Project 3
SPM course a.a. 23/24

May 17, 2024

LSH-based similarity join

The Similarity join is a fundamental operation in data analysis that consists of finding pairs of close tuples according to a given distance metric. Formally, a similarity join for two collections of data Γ_R and Γ_S is:

$$\Gamma_R \bowtie_\lambda \Gamma_S = \{(U,V) \in \Gamma_R \times \Gamma_S \mid dist(U,V) < \lambda\}$$

where *dist(U,V)* is a distance function between *U* and *V*, and λ is the threshold parameter.

Naively, similarity join computations can be performed by comparing all the data pairs, thus requiring the computation of the entire Cartesian product. By employing a locality-sensitive hash (LSH) function on tuples, we can divide the whole space into buckets of elements that are similar with high probability. Locality-sensitive hash is based on a hashing scheme that ensures that nearby data points are more likely to collide than distant ones. The random hashing function depends on the type of data and the chosen distance.

Consider the code *lshseq.cpp* code in /opt/SPMcode/LSH folder of the spmcluster fron-end. It implements the brute-force version of the LSH-based similarity join using the Fréchet algorithm for the distance function. Provide a parallel version of this code using FastFlow for a single multicore node and MPI for a cluster of nodes.

The input datasets are provided in the spmcluster machine (front-end node) in the folder /opt/SPMcode/LSH/datasets. For the distributed version, only one process is allowed to read the input dataset from the filesystem. Since the spmcluster is a diskless cluster, the time for reading big input datasets could be relevant. Therefore, for the performance evaluation, you could not consider such time for loading the dataset. The input dataset is a 3-column space-separated values file: the first column is an integer representing the trajectory's ID, the second column represents the collection's ID (0 for U, 1 for V), and the third column is the trajectory. Two tiny datasets (taxi1 and taxi2) are also present in the same folder for quick testing purposes. The validation and performance tests must be conducted considering the larger datasets (lsh1GB.dat, lsh5GB.dat, and lsh10GB.dat), which are synthetic datasets and were automatically generated. Note that the threshold parameter λ (and also the LSH resolution) is different between the synthetic dataset and the taxi datasets. In particular, for the taxi dataset, 0.01 should be used for the threshold and 0.08 for LSH resolution, whereas for the synthetic datasets, 10 and 80 should be used, respectively.

The developed code should be delivered in a tarball (tgz or zip) with a PDF document, a Makefile/Cmake for compiling the source code on the spmcluster machine, and all scripts for running the tests using SLURM. The files used for the tests should not be inserted in the project tarball.

The PDF document must describe the parallelization strategy adopted, the performance analysis, the plots of the speedup/scalability/efficiency obtained by the tests, comments, problems faced, etc. It should be at most 12 pages long.