Caching policies on Geolife Trajectory Data

Following Datasets are used for simulation :

- GeoLife GPS Trajectories
- Beijing City Road Map

GeoLife data :

GeoLife GPS Trajectories were collected within the (Microsoft Research Asia) Geolife project by 182 users over a period of nearly five years (from April 2007 to August 2012). [1,2,3] The GeoLife GPS Trajectories download contains many text files organised in multiple directories. The data files are basically CSVs with 6 lines of header information. They contain the following fields:

Field 1: Latitude in decimal degrees.
Field 2: Longitude in decimal degrees.
Field 3: All set to 0 for this dataset.
Field 4: Altitude in feet (-777 if not valid).
Field 5: Date — number of days (with fractional part) that have passed since 12/30/1899.
Field 6: Date as a string.
Field 7: Time as a string.

For more details about the dataset, visit: http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/

Beijing Road Network Data :

It contains 76k nodes, 85k edges of the Beijing City map.

Nodes are given as (x,y) pairs.

Data Preparation

***Step 1**: Extracting query logs from trajectory data*

For each trajectory, the start and end locations are chosen as source and destination nodes. Each query is a set of three attributes, i.e., source node, destination node, and timestamp (date & time of the start of the trip).

This dataset contains 18,670 trajectories.

For more details: https://heremaps.github.io/pptk/tutorials/viewer/geolife.html , http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/

***Step 2**: Mapping lon-lat to x-y (i.e. x <- lon, y <- lat) using UTM projection*

We have applied the UTM projection on GPS points in UTM zone 50 (Beijing's UTM zone).

UTM projection is able to convert lat-lon coordinates into units of meters while only introducing a relatively small amount of stretching/shrinking between points in the same UTM zone.

***Step 3**: Matching Queries with Road Network using Voronoi Diagram*

The x-y GPS queries are now matched with the road network nodes before moving on to the step. In the *Voronoi* approach, each query is matched to its nearest(*Euclidean distance*) possible network node. *(up to a threshold of 150 meters)*

```
Note: Any query with nearest node more than 150 meters away are
dropped. 150 meters is a reasonable value, we tried with other
value as well and then settled for this.
```

***Step 4**: Djikstra's Method to calculate the shortest path from source to destination*

Now, we have a **set of queries** and the **shortest path** for each of them.

Caching

Caching refers to storing some portion of the complete data for faster retrieval and saving cost.

Given the weights and values of n items, we need to put these items in a knapsack of capacity W to get the maximum total value in the knapsack.

```
Input:
Paths as (values, weight) pairs
Paths[] = {{20, 10}, {100, 20}, {120, 30}, {200,20}}
Cache Percent, W = 50%;
Output:
To achieve maximum possible value, we first calculate the Ratio of
values and weight and sort the paths in descending fashion of the
Ratio.

Ratio[] = {2, 5, 4, 10}
Sorted_Ratio[] = {10, 5, 4, 2}
Sorted_paths[] = {{200,20}, {100, 20}, {120, 30}, {20, 10}}
So, for W = 50% : Paths {{200,20}, {100, 20}} will be cached .
Maximum value = 300
```

Performance Evaluation on Test queries

Year-wise distribution of Queries :

Year	Queries
2007	599
2008	3652
2009	4304
2010	1059
2011	942

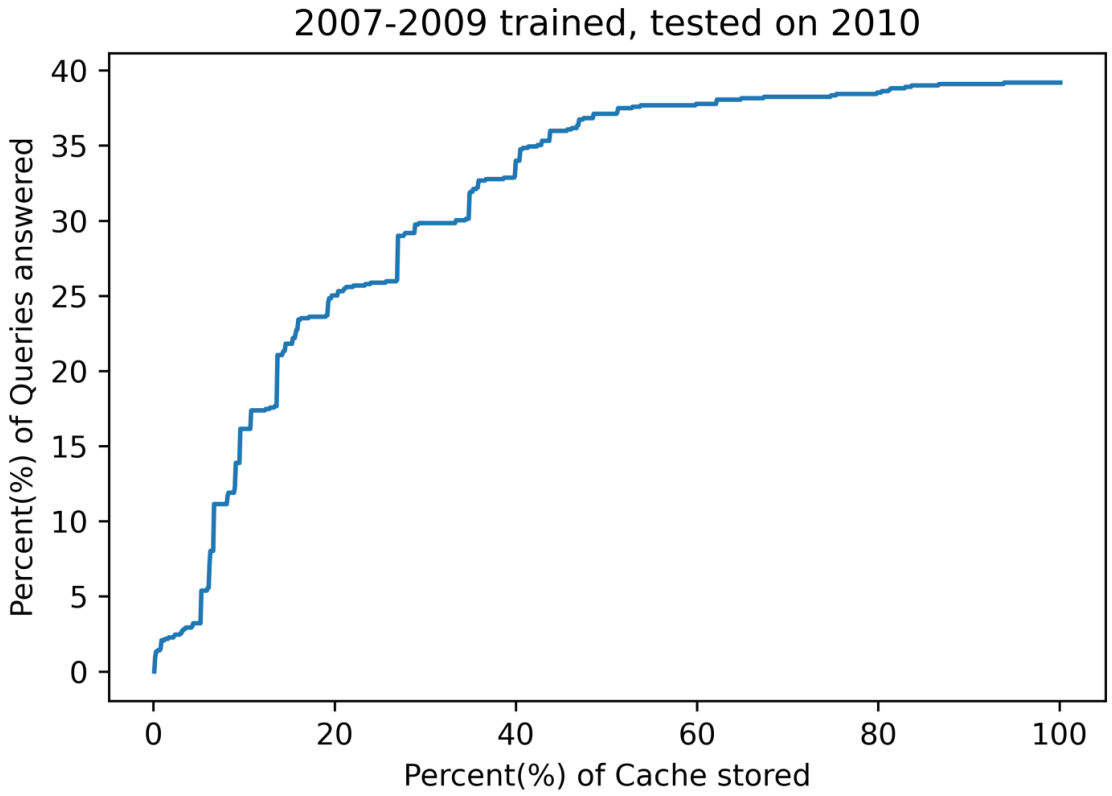
The first three years were used for training based on the caching method discussed above. Due to the nature of the data, the cache generated had some duplicate paths, which were removed before finalising the cache.

And then, the queries from 2010 are used for testing purposes.

Note: A query Q is said to be served by a path P if P contains both the source and destination nodes of Q.

Result

We got around **~40%** of queries that can be served, given complete cache from training data.



References

[1] *Yu Zheng, Lizhu Zhang, Xing Xie, Wei-Ying Ma. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of International conference on World Wild Web (WWW 2009), Madrid Spain. ACM Press: 791–800.*

[2] *Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, Wei-Ying Ma. Understanding Mobility Based on GPS Data. In Proceedings of ACM conference on Ubiquitous Computing (UbiComp 2008), Seoul, Korea. ACM Press: 312–321.*

[3] *Yu Zheng, Xing Xie, Wei-Ying Ma, GeoLife: A Collaborative Social Networking Service among User, location and trajectory. Invited paper, in IEEE Data Engineering Bulletin. 33, 2, 2010, pp. 32–40.*

Code

Code is available on this GitHub repo :

Link : https://github.com/mohit-iitb/contentCaching/tree/master