



UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería en Informática



TFM del Máster en Ingeniería
Informática

Sistema para el registro y análisis
de trayectorias semánticas
Documentación Técnica



Presentado por David Moreno del Hoyo
en Universidad de Burgos — 10 de febrero de 2017
Tutores: Bruno Baruque y Santiago Porras

Índice general

Índice general	I
Índice de figuras	III
Apéndice A Manuales	1
A.1. Introducción	1
A.2. Planificación temporal	1
A.3. Estudio de viabilidad	1
Apéndice B Especificación de Requisitos	2
B.1. Introducción	2
B.2. Objetivos generales	2
B.3. Catalogo de requisitos	2
B.4. Especificación de requisitos	2
Apéndice C Especificación de diseño	3
C.1. Introducción	3
C.2. Diseño de datos	3
C.3. Diseño procedimental	3
C.4. Diseño arquitectónico	3
Apéndice D Documentación técnica de programación	4
D.1. Introducción	4
D.2. Estructura de directorios	4
D.3. Manual del programador	4
D.4. Compilación, instalación y ejecución del proyecto	4
D.5. Pruebas del sistema	4
D.6. Instalación de herramientas	4
D.7. Manual del usuario	6
D.8. <i>Data cleaning</i>	10

Apéndice E Documentación de usuario	16
E.1. Introducción	16
E.2. Requisitos de usuarios	16
E.3. Instalación	16

Índice de figuras

D.1. Sección de continentes	6
D.2. Países mapeados de Sudamérica	7
D.3. Impresión en consola del resultado del comando.	8
D.4. Descarga de datos desde QGIS	10
D.5. Valores introducidos para la búsqueda de datos en QGIS	11
D.6. Éxito en la descarga de los datos.	12

Apéndice A

Manuales

A.1. Introducción

A.2. Planificación temporal

A.3. Estudio de viabilidad

Viabilidad económica

Viabilidad legal

Apéndice B

Especificación de Requisitos

- B.1. Introducción
- B.2. Objetivos generales
- B.3. Catalogo de requisitos
- B.4. Especificación de requisitos

Especificación de diseño

- C.1. Introducción
- C.2. Diseño de datos
- C.3. Diseño procedimental
- C.4. Diseño arquitectónico

Apéndice D

Documentación técnica de programación

D.1. Introducción

Esta sección pretende aclarar todos los conceptos relativos a las cuestiones técnicas y de programación llevadas a cabo en el presente Trabajo de Fin de Máster.

D.2. Estructura de directorios

D.3. Manual del programador

D.4. Compilación, instalación y ejecución del proyecto

D.5. Pruebas del sistema

D.6. Instalación de herramientas

Virtual Box

El software de virtualización seleccionado ha sido Oracle VM Virtual Box en su versión 5.1. Este programa ha sido instalado sobre Windows 10, sistema que actúa como anfitrión.

La descarga de este software se puede realizar desde la siguiente dirección web: <https://www.virtualbox.org/wiki/Downloads>

Ubuntu 16.04

Como servidor se ha optado por la versión 16.04 de Ubuntu en su versión de escritorio. El sistema ha sido instalado sobre una máquina virtual haciendo uso del software mencionado en la sección anterior.

Para descargar una imagen de este Sistema Operativo se ha de acceder al siguiente enlace: <https://www.ubuntu.com/download/desktop>

PostgreSQL 9.6 y PgAdmin 3

PostgreSQL puede ser instalado en Ubuntu desde la consola de comandos. Para ello se han de seguir unos sencillos pasos. El primer paso es actualizar el repositorio de Ubuntu mediante el siguiente comando sobre la consola:

```
apt-get update
```

Ahora se puede comenzar la instalación de Postgres mediante la siguiente línea en la consola de comandos:

```
apt-get install postgresql postgresql-contrib
```

Por último, si se desea hacer uso de PgAdmin 3, se podrá instalar mediante el siguiente comando:

```
apt-get install pgadmin3
```

Una vez que Postgres ha sido instalado, se ha procedido a la configuración del usuario por defecto (“postgres”) y de la creación de un nuevo usuario “rs”) para la adición de extensiones y creación de tablas necesarias.

PostGIS

PostGIS es una extensión espacial para las Bases de Datos PostgreSQL. Esta extensión añade soporte para objetos geográficos permitiendo consultas geográficas sobre la Base de Datos. PostGIS se encuentra distribuido bajo licencia GNU General Public License.

Para instalar PostGIS sobre Ubuntu 16.04 se ha de teclear el siguiente comando:

```
apt-get install postgis
```

Se puede obtener más información en la página web oficial de PostGIS: <http://www.postgis.net/>.

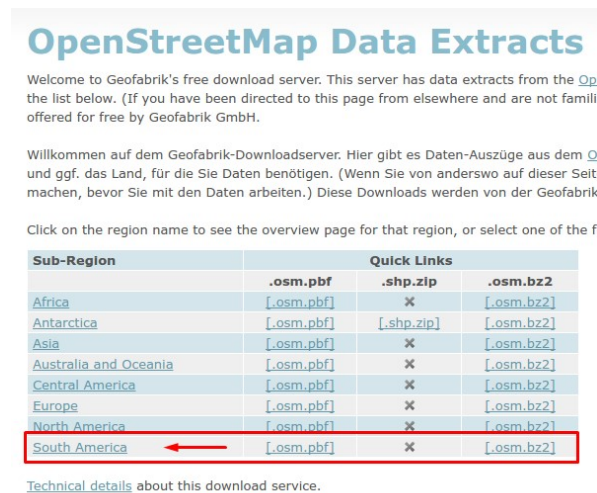
D.7. Manual del usuario

Osmosis

Osmosis permite extraer una zona concreta de un mapa con extensión osm o pbf. Para instalar esta herramienta sobre el Sistema Operativo Ubuntu se han de seguir los pasos indicados en su documentación oficial. El enlace a la misma es el siguiente: <http://wiki.openstreetmap.org/wiki/Osmosis>. En el caso de este trabajo, se ha descargado el mapa de la ciudad de Aracaju, capital del estado de Sergipe en Brasil. A continuación, se muestran los pasos necesarios para obtener la zona correspondiente a la ciudad mencionada partiendo de un mapa completo de Brasil.

Descarga del mapa de Brasil

Accediendo a la web de descargas de Geofabrik (enlace: <http://download.geofabrik.de/>) será posible elegir el país cuyo mapa que se desea descargar. Primero se seleccionará el continente sobre el que se encuentra el país a descargar (Figura D.1) y, después, se seleccionará el país propiamente dicho (Figura D.2).



OpenStreetMap Data Extracts

Welcome to Geofabrik's free download server. This server has data extracts from the [OpenStreetMap](#) project. The list below. (If you have been directed to this page from elsewhere and are not familiar with the data, please see the [FAQ](#) offered for free by Geofabrik GmbH.)

Willkommen auf dem Geofabrik-Downloadserver. Hier gibt es Daten-Auszüge aus dem [OpenStreetMap](#) und ggf. das Land, für die Sie Daten benötigen. (Wenn Sie von anderswo auf dieser Seite landen, bevor Sie mit den Daten arbeiten.) Diese Downloads werden von der Geofabrik angeboten.

Click on the region name to see the overview page for that region, or select one of the file formats.

Sub-Region	Quick Links		
	.osm.pbf	.shp.zip	.osm.bz2
Africa	[.osm.pbf]	✗	[.osm.bz2]
Antarctica	[.osm.pbf]	[.shp.zip]	[.osm.bz2]
Asia	[.osm.pbf]	✗	[.osm.bz2]
Australia and Oceania	[.osm.pbf]	✗	[.osm.bz2]
Central America	[.osm.pbf]	✗	[.osm.bz2]
Europe	[.osm.pbf]	✗	[.osm.bz2]
North America	[.osm.pbf]	✗	[.osm.bz2]
South America	[.osm.pbf]	✗	[.osm.bz2]

[Technical details](#) about this download service.

Figura D.1: Sección de continentes

Extracción de una ciudad

Para extraer una ciudad del plano que se acaba de descargar, se ha de abrir la consola del sistema y hacer uso del fichero de la aplicación Osmosis. Se necesitarán los siguientes argumentos (entre paréntesis las coordenadas que serán usadas en el caso de Aracaju):

APÉNDICE D. DOCUMENTACIÓN TÉCNICA DE PROGRAMACIÓN 7

• [.poly file](#) that describes the extent of this region.
 • [.osc.gz files](#) that contain all changes in this region, suitable e.g. for Osmosis updates
 • [raw directory index](#) allowing you to see and download older files

Sub Regions
 Click on the region name to see the overview page for that region, or select one of the file €

Sub Region	Quick Links		
	.osm.pbf	.shp.zip	.osm.bz2
Argentina	[.osm.pbf] (137 MB)	[.shp.zip]	[.osm.bz2]
Bolivia	[.osm.pbf] (57 MB)	[.shp.zip]	[.osm.bz2]
Brazil	[.osm.pbf] (451 MB)	[.shp.zip]	[.osm.bz2]
Chile	[.osm.pbf] (112 MB)	[.shp.zip]	[.osm.bz2]
Colombia	[.osm.pbf] (89 MB)	[.shp.zip]	[.osm.bz2]
Ecuador	[.osm.pbf] (54 MB)	✗	[.osm.bz2]
Paraguay	[.osm.pbf] (27.3 MB)	[.shp.zip]	[.osm.bz2]
Peru	[.osm.pbf] (66 MB)	[.shp.zip]	[.osm.bz2]
suriname	[.osm.pbf] (4.5 MB)	[.shp.zip]	[.osm.bz2]
Uruguay	[.osm.pbf] (20.5 MB)	[.shp.zip]	[.osm.bz2]

Figura D.2: Países mapeados de Sudamérica

- Nombre del fichero de entrada: se seleccionará el fichero de extensión osm sobre el que se obtendrá el recorte del mapa.
- Coordenada GPS que indique la parte **superior** del segmento de mapa que se va a recortar (-36.441).
- Coordenada GPS que indique la parte **izquierda** del segmento de mapa que se va a recortar (-11.0367).
- Coordenada GPS que indique la parte **inferior** del segmento de mapa que se va a recortar (-37.6971).
- Coordenada GPS que indique la parte **derecha** del segmento de mapa que se va a recortar (-10.537).
- Nombre del fichero de salida: se dará nombre al fichero de salida con extensión osm.

El comando a teclear incluirá los siguientes parámetros:

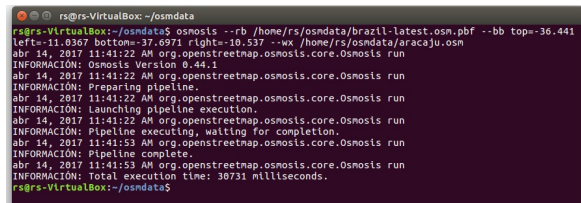
- rb: permite leer el contenido del fichero pbf descargado.
- bb: permite extraer los datos indicados en la caja definida por las coordenadas geográficas indicadas a continuación.
- top: coordenada geográfica superior.
- right: coordenada geográfica derecha.
- bottom: coordenada geográfica inferior.
- left: coordenada geográfica izquierda.

- wx: permite escribir los datos en un fichero osm.

Una vez descritos los parámetros y valores necesarios, se ejecuta el comando de la siguiente forma:

```
osmosis -rb /home/rs/osmdata/brazil-latest.osm.pbf -bb top=-36.441 left=-11.0367 bottom=-37.6971 right=-10.537 -wx /home/rs/osmdata/aracaju.osm
```

Una vez lanzado el comando, se verá la siguiente salida por consola:



```
rs@rs-VirtualBox:~/osmdata$ osmosis --rb /home/rs/osmdata/brazil-latest.osm.pbf --bb top=-36.441 left=-11.0367 bottom=-37.6971 right=-10.537 -wx /home/rs/osmdata/aracaju.osm
abr 14, 2017 11:41:22 AM org.openstreetmap.osmosis.core.Osmosis run
INFORMACIÓN: Osmosis Version 0.44.1
abr 14, 2017 11:41:22 AM org.openstreetmap.osmosis.core.Osmosis run
INFORMACIÓN: Preparing pipeline.
abr 14, 2017 11:41:22 AM org.openstreetmap.osmosis.core.Osmosis run
INFORMACIÓN: Launching pipeline execution.
abr 14, 2017 11:41:22 AM org.openstreetmap.osmosis.core.Osmosis run
INFORMACIÓN: Pipeline executing, waiting for completion.
abr 14, 2017 11:41:53 AM org.openstreetmap.osmosis.core.Osmosis run
INFORMACIÓN: Pipeline complete.
abr 14, 2017 11:41:53 AM org.openstreetmap.osmosis.core.Osmosis run
INFORMACIÓN: Total execution time: 30731 milliseconds.
rs@rs-VirtualBox:~/osmdata$
```

Figura D.3: Impresión en consola del resultado del comando.

Carga de mapas OSM en PostGIS

Una vez extraída la sección de mapa correspondiente a Aracaju se ha de cargar en PostGIS para poder hacer uso de toda su información posteriormente.

osm2pgsql

“osm2pgsql” es una herramienta que permitirá la carga de los datos obtenidos mediante Osmosis en PostGIS. Antes de comenzar, se ha de instalar siguiendo los pasos indicados en su página de referencia (enlace: <http://wiki.openstreetmap.org/wiki/Osm2pgsql>). Los parámetros que la herramienta necesita para su correcto funcionamiento son:

- s: habilita el modo “slim”, recomendado en la guía de referencia.
- U: indica el usuario de la Base de Datos. En este caso: “rs”.
- -d: indica la Base de Datos sobre la que se cargará el mapa. El valor debe ser “rutassemanticas”.
- nombre del fichero: inmediatamente después del nombre de la Base de Datos, se indica la ruta al fichero que contiene el mapa. Se seleccionará el fichero de extensión OSM obtenido en el paso previo: “aracaju.osm”.

El comando queda así:

```
osm2pgsql -s -U rs -d rutassemanicas /home/rs/osmdata/aracaju.osm
```

Una vez finalice, se crearán algunas tablas como “planet_osm_nodes”, “planet_osm_rels” o “planet_osm_ways”. Si ya existían, se añadirá el nuevo contenido.

QGIS

QGIS es una herramienta de información geográfica libre y de código abierto que permite conexiones con PostgreSQL y visualización de los datos obtenidos sobre un mapa físico. También permite la obtención de información geográfica desde Open Street Maps. De esta forma se pueden superponer rutas/trazas obtenidas desde las tablas de la Base de Datos con los mapas de Open Street Maps.

Para instalar QGIS desde ubuntu se han de seguir los pasos mostrados a continuación:

El primer paso es añadir la ruta a los ficheros fuente de QGIS en el fichero “sources.list”, para ello se puede abrir dicho fichero con el editor nano. Se encuentra en la siguiente ruta: “etc/apt/”. Las líneas a añadir son las siguientes y pueden ser incluidas en la parte final del fichero:

```
deb http://qgis.org/debian xenial main
deb-src http://qgis.org/debian xenial main
```

Ahora se puede proceder a la instalación de QGIS sobre el Sistema Operativo:

```
apt install qgis python-qgis qgis-plugin-grass
```

Una vez instalado, se puede configurar una conexión contra la Base de Datos instalada y configurada anteriormente. Para ello se ha de clicar en el icono de PostgreSQL ubicado en la columna central izquierda. Para ello se han de indicar valores como:

- **nombre:** valor que se dará al nombre de la conexión.
- **servidor:** servidor contra el que se realizará la conexión. En este caso será el servidor local.
- **de datos:** la Base de Datos de la que serán obtenidos los datos.
- **Autenticación:** nombre de usuario y contraseña del usuario de PostgreSQL usado para obtener la información.

Otro de los aspectos importantes a tener en cuenta es la facilidad que brinda al usuario para descargar contenido en formato indicando una zona

determinada mediante coordenadas geográficas. Esta característica ofrece el mismo resultado que el recorte del mapa de Brasil realizado con Osmosis. A continuación, se muestra un ejemplo con los mismos datos.

Descarga de contenido OSM con QGIS

Para comenzar, se ha de acceder a la sección “Vectorial”, después, situar el cursor sobre “OpenStreetMaps” y clicar sobre “Descargar datos” (Figura D.4).



Figura D.4: Descarga de datos desde QGIS

A continuación se indicarán las coordenadas sobre las que se desea obtener los datos. En la Figura D.5 se indican sus valores, son los mismos que los usados en Osmosis. Además, se ha de seleccionar un fichero destino. En este caso, se aprecia que es el mismo que el descargado mediante Osmosis.

Por último, se mostrará un mensaje de éxito si todo ha salido correctamente (Figura D.6).

Ahora se cuenta con el mismo fichero de extensión OSM y será posible su carga en PostGIS de la misma forma que se ha mostrado anteriormente.

Para obtener más información sobre QGIS se puede acceder al siguiente enlace: <http://www.qgis.org/es/docs/index.html>.

D.8. *Data cleaning*

En esta sección se describe el proceso de *data cleaning* o limpieza de datos aplicado a todas las rutas que se encuentran alojados en la Base de Datos.

Origen de datos

Antes de comenzar a describir el proceso de limpieza aplicado, se ha de tener en cuenta la procedencia de los datos usados para el presente Trabajo de Fin de Máster. El problema inicial al que hacer frente ha sido la falta de datos propios como punto de partida, por tanto, se ha tenido que realizar una búsqueda en distintos sitios de Internet para conseguir un conjunto de datos



Figura D.5: Valores introducidos para la búsqueda de datos en QGIS

lo más fiables posible y con la estructura adecuada. Se han establecido una serie de condiciones indispensables que el dataset ha de cumplir:

- **Coordenadas geográficas:** cada coordenada ha de contar con una latitud y longitud separadas en dos campos.
- **Marca de tiempo:** cada una de las coordenadas geográficas incluidas ha de contar con una marca de tiempo que permita localizar la coordenada en un espacio temporal concreto (un día, una semana, etc.).
- **Altitud:** para que la coordenada geográfica sea más precisa, debe contar con su altitud.
- **Diferenciación de rutas:** de igual manera que una coordenada geográfica se ha de poder diferenciar de la siguiente, cada ruta incluida en el dataset ha de poder ser diferenciada del resto. Es importante que cada una de las rutas se diferencien mediante un identificador único o se encuentren en ficheros separados.

Algunos de los sitios donde se pueden buscar este tipo de dataset incluyen páginas como el repositorio de Machine Learning de la Universidad Irvine de

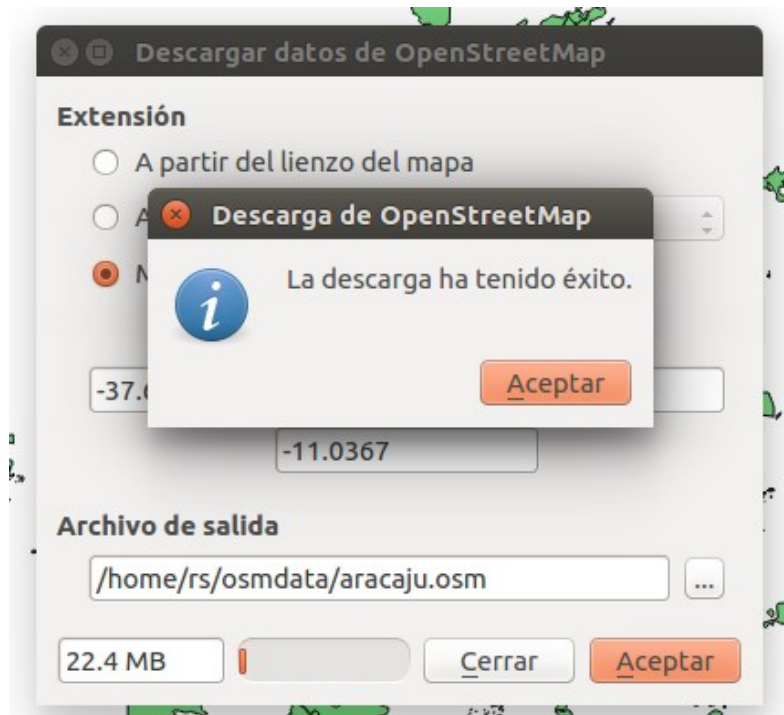


Figura D.6: Éxito en la descarga de los datos.

California (<http://archive.ics.uci.edu/ml/>). Donde se pueden encontrar los últimos dataset subidos o un buscador para poder encontrar el dataset buscado. En esta página se han encontrado y valorado dos conjuntos de datos.

Taxi Service Trajectory

El primer dataset encontrado ha sido el correspondiente a distintos trayectos en taxi sobre la ciudad de Oporto (Portugal). Este dataset se compone de un amplio fichero csv con información como el tipo de llamada realizada a la compañía de taxis, el taxi que ha respondido a la misma, la fecha y hora a la que se produjo la llamada y el conjunto de coordenadas geográficas tomadas durante el trayecto.

Aunque el conjunto de los datos parece interesante en un primer momento, no se cuenta con un indicador temporal por cada coordenada y los hace inviables para su posterior uso, siendo este el motivo por el que se ha descartado el uso de este conjunto de datos.

Este dataset puede ser descargado desde el siguiente enlace: <https://archive.ics.uci.edu/ml/datasets/Taxi+Service+Trajectory+-+Prediction+Challenge,+ECML+PKDD+2015>.

GPS Trajectories Data Set

El segundo de los conjuntos de datos analizados ha sido el correspondiente a las rutas realizadas por un conjunto de usuarios en la ciudad de Aracaju (Brasil) en taxi o autobús. Estos datos se encuentran divididos en dos ficheros csv. El primero de los ficheros cuenta con un identificador único de usuario, de esta forma se puede conocer qué usuario ha tomado cada ruta, un indicador de velocidad de la ruta (inicialmente no necesario), un campo de tiempo, uno de distancia recorrida, algunos de calificación de la ruta (rating) y por último, uno indicando la línea de autobús en la que se ha realizado el trayecto.

El segundo de los ficheros incluye cada una de las coordenadas geográficas correspondientes a cada ruta y su asociación al usuario que la ha tomado.

Aunque inicialmente fue uno de los dataset tenidos en cuenta, se ha desechado por no contener un número elevado de rutas que permitiese una mayor aleatoriedad en los datos.

Este dataset puede ser descargado desde el siguiente enlace: <https://archive.ics.uci.edu/ml/datasets/GPS+Trajectories>.

Geolife

El dataset seleccionado ha sido el denominado Geolife GPS Trajectories. Este dataset pertenece al repositorio que mantiene Microsoft. Este conjunto de datos viene acompañado por un documento en el que se indica cómo se han obtenido los datos, cuál es su estructura, etc. Se menciona que los datos han sido recolectados por unos 180 usuarios, en su mayoría en la ciudad de Pekín (China) pero también existen datos de viajes a otras ciudades o, incluso, a otros países.

En el dataset se incluye una carpeta por cada usuario y en cada carpeta se puede encontrar un fichero que toma su nombre de la marca de tiempo de la primera coordenada de cada ruta. Por tanto, puede que un usuario cuente con un único fichero, indicando así que solo ha realizado una ruta, o que cuente con múltiples ficheros en los que se almacenan distintas rutas.

Cada fichero contiene una línea por cada coordenada geográfica obtenida. Esta línea contiene distinta información: un par latitud longitud, que permite situar la coordenada geográfica sobre un plano, su altitud y su marca temporal (entre otras). Esta estructura hace que la diferenciación de las rutas sea sencilla ya que se encuentran separadas por usuarios y marca temporal en la que ha sido tomada. Debido a esto, este conjunto de datos es el que ha sido finalmente elegido y volcado a la Base de Datos.

Finalmente, comentar que otras universidades como la de St Andrews (<https://uk.crawdad.org/>) también prestan sus dataset aunque, en esta ocasión, no se ha revisado su repositorio en profundidad.

Proceso de limpieza

Aunque los datos provistos por el dataset de Microsoft son fiables, se ha implementado un proceso de limpieza de los mismos para poder afinar todavía más los resultados.

Se ha comentado en la sección anterior que este dataset cuenta con diferentes carpetas en las que se incluyen los ficheros correspondientes a las rutas efectuadas por cada usuario. Uno de los problemas encontrados ha sido la continuidad de una ruta dentro de un mismo fichero. Para explicar esto, primero hay que saber, que cada coordenada tomada dentro de una misma ruta suele diferenciarse por una cantidad de tiempo (ya sean segundos o minutos) o espacial (metros) constante.

Poniendo un ejemplo, teniendo las siguientes coordenadas geográficas extraídas de un mismo fichero:

```
40.004273,116.322334,87,2009-05-17,22:22:16
40.004285,116.322334,87,2009-05-17,22:22:21
40.004323,116.32235,8,2009-05-17,22:22:26
40.004387,116.322366,91,2009-05-17,22:22:31
40.004371,116.322317,91,2009-05-17,22:22:36
40.00441,116.322294,92,2009-05-17,22:22:41
40.004435,116.322279,102,2009-05-17,22:22:46
40.004451,116.322241,120,2009-05-17,22:22:51
40.004214,116.322215,110,2009-05-17,22:35:46
40.004171,116.322241,108,2009-05-17,22:35:51
40.004125,116.322269,108,2009-05-17,22:35:56
40.004107,116.322301,105,2009-05-17,22:36:01
40.004132,116.32232,103,2009-05-17,22:36:06
40.004154,116.32232,103,2009-05-17,22:36:11
40.004192,116.322331,105,2009-05-17,22:36:16
40.004248,116.32235,109,2009-05-17,22:36:21
```

Se puede observar que cada punto geográfico es recolectado cada cinco segundos, aproximadamente. Teniendo esto en cuenta, se aprecia un salto importante entre la octava y la novena coordenada. En concreto, hay un salto de 12 minutos y 55 segundos. Esta diferencia contrasta con los cinco segundos transcurridos entre cada toma de tiempo. Por tanto, se puede pensar que ha ocurrido algún problema en la toma de datos o que se tratan de rutas diferentes aunque se encuentren en el mismo fichero. En este caso, como se trata de un salto importante sobre las mediciones tomadas, se realiza un corte en la ruta y se tratará como dos rutas diferentes.

Este pequeño ejemplo muestra el problema que puede ocurrir con el resto de las rutas disponibles en los ficheros del dataset. Se ha implementado un algoritmo para poder separar las rutas que, por tiempo o diferencia geográfica,

parezcan pertenecer a dos o más rutas diferentes. Este algoritmo tiene en cuenta la mediana, tanto de la diferencia en términos de tiempo como de espacio.

El algoritmo comienza obteniendo todas las rutas de la Base de Datos y generando un mapa cuya clave es el identificador del usuario y valores las rutas del mismo. Después se recorre cada una de las rutas y se calcula una primera mediana (el usuario puede indicar si quiere la mediana sobre las marcas temporales o sobre las coordenadas geográficas). En base a esta primera mediana se empieza a analizar la ruta, siendo el siguiente paso, el cálculo de la diferencia temporal o geográfica entre cada dos de las coordenadas. Una vez obtenida esta lista de diferencias se comparará con la mediana obtenida anteriormente. En el caso de que no existan diferencias significativas, la ruta se mantendrá tal cual en la Base de Datos. Si, por el contrario, una posición de la lista contiene una diferencia apreciable, la ruta quedará separada, inicialmente, en dos tramos. El primer tramo se mantendrá en la Base de Datos pero el segundo tramo será re analizado calculando una nueva mediana y repitiendo el proceso descrito. Al final se podrán obtener varias rutas a partir de la inicial actualizando la Base de Datos con las nuevas rutas obtenidas.

Este algoritmo también es aplicable a ficheros de texto, que una vez procesados, generarán nuevas rutas en las tablas de la Base de Datos.

Apéndice E

Documentación de usuario

E.1. Introducción

E.2. Requisitos de usuarios

E.3. Instalación

A continuación se mostrarán los pasos a dar para instalar el software necesario.