Index Structures for Data Warehouses

D i s s e r t a t i o n
am
Fachbereich für Mathematik und Informatik
an der
Freien Universität Berlin

eingereicht am 22.12.1999

von

Marcus Jürgens

Betreuer
Prof. Dr. Hans-Joachim Lenz (Freie Universität Berlin)
Prof. Dr. Heinz Schweppe (Freie Universität Berlin)
Prof. Dr. Johann Christoph Freytag (Humboldt-Universität zu Berlin)

Tag der mündlichen Prüfung: 16. Feb. 2000

to Chris

Abstract

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 made 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 resign trying to correct my English.

Contents