7a.029.TUT - Very Fast Nearest Neighbor Retrieval in High-Dimensional Domains

Teemu Roos (PI), Petri Myllymäki, Gür Ersalan, Ville Hyvönen, Jukka Kohonen, Janne Leppä-aho
University of Helsinki, Department of Computer Science / HIIT

Project Start: 8/1/2018	End Date: 7/31/19	Project Budget: € 90 000	Spent:

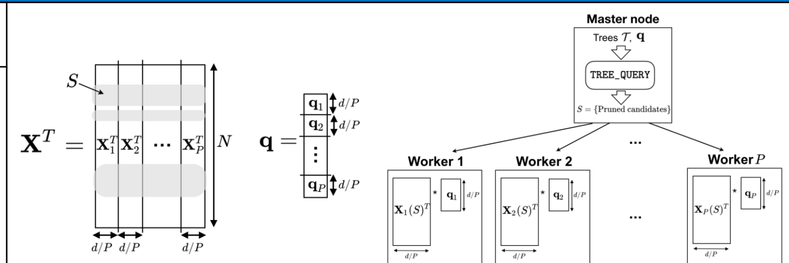
Project Summary:

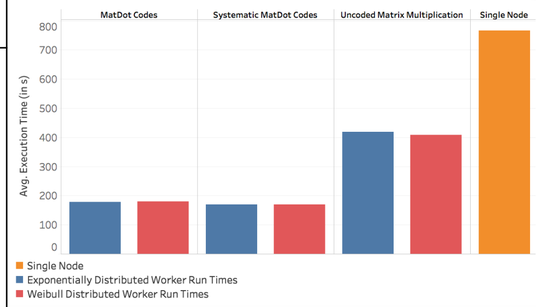
Nearest neighbor search is a core part of several machine learning algorithms. It finds applications in areas such as information retrieval, clustering and visualization. In many of the problem domains, the amount of samples and the dimensionality of data render the exact neighbor search infeasible.
This project investigates variants of approximate nearest neighbor (ANN) algorithms based on random projection trees. The objectives are to develop and evaluate: i) scalable algorithms for ultrahigh-dimensional data by exploiting sparsity; ii) incremental algorithms for dynamic (streaming) data; iii) robust (fault-tolerant) parallelization techniques on GPU, cluster and cloud platforms. Applications involve text, audio, image and other signal data.

Details of Progress/Achievements:

We have applied MRPT (multiple random projection trees) method to a large-scale text classification task. MRPT made nearest-neighbor based classification possible as the exact search was computationally infeasible. However, nearest-neighbor classifier itself was outperformed by other methods in this setting.

We have made progress in research regarding the fault-tolerant parallelization techniques and in the automatic tuning of the parameters of the method.





PROJECT DELIVERABLES

Deliverable	Achievements	Remaining To Do
1. 1–2 applications in large-scale document classification, clustering, etc	30 % Complete Method applied in a large-scale text classification task.	Applications with different data sets and tasks.
2. Open-source package for parallelized ANN	20 % Complete Initial version of the fault-tolerant parallelization scheme implemented.	Further testing and development required before a public release.
3. Theoretical framework for the study of parallel ANN and other machine learning procedures.	20 % Complete Research on fault-tolerance and auto-tuning. Two submitted publications.	More research.
4. Incremental algorithms for dynamic (streaming) data.	0% Complete	Research, design and implementation.