THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Data pre-processing

2.1 Data collection and preliminary cleaning

The taxi GPS data used in this project is collected from the Computational Sensing Lab[7] at Tsinghua University, Beijing, China. The data set contains approximately 83 million time-stamped taxi GPS records collected from 8,602 taxis in Beijing, from 1 May 2009 to 30 May 2009. The original data set consists of seven fields as shown in Table 2.1. "WGS-84" stands for "World Geodetic System" which is the reference coordinate system used by the GPS.

The original data set comes in a binary file format. After the data is decoded and imported into a MySQL database, the first step in cleaning data is to delete all records with zero value in the SPEED field, since when a taxi is stationary it yields no valuable information about the *trajectory* it is moving along. While being stationary could be due to a traffic jam, this kind of information is well captured by the time difference between the previous *non-stationary* data point and the next *non-stationary* data point.

Moreover, all records must have a *unique* pair of CUID and UNIX_EPOCH fields,

Field	**Explanation**
CUID	ID for each taxi
UNIX_EPOCH	Unix timestamp in milliseconds since 1 January 1970
GPS_LONG	Longitude encoded in WGS-84 multiplied by 10^5
GPS_LAT	Latitude encoded in WGS-84 multiplied by 10^5
HEAD	Heading direction in degrees with 0 denoting North
SPEED	Instantaneous speed in metres/second (m/s)
OCCUPIED	Binary indicator of whether the taxi is hired (1) or not (0)

Table 2.1: Fields in the original data set

since it is not possible for a taxi to appear in two different places at the same moment in time. This is possibly due to some errors in aggregating the original data set.

2.2 Reverse geo-encoding

After the preliminary cleaning of the data set, the next step is to map each GPS data point to a road segment, which is also known as *reverse geo-encoding*. A number of algorithms[3] have been proposed for that purpose, but most of them require an additional GIS[1] database of the road network in Beijing. This project adopts an alternative strategy which leverages on the existing public API[2] for reverse geo-encoding.

Currently, a number of online mapping platforms provide reverse geo-encoding services as part of their developer APIs. Amongst others, Google Maps and Baidu Maps offer relatively stable and fast reverse geo-encoding services. However, due to the “China GPS shift problem”[5] where coordinates encoded in WGS-84 format are required by regulations to be shifted by a large and variable amount when displayed

[1] Geographic Information System

[2] Application Programming Interface

on a street map, Google Maps is not able to display a GPS point correctly.

Baidu Maps, on the other hand, has been using their own coordinate system called BD-09 which is an improved version of the Chinese official coordinate system, GCJ-02. Baidu provides a set of APIs to convert WGS-84-encoded coordinates into BD-09-encoded ones.

In order to use Baidu APIs for reverse geo-encoding, the following system architecture has been set up as shown in Figure 2-1. The Apache HTTP server hides the MySQL database and sends HTTP POST request to Baidu Maps Web API to get converted coordinates. Then it updates the database through PHP *mysqli* utilitity.

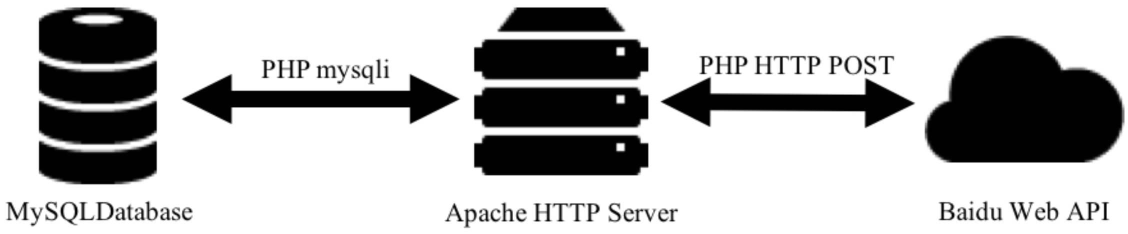

Figure 2-1: System architecture

After the conversion is completed, Baidu Maps API is used to reverse geo-encode all GPS data points.

2.3 Related work

The incentive for carrying out this project comes from a similar project[6], and similar procedures are followed in this project but with some modifications.