1 Planteamiento del problema

2 Datos y estructura de los datos suministrados

Se nos suministran dos bases de datos correspondientes a dos ciudades brasileñas distintas, Salvador de Bahía y Río de Janeiro. En cada de una de ellas encontramos posiciones de distintos sujetos estudiados identificados a través de un código. Cada base de datos contiene una tabla llamada *posicionesgps* en la que encontramos un registro por cada posición tomada por cada sujeto entre los días 2015-02-17 08:00:05 y 2015-03-04 08:18:05.

La estructura de los registros es la siguiente:

Parámetros	
Id	Identificador numérico de la posición (clave primaria)
IdServidor	Identificador numérico del servidor que realiza la inserción (PK)
Recurso	Nombre del recurso (tetra:1234567)
Latitud	Real que representa la latitud GPS
Longitud	Real que representa la longitud GPS
Velocidad	Entero que representa la velocidad instantánea
Orientación	Entero que representa la orientación respecto al norte en grados
Cobertura	Booleano que indica si hay cobertura
Error	Booleano que nos indica si ha habido algún error en la toma de la posición

En base de datos, el tipo de datos guardado es:

```
mysql> explain posicionesgps;
```

Field	Type	Null	Key	Default
id	bigint(10)	NO	PRI	0
idServidor	int(10) unsigned	NO	PRI	0
recurso	varchar(100)	YES	MUL	NULL
latitud	double	YES		NULL
longitud	double	YES		NULL
velocidad	tinyint(10) unsigned	YES		NULL
orientacion	smallint(10) unsigned	YES		NULL
cobertura	tinyint(10) unsigned	YES		NULL
error	tinyint(10) unsigned	YES		NULL
antigua	tinyint(10) unsigned	YES		0
fecha_timestamp	timestamp	NO	MUL	CURRENT_TIMESTAMP
automático	tiniyint(10) unsigned	NO	MUL	0

Para este estudio se ha trabajado sólo con los siguientes datos,

1. Id
2. Recurso
3. Latitud
4. Longitud
5. Velocidad
6. Fecha

2.1 Análisis de los datos

2.2 Espacio en disco

Con la cantidad de posiciones suministradas, cuánto ocupa cada posición en disco, para hacernos una idea de cuántas posiciones sería posible acumular en función de la frecuencia de éstas sobre un espacio en disco finito.

En nuestra base de datos llamada **Río de Janeiro** contamos con **6928467** posiciones y en **Salvador de Bahía** contamos con **4599974** posiciones.

El tamaño en disco de nuestras bases de datos es,

```
mysql> SELECT table_schema as 'Database',
            table_name AS 'Table',
            round(((data_length + index_length) / 1024 / 1024), 2)
            FROM information_schema.TABLES
            ORDER BY (data_length + index_length) DESC;
```

Database	Table	Size in MB	Size in KB
rio	posicionesgps	1205.64	120564000
bahia	posicionesgps	961.42	96142000

Lo cual nos da una idea de cuánto puede ocupar una toma de posición en disco.

El total de posiciones almacenadas en río es de 6928467 luego podemos estimar el tamaño de una posición en,

$$\frac{120564000}{6928467} = 17.4012519653KB$$

El total de posiciones almacenadas en bahía es de 4599974, luego

$$\frac{96142000}{4599974} = 20.9005529162KB$$

Podemos aproximar el tamaño de una posición por unos 19 KB.

Supongamos que una consola tiene unos 1GB de almacenamiento. Podemos almacenar unas 52631 posiciones en estos 30GB.

Los datos han sido recogidos entre las fechas 2015-02-17 08:00:05 y 2015-03-04 08:18:05, lo que hace una diferencia de 360 horas.

Tenemos 5014 distintos tipos de sujetos a estudiar en la base de datos de río:

```
mysql> USE rio;
mysql> SELECT COUNT(distinct(recurso))
            FROM posicionesgps;
```

count(distinct(recurso))
5014

Lo que nos da una frecuencia de toma de :

$$\frac{6928467}{5014 \cdot 360} = 3.83$$

posiciones a la hora.

Si aumentáramos esta frecuencia a una posición cada 30 segundos, conseguri̊iamos una frecuencia de 120 posiciones a la hora, luego un único sujeto, en una jornada laboral de 8 horas, ocuparía en espacio de 19.2 MB.

3 Nociones de distancia y vecindario

Con el fin de realizar los algoritmos de consolidación, hemos realizado un estudio acerca de distintos tipos de vecindarios a utilizar para los algoritmos de consolidación propios y los algoritmos de *clustering* utilizados en WEKA.

3.1 Vecindario simple

Utilizando la distancia euclídea, definimos un vecindario como aquel conjunto de puntos que se encuentran a una distancia euclídea menor que ϵ con respecto su centro p_0, es decir,

$$d_E(p_0, p) = \sqrt{lat_p - lat_{p_0})^2 + (long_p - long_{p_0})^2} < \epsilon$$

donde p es un punto con latitud lat_p y longitud $long_p$.

3.2 Vecindario involucrando el módulo de la velocidad

En el momento que se toma la posición p_0, aparte de la latidud y su longitud, se toma la velocidad instantánea del sujeto. Podemos considerar en este caso que, dado que nuestro sujeto se encuentra a mayor velocidad, puntos más alejados de lo que consideraríamos en el primer caso (fuera de nuestro vecindario simple), podrían estar dentro de nuestro nuevo radio, que dependería de la velocidad instantánea. Así, definimos nuestro nuevo vecindario:

$$d_E(p_0, p) = \sqrt{lat_p - lat_{p_0})^2 + (long_p - long_{p_0})^2} < \epsilon \cdot vel_{p_0}$$

donde vel_{p_0} es la velocidad instantánea de nuestro punto centro.

3.3 Vecindario involucrando el módulo de la velocidad y la orientación

Igual que contamos con la velocidad instantánea del sujeto, contamos también con el dato de la orientación respecto al norte de nuestro sujeto muestreado. Esta medida está tomada en grados sexagesimales en el sentido de las agujas del reloj respecto al norte.

Gracias a este dato, podemos calcular el vector dirección que contiene la información de la orientación de nuestro sujeto y como también conocemos el módulo de la velocidad, obtener el vector velocidad.

Nuestra componente x que identificaremos con el eje del vector dirección será el coseno de nuestra orientación,

$$cos(or_{p_0})$$

Y nuestra componente y del vector dirección será el seno de nuestra orientación,

$$sin(or_{p_0})$$

3.4 Vecindad t0-alcanzable

Si fijamos un intervalo de tiempo t_0, podemos definir una vecindad t_0-alcanzable como aquellos puntos que nuestro sujeto puede alcanzar en un tiempo t_0. Un sujeto que se desplace a velocidad reducida, tendrá una vecindad t_0-alcanzable más reducido que otro que se desplace a una velo$vel_{p_0} \cdot t_0$.

$$d_E(p_0, p) = \sqrt{lat_p - lat_{p_0})^2 + (long_p - long_{p_0})^2} < vel_{p_0} \cdot t_0$$

Éste es un caso concreto del vecindario involucrando la velocidad.

3.5 Vecindario involucrando el tiempo

Las posiciones de nuestros sujetos vienen muestreadas además con el instante en el que fueron tomadas. Podemos considerar que el tiempo entre tomas también es una distancia y definir un vecindario. Definimos esta distancia temporal como la resta de ambos instantes, y el vecindario como:

$$d_T(p_0, p) = time_p - time_{p_0} < \delta$$