SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

Bachelor's Thesis in Informatics

Improved Symbol Table Construction for FSST Compression

Hedi Chehaidar

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS
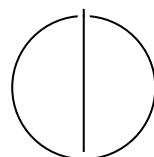
TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Improved Symbol Table Construction for FSST Compression

Verbesserte Symboltabellenkonstruktion für die FSST-Komprimierung

Author:	Hedi Chehaidar
Examiner:	Prof. Thomas Neumann
Supervisor:	Mihail Stoian
Submission Date:	January 7, 2026

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

Munich, January 7, 2026   Hedi Chehaidar

Acknowledgments

First and foremost, I would like to thank God for the guidance, strength and perseverance that made it possible to complete this work.

I would also like to express my sincere gratitude to all the people who supported, inspired, and challenged me throughout the work on this thesis.

I would like to thank my supervisor Mihail Stoian for the continuous guidance, insightful feedback, and trust in my ideas. The discussions we had, ranging from high-level algorithmic ideas to low-level implementation details, were invaluable and significantly shaped the direction and quality of this work. I am especially grateful for the encouragement to explore optimizations deeply and to question existing assumptions about compression performance.

I am deeply thankful for the practical insights, constructive criticism, and technical expertise, particularly regarding algorithmic correctness and performance evaluation. The feedback I received helped turn initial ideas into a well-structured and measurable contribution.

I am grateful to the broader research and open-source community whose work laid the foundation for this thesis. The availability of high-quality implementations, documentation, and discussions around compression techniques made it possible to build upon existing work and focus on meaningful improvements.

A special thanks goes to everyone who shared tools, datasets, or prior experience that helped accelerate experimentation and evaluation.

Finally, I want to thank my friends and family for their constant support, patience, and encouragement throughout this journey. Their motivation helped me stay focused and enthusiastic, especially during long debugging sessions and performance tuning phases.

This thesis would not have been possible without the collective contributions of all these people, and I am truly thankful for their support and inspiration.

Abstract

Modern analytical database systems process large volumes of string data, where efficient compression is essential for reducing memory footprint and improving performance. While the Fast Static Symbol Table (FSST) compression scheme provides excellent decompression speed and random-access capabilities, its effectiveness heavily depends on the quality of the constructed symbol table.

The original FSST algorithm relies on heuristic, greedy symbol selection, which can lead to suboptimal symbol choices and limit achievable compression ratios.

This thesis presents several enhancements to the FSST compression method that improve symbol table construction while preserving FSST's core advantages. The central contribution is a refined symbol selection process that systematically identifies more effective symbols and avoids redundant or conflicting choices.

First, a dynamic programming approach is introduced to evaluate and select higher-quality symbols within each generation, enabling a more globally informed optimization compared to the original greedy strategy. Second, an additional frequency counter is incorporated to accelerate the discovery of longer symbols and to explicitly favor them in subsequent generations, improving the exploitation of longer recurring patterns in the data. Third, a symbol pruning mechanism is applied to eliminate conflicting and redundant symbols, ensuring a more compact and effective symbol table.

Together, these techniques significantly improve the robustness and quality of the symbol table generation process. Experimental evaluation demonstrates that the proposed enhancements lead to consistently improved compression ratios compared to the original FSST algorithm, while maintaining its fast decompression and random-access properties. The results show that careful algorithmic refinement of symbol selection can yield substantial gains without altering the lightweight and practical nature of FSST, making the improved approach well suited for use in modern analytical systems.

Contents

1 Introduction

1.1 Section

List of Figures

List of Tables