed Data	81
6.7.2	Skewed Data	83
6.7.3	Normally Distributed Data	83
6.8	PISA Model for Dependent Data	85
6.9	Extension of Models	86
6.10	Applications of Models	87
6.10.1	Savings of R^*_a -tree Depending on the Query Box Size and Form	87
6.10.2	Savings of R^*_a -tree Depending on the Number of Dimensions	87
6.11	Summary	88
7	Techniques for Comparing Index Structures	91
7.1	Introduction	91
7.2	Experimental Parameters	91
7.2.1	Data Specific Parameters	91
7.2.2	Query Specific Parameters	92
7.2.3	System Specific Parameters	93
7.2.4	Disk Specific Parameters	93
7.2.5	Configuration	94
7.3	Index Structures and Time Estimators	95
7.3.1	Time Measures for Tree-Based Index Structures	95
7.3.2	Time Measures for Bitmap Indexing Techniques	97
7.4	Classification Trees	98
7.4.1	Applied Methods	99
7.4.2	Value Sets of Parameters	100
7.4.3	Results	101

7.5	Statistics in Two Dimensions	103
7.5.1	Sum Aggregation	104
7.5.2	Median Aggregation	105
7.5.3	Count Aggregation	105
7.5.4	Results.....	106
7.6	Summary.....	109
8	Conclusion and Outlook	113
A	List of Symbols	117
B	Approximation of PISA Model	123
	Bibliography.....	125
	Index	131