Parallelization of the corpus manager’s time-consuming operations

What is a corpus manager?

The corpus managers enable corpus exploration -- corpus is huge collections of texts. Corpora are used as a resource of the empirical language data, i.e. words, their meanings and contexts they occur in. Corpora can be employed in many fields of linguistics (morphology, syntax, semantics, sociolinguistics etc.). However, corpora cannot provide all informations that we, humans, have about a language. But the amount of missing pieces is smaller as the corpora are larger. Therefore, the future is in corpora with ten billions of words.

Time-consuming operations?

Each operation processing large amount of data can be time-consuming, which is unpleasant for users as they expect the results to be computed in a second. The time-consuming operations also consume hardware resources for a long time, so all other operations are slowed-down.

Parallelization?

Our parallelization model uses a cluster of cheap commodity computers. It provides faster computation than high-cost super computer and an operation can run in parallel on all computers of the cluster. Each computer works only with a small piece of the whole corpus, so the operation is executed fast. At the end, the results must be merged. Our model is very similar to the MapReduce model which was introduced by the Google company to handle internet search.

Do we use MapReduce?

Our implementation uses very simple model of communication that makes it faster than MapReduce for common operations. The thesis also contains comparison of our system with MapReduce.

Did we managed to solve problems with the corpus manager’s time-consuming operations?

Yes, we did! Our implementation can compute the corpus manager’s time-consuming operations in seconds instead of minutes or hours.

Publications

RÁBARA, Radoslav; RYCHLÝ , Pavel. Concurrent Processing of Text Corpus Queries. In Ninth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2015. pp. 49-58, 10 p. ISBN 978-80-263-0974-1.