

1 Problem Definition

Implement a nearest neighbor search technique on grid-indexed data. The algorithm should take as input: a query location q and an integer k and finds the k -nearest neighbors of q using the grid index.

2 Design and Implementation

2.1 Grid construction

Each data record is stored in a data structure named **Record**. The grids are stored in a multi-array for fast accesses. The grid construction algorithm is showed in Figure ?? . The construction needs two scan of the dataset. In the first scan, it finds the maximum latitude and longitude of all location. Then in the second scan, put the record to its corresponding grid.

```
1 visited
2 min_lat, min_long, max_lat, max_long are got in the first scan
3 foreach (r <- records)
4   if not r in visited
5     x = (r.lat - min_lat) / grid_size
6     y = (r.long - min_long) / grid_size
7     grids(x)(y) += r
```

2.2 k-nearest neighbour search on grids

A k-nearest neighbour search query accepts a parameter of location (latitude and longitude). It first finds the corresponding grid of the location, and then find a k-nearest locations in the grid. After that, it searches the layered cells. The algorithm is showed in Figure ?? .

```
1 def (lat, long, k):
2   x = (lat - min_lat) / grid_size
3   y = (lat - min_long) / grid_size
4   k_set = grid_search(grids(x)(y), k_set)
5   while (true):
6     min_d
7     for (cell <- cells layered around x, y):
8       d = min(cell, lat, long)
9       min_d = min(d, min_d)
10      if d < k_set.min or len(k_set) < k:
11        k_set = grid_search(cell, k_set)
12      if (len(k_set) >= k and k_set.max < d_min):
13        break
14      return k_set
15
16 def (grid_records, lat, long, k_set):
17   for (r <- grid_record):
18     if r < k_set.max or len(k_set) < k
```

```

19         k_set.insert(r)
20     return k_set

```

In this algorithm, if the minimum distance between the cell and the location is larger than the maximum value in `k_set`, then the cell does not need to be access. This algorithm stop when the minimum distance between the location and all cells in the layer (none of the cells in the layer has been accessed).

2.3 benchmark

Evaluation of the algorithm is focused of the time of constructing the grids and the average query time. Random locations were generated randomly and these locations are served as queries to the knn search system. The average query time for 10k queries is recorded. A simple linear scan algorithm is used for comparison. The result is showed in Table 1.

name	construction time	query time	cell accessed
grid search	10s	43ms	7800
linear scan	N/A	43ms	10000

Table 1: Evaluation result.