SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Tuning Linear Programming Solvers for Query Optimization

Sarra Ben Mohamed

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Tuning Linear Programming Solvers for Query Optimization

Anpassung von Linear Programming Solvern für Anfrageoptimierung

Author:	Sarra Ben Mohamed
Supervisor:	Prof. Dr. Thomas Neumann
Advisor:	Altan Birler
Submission Date:	15/10/2023

I confirm that this bachelor's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 15/10/2023 Sarra Ben Mohamed

Acknowledgments

Abstract

Contents

1 Introduction

Our aim with this project is to investigate and compare different methods and techniques to solve small linear programming problems representing the problem of cardinality estimation. Our goal is to estimate realistic and useful upper bounds on query sizes. Studies have shown that cardinality estimation is the major root of many issues in query optimization. [Ngo22] And yet, theoretical upper bounds that are way too large would be useless since we want practical estimation to choose the best from data plans to run efficient queries. For this purpouse, we will introduce a formal description of the cardinality estimation problem, represent it in the form of a packing linear programming problem with the intention of maximizing the size of the query under some constraints. The result is hundreds of relatively small LP that we collect in datasets and solve them with different methods and algorithms. We then draw conclusions based on the results of our experiments, benchmarks and the previous work done on similar packing LP problems. This should guide us into constructing a thorough analysis of the particularities of these LP problems, what's unique about their structure and if their solution process is following any patterns. We then discuss and draw hypotheses on the ways this analysis can be exploited to further optimize the solution process: which methods or combination of methods deliver the best time and memory complexity.

2 Related work

2.1 Background

or Fundamentals: the knowledge the reader needs to understand my contribution, mostly definition of mathematical concepts needed

2.1.1 Linear Programming

2.1.2 Cardinality Estimation

Defining the problem of upper bounding a multi-join query size as a packing linear programming problem. To illustrate the main ideas, we start with an example where the query is a simple join between two relations

$$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$$

In the context of our packing LP problem, we start with the inequality 2.1. Applying the natural logarithm to both sides yields 2.2. We then rename the variables, simplifying the inequality to 2.3. Normalizing by dividing both sides by r', we obtain 2.4. This leads us to the objective function for our packing LP problem.

$$|a| \cdot |b| \leq |R| \tag{2.1}$$

$$\ln |a| + \ln |b| \leq \ln |R| \tag{2.2}$$

$$a' + b' \leq r' \tag{2.3}$$

$$\frac{1}{r'}a' + \frac{1}{r'}b' \leq 1 \tag{2.4}$$

$$\text{maximize } a' + b' + c' + d' \quad \text{s.t.} \quad \frac{1}{r'}a' + \frac{1}{r'}b' \leq 1 \tag{2.5}$$

2.1.3 The Simplex Algorithm

2.2 Previous Work

alternative approaches that are superseded by my work

2.2.1 Comparative studies of different update methods

2.2.2 Other techniques

3 Tuning Linear Programming Solvers for Query Optimization

This is the body

3.1 Proposal

3.2 Experimental Design

3.2.1 Analysis of dataset properties

In this subsection we will conduct an analysis of our dataset properties. What are the particularites of the structure of these LP problems, is their any patterns in their solution process. This anaylsis is based on observing the statistical results we obtained from running different solvers on these problems. This will later provide us with insight regarding optimization of these problems.

3.2.2 Dataset Structure

Our dataset stucture: as opposed to what the linear programming research has dealt with, which is very large problems, we are dealing with hundreds of small problems. These are represented in the revised simplex algorithm by sparse matrices but not as sparse as it would have been if the problem was large, small matrices that are not small enough to be dense. (they still have quite a number of non-zeroes).

3.3 Analysis

3.4 Results

4 Evaluation

4.1 Setup

4.1.1 Evaluation metrics

4.1.2 Evaluation baselines

4.2 Results

4.3 Discussion

5 Conclusion

List of Figures

List of Tables

Bibliography

[Ngo22] H. Q. Ngo. “On an Information Theoretic Approach to Cardinality Estimation (Invited Talk).” In: *25th International Conference on Database Theory (ICDT 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2022.