Ain Shams University

Faculty of Engineering

Query Optimization For Distributed DataBase Systems

By

Mona Ahmed Fahmy Abdel Baky

A Thesis

Submitted in fulfilment of the

requirements of the degree of PH.D.

In Electrical Engineering

(Computer and Systems Engineering)

Supervised By

Prof. DR. M. Adeeb R. Ghonaimy

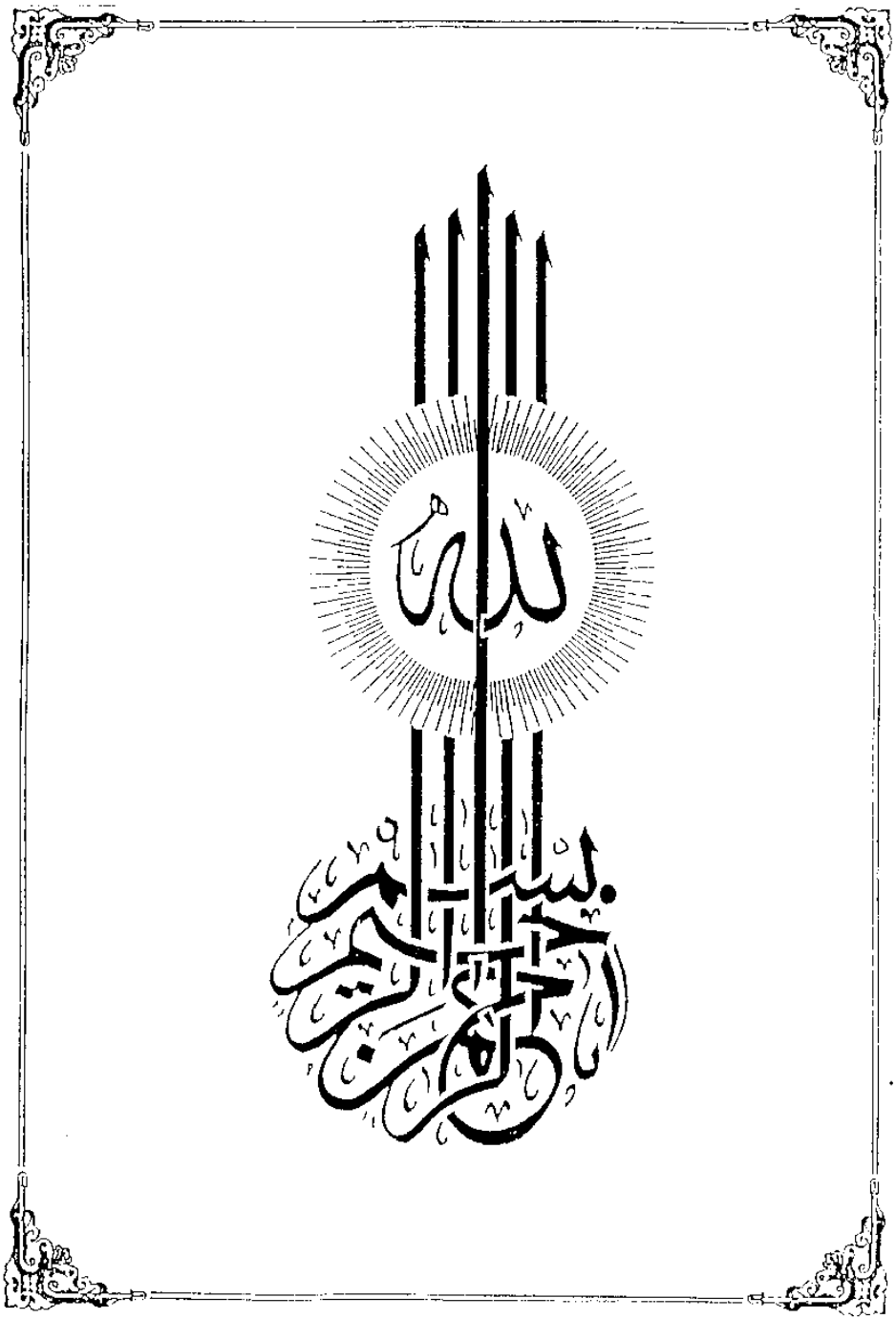
Prof. DR. Osman A. Badr

CAIRO
(1992)

Examiners Commitee

Name, Title and Affiliation	*Signature*
1. *Prof.DR. R.A. Amer* *Prof. at King Abdel Aziz University, Saudi Arabia*	
2. *Prof.DR. F.H.Saleh* *Prof. at Cairo University, Egypt*	
3. *Prof.DR. M.A.R. Ghonaimy* *Prof. at Ain Aham University, Egypt*	

Date : 26/4/1992



Statement

This dissertain is submitted to Ain Shams University for the degree of PH.D. in Electrical Engineering (Computer and Systems Engineering).

The work included in this thesis, was carried out by the auther in the Depertment of Computer and Systems Engineering , Faculty of Engineering, Ain shams University from 11/1/1986 to 2/4/ 1992.

No part of this thesis has been submitted for a degree or a qualification at any other university or institution

Date :

Name :

Signature :

Acknowledgement

*It is hard to express my gratitude to Professor Dr. **M. Adib Ghoneimy** for his invaluable advice, guidance, and supervision through his ever-flowing information represented in references, text books, and papers, and his comprehensive and constructive comments which were great provisions that helped me to achieve my goal and complete my work.*

*I would like also to express my sincere thanks to Professor Dr. **Osman Badr** for his support during my research, and encouragement that helped me to carry on through critical moments.*

*Also, I would like to thank my colleagues; especially Eng. **Hoda Korashy** for her sincere cooperation and the mutual technical constructive discussions that helped a lot, and to Dr. Hassan **Shehata** for his aid in facilitating and providing for me resources, especially in critical times.*

Finally, I would like to express my deepest feelings of gratitude to my whole family, especially my mother, my husband, and my son and daughter.

Abstract

Most of the work done in this area give a good sequence or a close to optimal sequence for join queries, by using *heuristics* based on the semi-join operations. Most of these algorithms are implemented for a certain environment and can not solve the whole problem. Other works use *dynamic programming*, this approach is only good when queries reference a small number of relations because the search space becomes large quickly as the number of relations increase. In fact this rapid growth of the search space limits the usefulness of dynamic programming.

The suggested work is based on using an A^* algorithm which is a known heuristic search technique in Artificial Intelligence. In A^* algorithm the solution can be found without necessarily having to compute the cost of all states or even having to construct them.

The suggested work has been designed to cover the following:

1. **Selecting one copy of each relation referenced by the query.**
2. **Elimination of sending unnecessary relations to assembly site.**
3. **Constant and variable Communication cost between sites, and local processing costs.**

Simulation has been utilized to study the effect of the following parameters on response time of the query:

1. **Input parameters, such as, rate of queries, and distribution of queries among sites.**
2. **System parameters, such as, communication channel capacity and protocols, and computers.**
3. **Configuration parameters, such as, allocation of relations.**

Table Of Contents

CHAPTER 1:Thesis Objective And Overview 1

CHAPTER 2:Query Processing Optimization In Distributed DataBase Systems 3
2.1. Database Models 4
2.2. Computer Network 6
2.3. Queries 10
2.4. Relational Algebra Operators 17
2.5. Distributed Database 25
2.6. Framework For Query Processing Optimization 32
2.7. Factors Affecting The Query Processing Optimization 36
2.8. Join-Query Processing Optimization 39
2.9. Database Machine 42
2.10. Multiple Query 47

CHAPTER 3:Models Of Distributed DataBase For Query Processing Optimization 51
3.1. Problem Formulation 52
3.2. The Transformation Approach 58
3.3. Three Phases 60
3.4. Copy Identification Phase 62
3.5. Fragment Processing 64
3.6. Optimal Strategy For Simple Queries 67
3.7. Heuristics Algorithms Based On Semi-joins . . 68
3.8. Query Processing Optimization Using Algorithms Based On Joins 82
3.9. Query Processing Optimization Using Dynamic Programming 87

CHAPTER 4:Suggested Model For Query Processing Optimization Of Distributed DataBase 93
4.1. The Suggested Model Overview 94
4.2. A*-Optimal search For An Optimal Solution 99
4.3. A Heuristic Function h Developed For The Suggested Model 102
4.4. Suggested Method 108
4.5. Illustrative Examples 115
4.6. Computer Outputs Of Suggested Method 123

CHAPTER5:Sensitivity Analysis of Query Processing Optimization In Distributed DataBase Systems . . . 127
5.1. Why Simulation 128
5.2. Simulated Model 128
5.3. Confidence Level 133
5.4. NETWORK II.5 136
5.5. The Results Of Sensitivity Analysis 139

CHAPTER 6:Conclusion and Further Work 172
6.1.Conclusion 172
6.2.Further Work 174

REFERENCES: 175

APPENDIX A: A/1

APPENDIX B: B/1

CHAPTER 1

CHAPTER 1: Thesis Objective And Overview 1

CHAPTER 1

Thesis Objective And Overview

Distributed databases are important for economical, organizational and technological reasons. They can be implemented in large geographical computer networks and in small local networks. The number of applications of distributed databases will grow in the next years, as distributed database management systems will become available.

Distributed database technology extends traditional database technology in a nontrivial way. In this environment, several technical problems require different solutions, and several completely new issues arise.

This research is concerned with the problem of query processing optimization in distributed database. The goal of query optimization is to find an execution strategy which minimizes the cost of a query, given by the sum of the transmission costs and of the local processing costs. Depending on the characteristics of the communication network and of the local systems, local processing costs can be disregarded in the optimization of the execution strategy of the query. This is typically the case of large geographical networks with low bandwidth, which have transfer rates

of an order of magnitude lower than the disk-to-memory transfer rate. On the Contrary, for fast local networks local processing costs must be considered.

Most of the work done in this area gives a good sequence or a close to optimal sequence for join queries, by using *heuristics* based on the semi-join operations. Most of these algorithms are implemented for a certain environment and cannot solve the whole problem. Other works use ***dynamic programming***, this approach is only good when queries reference a small number of relations because the search space becomes large quickly as the number of relations increase. In fact this rapid growth of the search space limits the usefulness of dynamic programming.

The suggested work is based on using **A*** algorithm which is a known heuristic search technique in Artificial Intelligence. In **A*** algorithm the solution can be found without necessarily having to compute the cost of all states or even having to construct them.

This thesis is organized as follows:

Chapter 2 discusses the aspects of databases and computer networks that are required in order to understand the rest of the thesis. It also discusses the query processing optimization in distributed database systems.

Chapter 3 illustrates the formulation of the query processing optimization problem and the different models of distributed database for query processing optimization.

Chapter 4 is concerned with the suggested model for query processing optimization in distributed database.

Chapter 5 studies the sensitivity analysis of query processing optimization in distributed database. The study is done using simulation techniques. NETWORK II.5 version 3 is the simulation package used.

Chapter 6 concludes the thesis and points to further work.

CHAPTER 2

CHAPTER 2:Query Processing Optimization In Distributed DataBase Systems 3

2.1. Database Models 4

2.2. Computer Network 6
2.2.1. Types of Communication Network 6
2.2.2. Network Architectures 8

2.3. Queries 10
2.3.1. Query Language Types : 10
2.3.2. Query Classification 13
2.3.3. Tree Queries Versus Cyclic Queries . 14

2.4. Relational Algebra Operators 17
2.4.1. Definition of Relational Algebra Operators 17
2.4.2. Estimation of Relational Algebra Operators 18

2.5. Distributed Database 25
2.5.1. Data distribution 25
2.5.2. Distributed Database Classification 27
2.5.3. Fragmentation 28
2.5.4. Data Redundancy 31
2.5.5. Data Dictionary 31

2.6. Framework For Query Processing Optimization 32
2.6.1. Query Processing Optimization Objectives 32
2.6.2. Query Processing Optimization Problems 34

2.7. Factors Affecting The Query Processing Optimization 36
2.7.1. Communication Networks And Query Processing Optimization 36
2.7.2. Computer Hardware And Query Processing Optimization: 36
2.7.3. Database And Query Processing Optimization: 37

2.8. Join-Query Processing Optimization 39
2.8.1. Algorithms which use semi-joins . . . 40
2.8.2. Algorithms which use joins 41

2.9. Database Machine 42

2.10. Multiple Query 47

CHAPTER 2

Query Processing Optimization In Distribution DataBase Systems

The collection of data is usually referred to as the *database*. The *DBMS*s are designed to improve the productivity of application programmers and facilitate data access by computer end users.
The database that is stored on, or manipulated from one computer only is called a *centralized database*, while if the database is resident in, or manipulated from several computers, it is called a *distributed database*.
A language expression used to enable users to retrieve data from a database is called a *query*. When a query is presented to the system, it is necessary to find the best method of finding the answer using the existing database structure. There are a large number of possible strategies for processing a query , especially if the query is complex. The goal of this research is the query processing optimization in distributed databases.

In this chapter, first, I will review the data models, the basic concepts about computer networks, queries, relational algebra operators, and distributed databases. Then, I will talk about the framework for query processing optimization, and factors affecting it. Then, I present the optimization of Join-Query. Finally, I will talk about special problems of the database machine, and multiple query processing optimization.

2.1. Database Models:

The database models are used in describing data at the conceptual and view levels. There are three database models.

Relational model:

The data and the relationships among data are represented by a collection of tables each of which has a number of columns with unique names. Figure 2.1 is a sample relational database. For more details about relational model and its normal forms (first, second, third, boyce/codd, fourth, and fifth normal forms) see [DATE74], [MART75], [KRON77], and [DATE86].

Name	Street	City	Number
Lowery	Maple	Queens	900
Shiver	North	Bronx	556
Shiver	North	Bronx	647
Hodges	Sidehill	Brooklyn	801
Hodges	Sidehill	Brooklyn	647

Number	Balance
900	55
556	100000
647	105366
801	10533

Figure 2.1
A sample relational database.

Network model:

Data in the network model are represented by a collection of records and relationships among data are represented by links, which can be viewed as pointers. The records in the data base are organized as collections of arbitrary graphs. Figure 2.2 is a sample network database that has the same information as in