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

| Report Date | Project Start | Project End | Project Budget | Amount Spent To Date |
|---|---|---|---|---|
| 1/4/2019 | 8/1/2018 | 7/31/2019 | $ 90 000 | |

### PROJECT SUMMARY


This project investigates variants of approximate nearest neighbor (ANN) search 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.


### PROJECT TEAM

| Team Member Name | Team Role *(PI, Co-PI, Student, Researcher)* | Academic Site |
|---|---|---|
| Teemu Roos | PI | University of Helsinki / HIIT |
| Petri Myllymäki | Co-PI | University of Helsinki / HIIT |
| Gür Ersalan | Student | University of Helsinki / HIIT |
| Ville Hyvönen | Student | University of Helsinki / HIIT |
| Jukka Kohonen | Researcher | University of Helsinki / HIIT |
| Janne Leppä-aho | Student | University of Helsinki / HIIT |
| | | |
| | | |

### IAB PROJECT MENTOR(S)

| IAB Project Mentor Name | IAB Organization |
|---|---|
| Kimmo Valtonen | M-Brain |
| | |

### PROJECT FUNDED BY

| IAB Organization(s) |
|---|
| M-Brain |

# CVDI Year 7 Mid-Year Report

07/01/18 – 12/31/18

**OVERALL PROGRESS/ACHIEVEMENTS TO DATE**

We have made progress in research regarding the fault-tolerant parallelization techniques, and in the automatic tuning of the parameters of the multiple random projection trees (MRPT) method. Research regarding the automatic tuning is already incorporated to an open-source package for ANN search available on GitHub (https://github.com/vioshyvo/mrpt). This work was/will be presented at IEEE Big Data Conference 2018 [1] and PAKDD-2019 [2].

We have applied MRPT method to a large-scale text classification task. The data was provided by our industry partner M-Brain. 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. This work will be reported in the form of a

[1] U. Sheth, S. Dutta, M. Chaudhari, H. Jeong, Y. Yang, J. Kohonen, T. Roos, and P. Grover (2018). *An application of storage-optimal MatDot codes for coded matrix multiplication: Fast k-nearest neighbors estimation*, IEEE Big Data Conference, Seattle, December 10–13, 2018.

[2] E. Jääsaari, V. Hyvönen, T. Roos (2019). *Efficient autotuning of hyperparameters in approximate nearest neighbor search*, to appear in Proc. 23rd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-2019), Macau, China, April 14–17, 2019.

**PROJECT DELIVERABLES**

| Deliverable | Achievements | Remaining To Do |
|---|---|---|
| An open-source package for parallelized ANN. | • A light-weight and user-friendly library for (un-parallelized) ANN search released.<br>• Initial version of the fault-tolerant parallelization scheme implemented. | Further development and maintenance. |
| Concrete applications of ANN in large-scale document classification, clustering, etc. | Method applied in a large-scale text classification task. The data set was provided by our industry partner. | More applications with different data sets and tasks. |
| Theoretical framework for the study of parallel ANN and other machine learning procedures. | Research on fault-tolerance and auto-tuning. Published as [1] and [2] | More research to follow. |
| Optional deliverable: Incremental algorithms for dynamic (streaming) data. | Basic updating protocols exist (on paper). | Research, design, and implementation (if resources allow) |