ON OPTIMAL CONSTRAINT DECOMPOSITION, MONITORING, AND MANAGEMENT IN DISTRIBUTED ENVIRONMENTS

by

Samuel E. Varas G.
A Dissertation Submitted
to the Graduate Faculty
of
George Mason University
in Partial Fulfillment of
the Requirements for the Degree
of
Doctor of Philosophy
Information Technology

Committee:

_____   Larry Kerschberg, Dissertation Director

_____   Alex Brodsky, Dissertation Director

_____   Daniel Barbará

_____   Daniel Menascé

_____   Edgar H. Sibley

_____   Carl M. Harris, Associate Dean for
Graduate Studies and Research

_____   Lloyd J. Griffiths, Dean, School of
Information Technology and Engineering

Date: _____   Summer, 1998
George Mason University
Fairfax, Virginia

On Optimal Constraint Decomposition, Monitoring, and Management in Distributed Environments

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy Information Technology at George Mason University.

By

Samuel E. Varas G.
B.S. Industrial Engineering, University of Chile, Chile, 1989
M.S. Industrial Engineering, University of Chile, Chile, 1989

Directors: Dr. Larry Kerschberg and Dr. Alex Brodsky
School of Information Technology and Engineering

Summer 1998
George Mason University
Fairfax, Virginia

Dedication

I would like to dedicate this work to my parents, Carmen and Samuel, whom always taught me to give my best effort, and to my wife Paula, and our daughter Paula Antonia who understood me and shared their life with me.

Acknowledgments

I would like to thank my advisors Dr. Larry Kerschberg and Dr. Alex Brodsky for their support and exceptional guidance and encouragement to understand new paradigms. I would also like to thank the other members of my dissertation committee, Dr. Daniel Barbará, Dr. Daniel Menascé, and Dr. Edgar H. Sibley for their helpful comments and support during the research and writing of this dissertation.

I would like to thank the Industrial Engineering Department, of the Physical Sciences and Mathematics School, of the University of Chile for its support. I would also like to thank the Government of Chile for the Scholarship "Beca Presidente de la República" granted to me.

I would also like thank to the Defense Advanced Research Project Agency (DARPA) and the Advanced Logistics Program under contract number N00600-96-D-3202 for supporting, in part, this dissertation.

Table of Contents

List of Figures

Abstract

ON OPTIMAL CONSTRAINT DECOMPOSITION, MONITORING,
AND MANAGEMENT IN DISTRIBUTED ENVIRONMENTS

Samuel E. Varas G.

George Mason University, 1998

Dissertation Directors: Dr. Larry Kerschberg and Dr. Alex Brodsky

This dissertation addresses three distributed database management problems: integrity constraint management, optimal view materialization, and quasi-view optimal decomposition. First, it considers the problem of decomposing global integrity constraints in a distributed database. Decompositions are performed in order to save communication and other distributed processing costs, since if during a local update the corresponding local constraint is satisfied, no distributed global constraint checking is necessary. This dissertation addresses the problem of deriving the best possible decompositions, both during database design and at update time. It formulates a generic powerful framework for finding optimal decompositions for a range of design and query-time scenarios. It also provides a comprehensive solution for the family of unrestricted linear constraints. Linear (integrity) constraints are widely used in (distributed) applications such as: resource allocation, ticket reservations, financial transactions, and logistics.

The comprehensive optimization-based solution includes: (1) reducing the problem to one of mathematical programming, (2) developing effective algorithms for it, and (3) providing a distributed a protocol to manage updates and constraint decompositions, while guaranteeing desirable properties of consistency, availability, and optimality. Second, the optimal view materialization problem is addressed, where for set of views (queries), one must decide which additional (intermediate) views should be materialized in order to reduce the computational effort of maintaining the views updated. It introduces a generic optimization framework to decide optimally those (intermediate) materialized views. It uses expression-DAG (Directed Acyclic Graph) as a mechanism to represent equivalent view evaluation plans, and shows that the optimal view materialization problem, under certain objective function conditions, is equivalent to finding a constrained shortest path in an expression-DAG. For those cases where the optimal solution is an expression-DAG path, a linear-time algorithm is presented. Third, the concept of quasi-view (a view with explicit update conditions) is extended considerably in this dissertation. The problem of deciding on the optimal quasi-view decomposition is addressed. This problem reduces to the optimal view materialization and constraint decomposition problems. Although the quasi-view decomposition problem is not a separable one, a solution strategy is presented in terms of the view materialization and constraint decomposition problems.

Chapter 1

INTRODUCTION

Decentralized and distributed architectures pose unique problems for database management systems, particularly in the area of constraint management. Distributed architectures can be characterized by: (1) autonomous and distributed data sources connected by loosely-coupled networks, (2) an increasing number of users with complex requirements, and (3) large amounts of multimedia data and information being gathered, cataloged and stored. Distributed database solutions assigned most of the work to local processing, thereby saving on communication costs. However, this decentralization requires additional costs in coordination and control to ensure proper system behavior.

This dissertation investigates and develops optimal solutions for three distributed database management problems: integrity constraint management, optimal view materialization, and optimal quasi-view decomposition. These problems are interrelated in that the solution techniques use results obtained for other problems as shown in Figure 1.1. The arrow in Figure 1.1 denote the "uses" relationship, e.g., the solution

1

to the "Distributed Constraint Management" problem uses the results and machinery developed to solve the "Optimal Constraint Decomposition" problem. The next section describes a general problem characterization and provides an overview of the contributions for each problem addressed, the complete explanation is found at the respective chapters. Finally, Section 1.2 presents the overall organization for this dissertation.

1.1 Problem Characterization

The problems addressed in this dissertation are interrelated, but each problem can be addressed individually. Figure 1.1 presents the relation among the problems. The research begins by providing a framework for distributed constraint management. Next, it is shown that under certain conditions, one may use the framework to obtain optimal constraint decompositions. The problem of optimizing materialized views is an independent problem. Finally, it is shown that the optimal quasi-view materialization problem involves the novel use of both optimal constraint decomposition and optimal view materialization.

Now, each problem is characterized and an overview of the contributions is provided.

1.1.1 Optimal Integrity Constraint Management

Traditional protocols used to manage global distributed constraints (e.g., two-phase commit) incur enormous overhead and have limitations. In order to reduce those costs and limitations, the idea of local constraint verification has been studied in
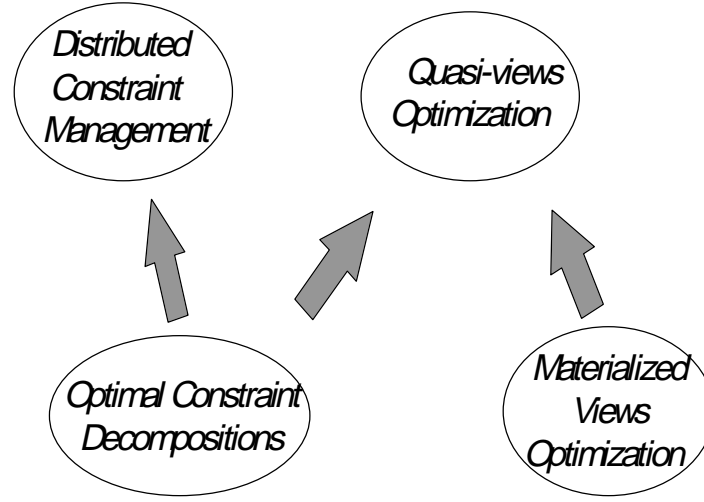
Figure 1.1: Problems Addressed

[BGM92, GM91, GW93, Maz93, Qia89, SV86, GSE$^+$97, Huy97]. This idea decomposes a global constraint into a set of local ones, such that they provide a sufficiency test of the global constraint, i.e., if the local constraints are satisfied, then the global constraint is also satisfied. Furthermore, those sufficiency tests allow for autonomous operation, which is important when some sites are unavailable or the network is partitioned. This problem is important but has not been addressed effectively: (1) most of the previous work has been oriented to finding feasible tests, but they have not addressed the problem of finding the *optimal* one, (2) sufficiency tests require frequent re-definition, because certain database instances may satisfy global constraints, while violating current sufficiency tests, and (3) this dynamic re-definition of sufficiency tests requires a protocol to coordinate and guarantee correct system operation. This moti-

vates the research presented in this thesis.

The first problem addressed in this dissertation corresponds to optimal constraint decomposition. Here a global constraint Ω is decomposed into a set of optimal localized constraints $C_1, \dots, C_M$, such that they constitute sufficient satisfiability tests for Ω, under various decomposition scenarios. In general, this optimization problem is intractable, since there is no bounded characterization for such decompositions. However, this dissertation proposes a characterization of feasible solutions within a general optimization framework. Moreover, for the case where Ω is represented by a set of general linear constraints, a comprehensive solution is provided which reduces the decomposition problem to a standard mathematical programming problem, and efficient algorithms are provided.

The second problem addressed corresponds to that of distributed constraint management, i.e., the protocol to manage updates and effect concurrent linear integrity constraint decompositions. Even though this problem has been addressed before in [BGM92], this dissertation extends considerably its results by providing a protocol that guarantees properties of consistency, safety, last-resort update refusal, and optimality under communication and site failures. The protocol makes use of previous machinery to decompose and re-decompose a global constraint.

1.1.2 Optimal View Materialization

The evaluation of materialized views (queries) may require considerable computational effort because some materialized views (1) can share some intermediate results (views) with other materialized views, or (2) are complex enough to justify some intermediate pre-computed views. This optimal view materialization problem can be formulated as follows: Given a set of materialized views (queries) $\mathcal{V}$, defined over a set of base-relations or views $\mathcal{R}$, then, one must decide which additional (intermediate) views $\mathcal{V}^*$ should be materialized in order to optimize some criterion (e.g., maintenance costs, response time, etc.), while a set of materialization constraints is satisfied (e.g., maintenance time, available storage, budget, etc.). In general, this problem is NP-hard [Gup97], because, for a given set of possible additional views, it corresponds to selecting a subset of elements, in which the number of subsets is exponential in the number of additional views.

This dissertation solves the optimal view materiazation problem by providing: (1) a mechanism to compactly represent intermediate views (queries) called expression-DAG (Direct Acyclic Graphs) is introduced and extended from [RSS96] [1], (2) an optimization framework is proposed to decide the set $\mathcal{V}^*$ based on a dynamic programming approach, and (3) when a complete path (i.e., expression-path or AND-path) characterizes the solution of this problem, a linear-time algorithm (in terms of the expression-DAG size) is presented.

---

[1]Expression-DAG and AND-OR graphs are used to represent intermediate views.

Note that this special case is found in important applications such as the case where the evaluation time is the critical variable, and therefore, although more intermediate views are materialized, the complete evaluation should be more efficient.

1.1.3   Optimal Decomposition of Quasi-Views

Many applications must support the monitoring of distributed data for the occurrence of critical events or complex conditions among data items. The concept of quasi-view (i.e., views with explicit re-materialization conditions called *refresh conditions*) has been introduced as a mechanism to refresh views based on the refresh conditions (monitoring conditions). The concept of quasi-view, introduced in [Sel94], is based on the notion of quasi-copies [ABGM90]. However, [Sel94, SK97] did not address the problem of multiple data sources and global events, such as those found in distributed environments.

The problem addressed in this dissertation is how a quasi-view is evaluated efficiently over a distributed environment. In general, this problem is complex because it requires the efficient coordination of the view materialization strategy with the refresh condition strategy. This dissertation proposes a general solution strategy, where the problem is reduced to the optimal view materialization and constraint decomposition problem, and for the special case of refresh conditions represented as disjunctive arithmetic linear constraints, the refresh condition part corresponds to a special case of the "compact split decomposition" problem formulated in Chapter 2.

1.2   Organization

Chapter 2 addresses both the optimal constraint decomposition and distributed constraint management problems. It proposes an optimization framework to find the best decomposition under multiple scenarios. For the general linear constraint case, a mathematical programming reduction and algorithm are provided. In addition a distributed protocol to manage global linear integrity constraints is presented. Chapter 3 addresses the problem of optimal view materialization, which is formulated as an optimization problem, and for some special cases (where a complete materialization path should be selected) a linear-time algorithm is proposed. Chapter 4 addresses the problem of optimal quasi-view decomposition. An optimization framework is proposed to decompose quasi-views, which is based on both the constraint decomposition and the view materialization problems. Finally, Chapter 5 presents the conclusions, a summary of contributions and suggestions for future research.