Overview

The goal of this master thesis is to investigate string sorting on distributed systems. Sorting strings or vectors is a basic algorithmic challenge different from integer sorting because it is important to access components of the keys to avoid repeated operations on the entire string.

In a distributed scenario, each machine can first sort its local set of strings. But then the global order has to be determined by exchanging prefixes or hashes of prefixes of the remaining distinguishing characters. Multiple algorithms are conceivable to solve this problem.

For example, the complete remaining strings could be exchanged with compression of common prefixes. Alternatively, a prefix doubling approach with order-preserving hashing and distributed duplication detection [1] can be applied.

The newly designed algorithms should be analyzed theoretically if possible, and implemented and evaluated on a cluster system. Depending on the algorithm and setting, either Thrill [2] or MPI should be used as communication framework.

This master thesis builds on previous work on sequential string sorting and parallel string sorting on multi-core systems [3, 4].

Requirements

- Interest in string algorithms and distributed big data processing
- Excellent programming skills in C++
- Ability to think for yourself

Info: Other distributed algorithm topics are available.


