CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR DE ENSENADA, BAJA CALIFORNIA

PROGRAMA DE POSGRADO EN CIENCIAS EN CIENCIAS DE LA COMPUTACIÓN

Predicción espacio-temporal de la movilidad del usuario

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Doctor en Ciencias

Presenta:

Jorge Álvarez Lozano

Ensenada, Baja California, México

2015

Tesis defendida por

Jorge Álvarez Lozano

y aprobada por el siguiente comité

Dr. José Antonio García Macías
Director del Comité

Dr. Jesús Favela Vara
Miembro del Comité

Dr. Edgar Leonel Chávez González
Miembro del Comité

Dr. José Alberto Fernández Zepeda
Miembro del Comité

Dra. Mónica Elizabeth Tentori Espinosa
Miembro del Comité

Dr. Luis Enrique Palafox Maestre
Miembro del Comité

Dra. Ana Isabel Martínez García
Coordinador del Programa de Posgrado en Ciencias de la Computación

Dr. Jesús Favela Vara
Director de Estudios de Posgrado

Febrero, 2015

Resumen de la tesis que presenta Jorge Álvarez Lozano como requisito parcial para la obtención del grado de Doctor en Ciencias en Ciencias de la Computación.

Predicción espacio-temporal de la movilidad del usuario

Resumen elaborado por:

Jorge Álvarez Lozano

Predecir la ubicación del usuario de manera precisa es importante en diferentes dominios y áreas de investigación. Actualmente, con la proliferación de dispositivos móviles y los diversos sensores incorporados en éstos, es posible obtener una gran cantidad de datos asociados a la movilidad del usuario. A la fecha, varias investigaciones han encontrado que los usuarios exhiben un alto grado de repetición al visitar ciertos lugares durante sus actividades cotidianas. Así, algunos trabajos han aprovechado la regularidad de los movimientos registrados del usuario para predecir el próximo o próximos lugares donde estará. La predicción espacial se enfoca en predecir las próximas ubicaciones del usuario, mientras que la predicción espacio-temporal toma también en cuenta el tiempo, para predecir las próximas ubicaciones y el tiempo que estará en dichas ubicaciones. Actualmente, los trabajos reportados en la literatura realizan la predicción espacio-temporal en el corto y largo plazo (e.g., 10 minutos y 1 año, respectivamente). Sin embargo, no se encontraron trabajos que resulten adecuados para predicciones espacio-temporales en el mediano plazo (e.g., unas cuantas horas adelante).

El modelo de predicción aquí propuesto toma como base los puntos de interés, los cuales son aquellos lugares que los usuarios visitan de manera regular y en los cuales pasan al menos un tiempo determinado. Se modela la movilidad de los usuarios entre los puntos de interés como una cadena de Markov, puesto que la ubicación actual del usuario determina con cierta probabilidad la siguiente ubicación de éste. Además, al considerar la relación de la propiedad Markoviana entre los puntos de interés y los tiempos de estadía del usuario en dichos puntos, la movilidad del usuario se modela como un modelo oculto de Markov. Así, es posible predecir los próximos puntos de interés que el usuario visitará y los tiempos de arribo a éstos.

Para evaluar la eficiencia del modelo de predicción se utilizaron dos conjuntos de datos públicos, los cuales son utilizados en los trabajos relacionados. Los resultados experimentales mostraron que el modelo de predicción espacio-temporal resulta ser eficiente en periodos de hasta 7 horas. Para un periodo de predicción de 30 minutos se obtuvo hasta un 81.75 % de precisión y al considerar un periodo de 7 horas, se obtuvo una precisión de 66.25 %.

Palabras Clave: **Predicción espacio-temporal, puntos de interés, análisis de la movilidad.**

Abstract of the thesis presented by Jorge Álvarez Lozano as a partial requirement to obtain the Doctor in Sciences degree in Computer Science.

Spatio-temporal prediction of user mobility

Abstract by:

Jorge Álvarez Lozano

Predicting user location accurately is important for different domains and research areas. Nowadays, with the proliferation of mobile devices and the various sensors they have, it is possible to obtain a great amount of contextual data about user mobility. Up to now, several studies have found that users exhibit a high degree of repetition by visiting certain places during their daily activities. Thus, some works have taken advantage of the regularity of past movements to forecast the next locations of users. Spatial prediction focuses on predicting the next location of the user, while spatio-temporal prediction also considers time and as a result not only predicts the next location but also the time when the user will be there. Currently, works in the literature make spatio-temporal predictions for short and long-term (e.g., 10 minutes and 1 year, respectively). However, there are no works that are suitable for spatio-temporal predictions in the medium term (e.g., a few hours later).

The prediction model proposed here takes as reference the points of interest, which are those places that users visit regularly and where they spend at least a certain amount of time. User mobility is modelled between the points of interest as a Markov chain, since the current location of the user determines with some probability his following location. Moreover, considering the relationship of Markovian property between points of interest and user staying times at these points, user mobility is modelled as a hidden Markov model. Consequently, it is possible to predict the next points of interest where the user will be and the arrival times to these places.

To evaluate the efficiency of the prediction model, two public datasets were used. The experimental results showed that the spatio-temporal prediction model is efficient for periods of up to 7 hours. For a prediction period of 30 minutes the model presented an accuracy of up to 81.75 % and for a prediction period of 7 hours, the accuracy was of up to 66.25 %.

Keywords: **Spatio-temporal prediction, points of interest, user mobility analysis**

Dedicatoria

A mis padres:

Pascual Álvarez Sánchez

Otilia Lozano Hernández

A mi esposa:

Mary Salazar Soria

Agradecimientos

Quiero agradecer a mi asesor, el Dr. José Antonio García Macías, por su apoyo incondicional, sus conocimientos, sus críticas, y comentarios durante el desarrollo de esta investigación. Le estaré agradecido por su confianza, paciencia y el haberme aceptado como estudiante de doctorado.

De igual manera, agradezco de manera especial al Dr. Edgar Leonel Chávez González, quien fue parte importante durante el desarrollo de este trabajo. Sus comentarios, ideas, y sugerencias fueron de suma importancia.

A mis amigos René, Raymundo, Valeria, Jorge Mario, Ariel, Daniel, Amado, entre otros, por sus comentarios, y amistad. Gracias por esos grandes momentos.

También quisiera agradecer a los miembros de mi comité de tesis: Dr. Jesús Favela Vara, Dr. José Alberto Fernández Zepada, y Dra. Mónica E. Tentori Espinosa, por su tiempo, consejos, correcciones e invaluables comentarios que me ayudaron a concluir esta investigación.

Mi agradecimiento al Centro de Investigación Científica y de Educación Superior de la ciudad de Ensenada, Baja California. Muchas gracias al personal del departamento de Ciencias de la Computación, sobre todo a Caro Amador Tavares y Lydia Salazar Ochoa, por el apoyo moral y logístico recibido durante el trabajo de investigación.

Al Centro de Investigación Científica y de Educación Superior de Ensenada.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de maestría.

Tabla de contenido

Página

Tabla de contenido (continuación)

Tabla de contenido (continuación)

Tabla de contenido (continuación)

Tabla de contenido (continuación)

Lista de figuras

Figura Página

Lista de figuras (continuación)

Lista de figuras (continuación)

Lista de tablas

Capítulo 1. Introducción

La ubicación es un aspecto central en vida de las personas; los lugares que visitan reflejan sus gustos, estilo de vida, relaciones sociales y determina en cierta medida las actividades que realizan; las actividades que se realizan en el hogar difieren de aquellas que realizan en la oficina. Del mismo modo, las personas con las que se interactúa en el lugar de trabajo son diferentes de aquellas con la que se tiene contacto en el gimnasio, o en el centro comercial.

Es entonces que al conocer la ubicación de las personas resulta factible realizar diversas inferencias acerca de sus actividades, entorno social, y otros aspectos. Debido a ello, la ubicación es un factor importante en los sistemas conscientes del contexto, el cual se define como cualquier información que puede ser utilizada para caracterizar la situación de una entidad, donde una entidad es una persona, lugar, u objeto que es considerado relevante para la interacción entre un usuario y una aplicación, incluyendo al usuario y a la aplicación (Abowd *et al.* (1999)). Por lo tanto, un sistema consciente del contexto es aquel cuya funcionalidad o comportamiento se encuentra en función del entorno de las personas.

El uso del contexto permite la creación de sistemas reactivos y proactivos. Un sistema reactivo es aquel que reacciona al contexto actual del usuario y ofrece un servicio o información (i.e., estado del clima de la ubicación del usuario). Sin embargo, un sistema no sólo puede reaccionar al contexto actual del usuario (reactivo), también puede anticiparse al contexto futuro del usuario y actuar en consecuencia (i.e., información del estado del clima de la ubicación de la persona en 12 horas). Este tipo de sistemas se conocen como proactivos (Satyanarayanan (2001)); la implementación de éstos es más compleja que la de los sistemas reactivos ya que se requiere predecir o inferir el contexto futuro de las personas.

A la fecha existen diversos sistemas que con base en la ubicación actual del usuario realizan alguna acción. Esto es, sistemas basados en localización del tipo reactivos. Con el propósito de contar con sistemas proactivos, los servicios basados en localización deben estimar la ubicación futura del usuario. Como menciona Bo Begole (Begole (2010)), la idea es pasar de la *consciencia del contexto* a la *inteligencia contextual*, es

Figura 1: La ubicación del usuario determina varios aspectos de su vida.

decir, combinar la información del contexto actual y la información histórica de la persona para predecir situaciones futuras que conlleven a ofrecer información o un servicio acorde a dicha predicción.

La capacidad para predecir la ubicación del usuario (*dónde*) es de interés y de beneficio para diversas áreas de investigación y dominios de aplicación como planeación urbana (Yuan *et al.* (2010, 2013)), el cuidado de la salud (Chang *et al.* (2009)), redes de computadoras (Cheng *et al.* (2003)), sistemas de recordatorios (Ashbrook (2002)), sistemas de recomendación (Aalto *et al.* (2004); Marmasse y Schmandt (2000)), y atención en situaciones de desastre (Gao *et al.* (2011a,b)), sólo por mencionar algunos.

Así, también, tener conocimiento preciso de las próximas ubicaciones del usuario permite la creación de diversas aplicaciones y sistemas novedosos; un sistema puede proveer información importante relacionada con el lugar que el usuario va a visitar, así como publicidad, recomendaciones, y noticias, evitando con ello la entrega de información que

no es relevante a dicho lugar, o bien la entrega de información a usuarios que no van a visitar estos lugares y, por lo tanto, no tienen interés en ella.

Además, el beneficio se incrementa al tener conocimiento de *cuándo* (*hora*) el usuario arribará a la próxima o próximas ubicaciones. De esta manera, un sistema puede proporcionar información específica asociada al lugar que el usuario visitará en un tiempo dado. Por ejemplo, una promoción válida en un restaurante de las 20:00 a las 22:00 horas.

En la actualidad resulta factible realizar la predicción de la ubicación de las personas debido a las diferentes tecnologías que permiten conocerla, en especial las tecnologías que incorporan los dispositivos móviles.

1.1. Sensado móvil

Debido a la disminución de costos en cuanto a hardware, producción y la utilidad que proveen, los dispositivos móviles han tenido un auge significativo en los últimos años. Desde la aparición del iPhone en el año 2007, anualmente se ha registrado un incremento aproximado del 40 % [1] en el número de dispositivos (i.e., teléfonos, tabletas). En el año 2012 había 6500 millones de dispositivos móviles en el planeta [2], en el año 2013 habían aproximadamente 6800 millones de dispositivos móviles; prevalecían los teléfonos celulares y las tabletas (Figura 2). Se estima que para el año 2018 el número de dispositivos móviles alcance la cifra de 10000 millones.

Actualmente, los dispositivos móviles son parte imprescindible en las actividades cotidianas de los usuarios de estos dispositivos. Ya sea para cuestiones laborales, educativas, de investigación u ocio, éstos se encuentran al alcance de los usuarios (Figura 3(a)). Los dispositivos móviles han pasado de ser un instrumento de comunicación a convertirse en una tecnología ubicua. Gradualmente los dispositivos móviles han incorporado nuevas y mejor capacidad en cuestión de procesamiento, energía, comunicación, y sensado del ambiente. Debido a ello, ahora estos dispositivos se denominan *smartphones* o *teléfonos inteligentes*. A la fecha, del total de la población mundial, el 56 % posee un teléfono inteligente, este porcentaje aumentó considerablemente ya que en 2011 sólo era el 35 % y en

[1] http://readwrite.com/2013/05/13/mobile-is-taking-over-the-world#awesm= oCzvaODHRBKQJk

[2] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper _c11-520862.html

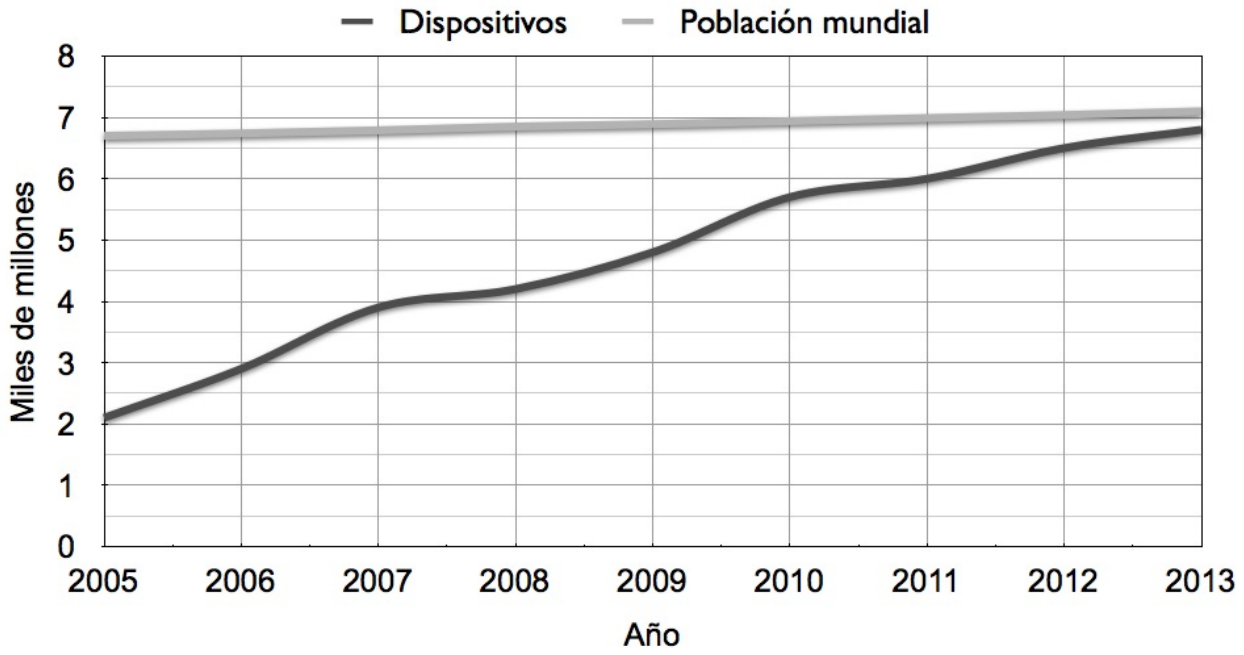


Figura 2: Proliferación de los dispositivos móviles en el periodo 2005-2013.

el 2012 llegó a 46 % (Figura 3(b))[3] [4].

Un aspecto importante de los teléfonos inteligentes, es la incorporación de diversos sensores, los cuales permiten obtener datos relacionados a localización, iluminación, audio, movimientos, orientación, proximidad, sólo por citar algunos (Tabla 1). Por lo tanto, al considerar la ubicuidad de los dispositivos móviles y las capacidades que éstos integran, es posible obtener y almacenar una gran cantidad de datos que caracterizan el ambiente o al usuario mismo. Esto es, los dispositivos móviles se convierten en una fuente importante de datos contextuales.

1.2. Sistemas basados en localización

La proliferación de los dispositivos móviles equipados con tecnología GPS permitió hacer realidad el *cómputo consciente de la ubicación* y, por consiguiente, los sistemas basados en localización. Aun en los casos en donde los dispositivos móviles no cuentan con tecnología GPS, la ubicación se calcula a través de las técnicas de triangulación o trilateración al utilizar información de las estaciones base detectadas por el dispositivo móvil

[3] http://www.digitalbuzzblog.com/infographic-2013-mobile-growth-statistics/ Cabe destacar que estos porcentajes se obtienen al considerar el número de SIM registradas, no el número de personas utilizando un dispositivo móvil; algunas personas pueden tener varias SIM activas.

[4] http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/a

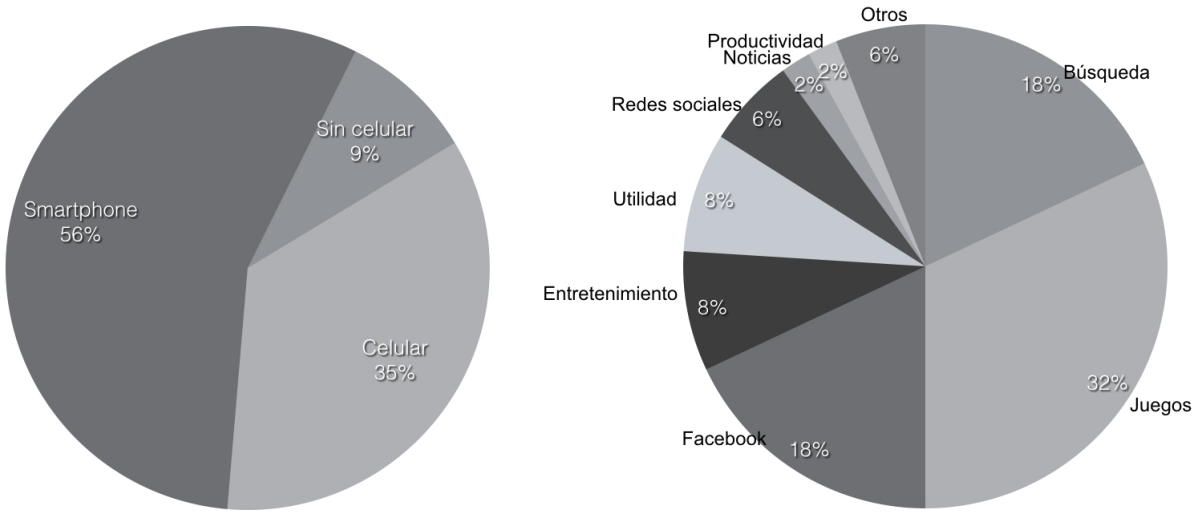


Figura 3: Dispositivos móviles en la actualidad; a) Principales usos de los dispositivos móviles a nivel mundial b) Porcentaje de la población mundial que cuenta con telefonía celular

Tabla 1: Interfaces de comunicación y sensores: fuentes de datos contextuales.

Sensor/Interfaz	Datos relacionados a
GPS	Localización
GSM	Localización
Wi-Fi	Localización
Acelerómetro	Movimientos
Micrófono	Audio
Cámara	Imágenes
Brújula	Orientación
Bluetooth	Proximidad
NFC	Proximidad
Luz	Iluminación

en las redes GSM (Lin *et al.* (2004)), o bien al utilizar la información de los puntos de acceso 802.11 que se encuentran en la periferia del dispositivo (LaMarca *et al.* (2005)), éstos sólo por citar algunos métodos para identificar la ubicación del usuario.

Al considerar las diversas interfaces de comunicación de los dispositivos móviles, la ubicación se puede transmitir a un servicio en la red, o bien, procesarse de manera local en el dispositivo para ofrecer información o algún servicio relevante al usuario. De esta manera, al combinar la funcionalidad de las interfaces de comunicación y los sensores permite la creación de diversos sistemas basados en localización.

En la actualidad existen sistemas basados en localización que ofrecen servicios de emergencia (ambulancia, policía, bomberos), otros que permiten encontrar servicios o lugares de interés que se encuentran en la cercanía. Así, también, hay servicios que permiten al usuario trasladarse de un lugar a otro, compartir la ubicación del usuario y/o las actividades asociadas a ésta, conocer la ubicación de sus amistades, llevar un registro de sus actividades físicas, entre otros. De igual manera, mediante estos servicios los usuarios reciben publicidad, o algún cobro asociado a su ubicación. Resulta relevante mencionar que actualmente 1400 millones de personas hacen uso de sistemas basados en localización [5].

En la Tabla 2 se presenta de manera general un listado de las categorías de los sistemas basados en localización y algunos ejemplos de cada una de ellas.

1.2.1. Sistemas reactivos

La mayoría de los sistemas basados en localización son del tipo *reactivo*. Es decir, estos sistemas ofrecen información al usuario o bien realizan alguna acción cuando éste la solicita de manera explícita. Tanto la información que estos sistemas presentan como las acciones que realizan se encuentran en función del contexto actual del usuario.

[5] http://www.mobileapp-development.com/blog/mobile-trends-in-2013-?-how-to-be-part-of-the-future.aspx

Tabla 2: Categorías de los sistemas basados en localización.

Categoría	Aplicaciones
Publicidad	Placecast, Eventful
Servicios de Emergencia	Ushahidi Platform, PDX Reporter, E911
Sistemas de información	Waze,Yelp, Gas Buddy, Yahoo Local, Google Places
Navegación	Google Maps
Redes sociales	Foursquare, Gowalla, Facebook, Google +
Juegos	GeoSocial, GeoHunters, CitySecret
Deportes	Nike+, Nokia Sports Tracker, Endomondo
Cobro de servicios	ZoneWise, O2 Genion Home-Zone Service
Realidad aumentada	Layar, Wikitude,
Rastreo	PDX Bus, UPS

1.2.2. Sistemas proactivos

De manera general, los dispositivos se encuentran en un estado de *reposo* en espera de que los usuarios soliciten algún servicio. Sin embargo, también existen aplicaciones que integran mecanismos que permiten anticiparse al contexto del usuario y ofrecer información y/o realizar acciones sin que el usuario las solicite de manera explícita.

La aplicación más famosa en este ámbito es Google Now [6], la cual estima cuándo un usuario realizará ciertas acciones y así ofrecer ayuda oportuna. Para ello, Google Now utiliza información del correo electrónico, calendario y las búsquedas que el usuario ha realizado en la Web a fin de conocer las próximas acciones de éste. Las aplicaciones Osito [7] y Donna [8] utilizan las mismas fuentes de información que Google Now, éstas presentan información asociada a las próximas acciones del usuario (i.e., alertan al usuario cuando éste se encuentra retrasado para llegar a una reunión).

1.3. Motivación y planteamiento del problema

Las aplicaciones que se mencionan son de utilidad en diversos dominios de aplicación, sin embargo, la funcionalidad o información que brindan estas aplicaciones puede ser mayor, ya que la *proactividad* de estas aplicaciones o sistemas depende únicamente de la información que el usuario provee de manera explícita. Debido a ello, resulta viable utilizar

[6] http://www.google.com/landing/now/
[7] http://www.getosito.com
[8] http://don.na

los datos históricos y del contexto actual del usuario con el fin de predecir el contexto futuro del mismo y actuar en consecuencia, lo que permite ofrecer mejor información y/o servicios al usuario. Específicamente, resulta de especial interés predecir la movilidad del usuario, es decir, conocer los próximos lugares o ubicaciones en donde estará el usuario en un periodo de tiempo dado ([T, T+ Δ T]).

Figura 4: Fuentes de información contextual

Realizar la predicción de la movilidad del usuario resulta factible, ya que al considerar las diversas interfaces de comunicación y sensores incorporados en los dispositivos móviles es posible recolectar tanto datos continuos como discretos de la movilidad del usuario. También, resulta viable obtener datos de localización a través de las aplicaciones basadas en localización que los usuarios utilizan en sus dispositivos móviles. Por lo tanto, como se muestra en la Figura 4, existen diversas fuentes de datos contextuales a partir de las cuales es posible recolectar datos de localización y, así, analizar estos datos para identificar patrones de movilidad y, posteriormente definir un modelo de predicción de la movilidad del usuario.

A la fecha, se han realizado diversos estudios para entender, caracterizar e identificar patrones de movilidad. Las investigaciones han demostrado que los usuarios poseen un alto grado de regularidad al visitar ciertos lugares durante sus actividades cotidianas (Gonzalez *et al.* (2008); Eagle y Pentland (2006); Zheng *et al.* (2008); Furletti *et al.* (2013); Calabrese *et al.* (2013)). Esta regularidad se ha explotado para definir modelos de predicción espacial y estimar la próxima o próximas ubicaciones del usuario.

Este enfoque resulta útil en escenarios donde únicamente es de interés el aspecto espacial, sin embargo, un escenario más complejo requiere también conocer *cuándo* el usuario estará en la próxima o próximas ubicaciones. Para ello, se requiere considerar no sólo el aspecto espacial, sino también el temporal, y de esta manera realizar la predicción espacio-temporal de la movilidad del usuario.

Con respecto a la predicción espacio-temporal, a la fecha se han propuesto diferentes modelos de predicción. Algunos trabajos representativos son presentados por Scellato *et al.* (2011) y Sadilek y Krumm (2012b). Estos enfoques tienen como objetivo predecir la movilidad del usuario en el corto y largo plazo, respectivamente. Esto es, los modelos de predicción pueden estimar dónde estará el usuario en los próximos minutos, o bien, en un par de años.

Al conocer el enfoque de estos trabajos, la incógnita que surge es ¿qué sucede si se desea conocer dónde estará el usuario en las próximas horas?, y además, si se desea tener conocimiento de los tiempos de arribo a los lugares que visitará en el periodo de predicción. Esta interrogante es importante ya que en diversos escenarios se requiere conocer la ubicación del usuario en el mediano plazo (i.e., cinco horas), por lo que la funcionalidad de los modelos de predicción actuales no es de utilidad.

Para enfatizar la necesidad de un modelo que permita predecir de manera eficiente la ubicación del usuario en el mediano plazo, se describen los siguientes escenarios de aplicación:

- Redes oportunistas. Un escenario interesante que requiere contar con la predicción espacio-temporal es la comunicación oportunista de datos entre regiones desconectadas. Un problema asociado a las redes oportunistas es el retardo en la entrega de los datos, y la tasa de entrega de éstos. Al no contar con información de ruteo y/o estructura de la red, no es posible asegurar la entrega de los datos en un tiempo determinado. Por lo tanto, al considerar la predicción espacio-temporal de la movilidad de los usuarios, los protocolos de las redes oportunistas pueden tomar ventaja de las predicciones de los usuarios en el mediano plazo para elegir al o a los mejores portadores o transportadores de los datos. De esta manera se maximiza la proba-

bilidad de entrega y se disminuye el tiempo de retardo para la entrega de los datos; éstos representan los mayores retos en las redes DTN y de las redes oportunistas (Pirozmand *et al.* (2014); Cardei *et al.* (2008); Nguyen y Giordano (2012)).

- Reserva de recursos. La reservación de recursos es útil en las redes celulares o bien redes de comunicación. Al conocer la cantidad de personas que habrá en un tiempo determinado, es posible asignar una mayor o menor cantidad de recursos a un lugar o área geográfica (Prasad y Agrawal (2010); Lee *et al.* (2003a)). De igual manera, la reservación de recursos es factible para un sistema proactivo de índole social. Actualmente, diversos establecimientos permiten realizar reservaciones para un determinado día y hora. De esta manera, un sistema proactivo realizará una reservación anticipada de acuerdo a la predicción de un usuario dado.

- Domótica. A la fecha, algunos trabajos relacionados han tomado como base la regularidad del usuario al visitar ciertos lugares para automatizar o controlar diversos aspectos del hogar. Por ejemplo, controlar la calefacción o el aire acondicionado de acuerdo al próximo arribo o partida de los usuarios (Ellis *et al.* (2012); Scott *et al.* (2011); Das *et al.* (2002); Krumm y Brush (2011)). De esta manera, resulta interesante realizar diversas acciones considerando los tiempos de arribo a ciertos lugares.

- Sistema de recordatorios. Otro nicho de oportunidad en el cual la predicción en el mediano plazo toma relevancia, son los sistemas de recordatorios (Ashbrook (2002); Sadilek y Krumm (2012b)). Actualmente, como ya se mencionó, existen diversos sistemas que ofrecen recordatorios de una manera proactiva. Sin embargo, esta cualidad se puede mejorar o aumentar al considerar no sólo la información que el usuario proporciona, sino también la información que el sistema aprende al analizar los datos del usuario; en este caso particular, la información relacionada a la movilidad de los usuarios en el mediano plazo. De esta manera, un sistema proactivo puede proporcionar información relevante para que el usuario administre sus actividades u organice su jornada.

Debido a lo anterior, en este trabajo se tiene interés en predecir la ubicación del usua-

rio en el mediano plazo. Esto es, considerando un tiempo actual T se desea conocer dónde estará el usuario en las próximas N horas ($(T + \Delta T)$ para algún valor ΔT). A la fecha no se ha encontrado un modelo de predicción que permita estimar la ubicación del usuario en el mediano plazo (varias horas) de manera efectiva.

El modelo de predicción propuesto toma como referencia los patrones repetitivos de las visitas de los usuarios a ciertos lugares que son importantes en sus actividades cotidianas. Además, el modelo supone la existencia de la propiedad Markoviana al considerar la transición del usuario entre los lugares que son importantes para él. Posteriormente, considera la propiedad Markoviana de la movilidad del usuario entre los lugares, la relación de los tiempos de estadía en estos lugares con tiempos específicos del día para modelar la movilidad del usuario como un modelo oculto de Markov. Este modelado permite realizar la predicción espacio-temporal de la movilidad del usuario.

1.4. De la predicción individual a la colectiva

Más allá de analizar los datos de localización de cada usuario, e identificar los patrones de movilidad del usuario, resulta de interés el combinar las predicciones de cada uno de los usuarios para definir e implementar sistemas y/o servicios que sean de utilidad para la población en general. De esta manera, la población puede consultar un servicio de predicción para conocer la capacidad de una ubicación en particular. Por ejemplo, resulta realista pensar que los usuarios desean tener conocimiento de la cantidad de personas que habrá en un restaurante a una hora determinada a fin de decidir si es adecuado ir a dicho lugar o a algún otro; o bien, consultar la cantidad de personas que habrá en una oficina gubernamental a fin de realizar el pago de servicios.

Por lo tanto, este enfoque resulta adecuado para consultar cualquier ubicación que sea de carácter público. De igual manera, a partir de la unión de las predicciones individuales, el servicio puede prevenir o avisar al usuario que un lugar que va a visitar en las próximas horas estará congestionado, así el usuario puede tomar una acción al respecto. Esta situación puede resultar paradójica, dado que los usuarios que utilizan el servicio pueden evitar el lugar, y al final el lugar no estará congestionado. Aunque este aspecto es de sumo interés, en el enfoque de este trabajo la predicción de la movilidad se realiza sin

considerar el comportamiento paradójico.

1.5. Objetivo de la investigación

Con base en lo discutido hasta el momento, el objetivo de investigación de este trabajo reside en analizar los datos de la movilidad del usuario a fin de definir un modelo de predicción que permita estimar la siguiente o siguientes ubicaciones donde estará el usuario en un periodo de tiempo dado, además del tiempo de arribo a dichos lugares.

1.5.1. Preguntas de investigación

Con el propósito de cumplir con el objetivo de investigación, este trabajo respondió las siguientes preguntas de investigación:

- Los dispositivos móviles proveen diversos datos contextuales del usuario, los cuales son utilizados en diversas investigaciones para identificar y definir patrones de comportamiento y modelos de predicción espacial o espacio-temporal. Con el objetivo de definir el modelo de predicción que permitiera conocer la movilidad del usuario en el mediano plazo, la pregunta que se respondió fue *¿cuáles son los datos contextuales que se deben considerar para definir el modelo de predicción espacio-temporal?*
- Un factor importante al momento de definir el modelo de predicción es identificar aquellos lugares que son importantes para el usuario en sus actividades cotidianas. Estos lugares se toman como referencia para realizar la predicción, y así conocer en cuál o cuáles de estos lugares estará el usuario y la hora del arribo a ellos.Por lo tanto, una de las preguntas que se respondió fue *¿cuáles aspectos se deben de considerar para identificar de manera adecuada los lugares que son importantes en las actividades cotidianas de los usuarios?*
- Para definir de manera precisa el modelo de predicción espacio-temporal, es necesario determinar la cantidad de datos que se deben considerar para realizar el entrenamiento del modelo de predicción, de esta manera se respondió la siguiente pregunta: *¿cómo determinar la cantidad de datos necesarios para realizar el entrenamiento del modelo de predicción espacio temporal?*

- A la fecha hay diversas investigaciones que se enfocan en realizar la predicción espacio-temporal, cada una de las cuales utiliza un método distinto (i.e., modelos ocultos de Markov, series de tiempo), por consiguiente, otra pregunta que se respondió fue: *¿cuál técnica resulta adecuada para modelar la movilidad del usuario a fin de realizar la predicción espacio-temporal?*

- Al tomar en cuenta que la movilidad del usuario varía a lo largo del tiempo, es necesario considerar los cambios en el patrón de movilidad, y así contar con un modelo preciso a lo largo del tiempo. Por lo tanto, otra pregunta que se respondió fue: *¿cómo el modelo de predicción espacio-temporal puede identificar e incluir los cambios en el comportamiento de la movilidad del usuario?*.

- Finalmente, debido a que no siempre se cuenta con una gran cantidad de datos para realizar el análisis de la movilidad del usuario, se respondió a la pregunta: *¿qué mecanismo debe incorporar el modelo predicción espacio-temporal a fin de compensar la falta de datos de localización?*

1.6. Metodología

Para responder las preguntas de investigación, y así lograr el objetivo de investigación de este trabajo, se siguió la siguiente metodología:

1.6.1. Análisis de la literatura

Como etapa inicial, se realizó un análisis exhaustivo de la literatura referente al análisis de la movilidad de los usuarios, y modelos de predicción espacial y espacio-temporal. El objetivo de este análisis fue identificar los datos contextuales que utilizan estos modelos de predicción, el enfoque y dominio de aplicación que abordan, así como también identificar las ventajas y limitaciones de éstos.

1.6.2. Análisis de la movilidad

El objetivo de esta etapa fue analizar la movilidad del usuario a fin de identificar aquellas características que son importantes al momento de modelar la movilidad. Específica-

mente, se presta atención a las características espaciales y temporales de la movilidad, y la relación entre éstas, con el fin de definir el modelo de predicción espacio-temporal.

1.6.3. Identificación de los lugares significativos o puntos de interés

Una parte importante al momento de definir el modelo de predicción fue identificar aquellos lugares que se toman como referencia para realizar la predicción espacio-temporal. Por lo tanto, en esta etapa se procedió a identificar aquellos lugares que son importantes o significativos para el usuario. Para ello, se realizó un análisis de los métodos existentes para identificar estos lugares, los factores que consideran y el enfoque de cada uno de los métodos. En este trabajo, a estos lugares se les denominará *puntos de interés.*

1.6.4. Definición del modelo de predicción

Luego de conocer las características espaciales y temporales de la movilidad, y una vez que se identificaron los lugares que son importantes para el usuario, se definió el modelo de predicción espacio-temporal utilizando los modelos ocultos de Markov. En esta etapa se especifica cómo se definieron cada uno de los elementos de los modelos ocultos de Markov a fin de contar con el modelo de predicción espacio-temporal.

1.6.5. Evaluación

Con la finalidad de conocer la eficiencia del modelo de predicción propuesto, en esta etapa se definió la manera en que el modelo de predicción fue evaluado. A partir de las evaluaciones se identificaron los mecanismos que ayudaron a incrementar la precisión, y fue posible identificar aspectos de la movilidad que no se han considerado en trabajos previos. Por lo tanto, con el conocimiento que se obtuvo en cada una de las evaluaciones se redefinió el modelo de predicción.

De manera gráfica, la metodología se presenta en la Figura 5.

1.7. Contribución del trabajo

De manera general, la contribución de esta investigación se puede resumir de la siguiente manera:

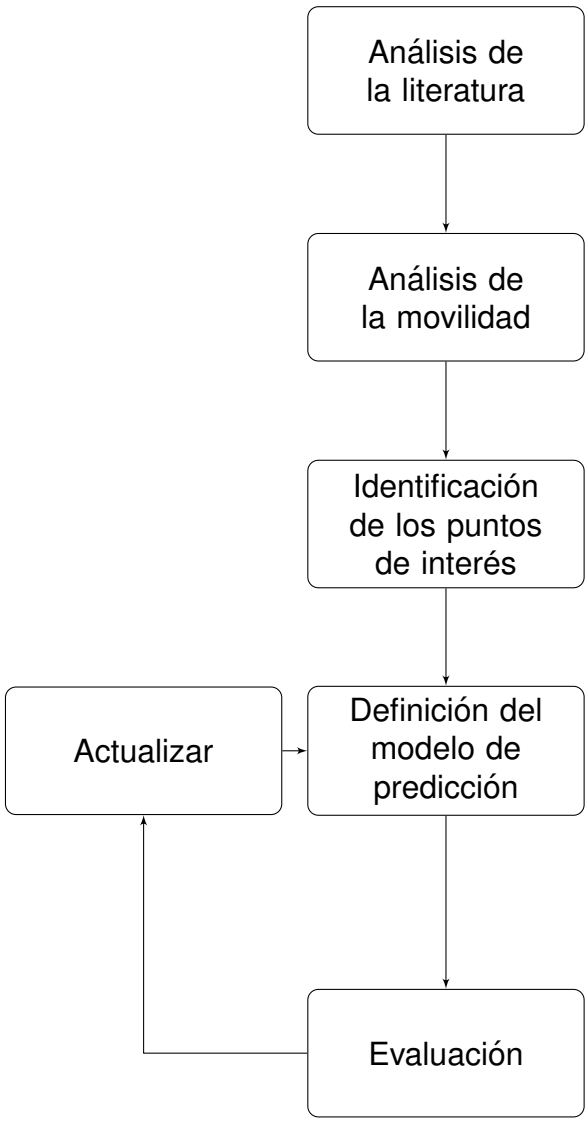


Figura 5: Metodología

- Se presenta un modelo para predecir la movilidad del usuario en los aspectos espacial y temporal. Este modelo se basa en los modelos ocultos de Markov, y toma como referencia la hipótesis de que la movilidad del usuario entre los puntos de interés se puede representar como una cadena de Markov; la existencia de la propiedad Markoviana se verifica de manera experimental.

- Se presenta un mecanismo que identifica el periodo de tiempo que abarca el patrón de movilidad actual del usuario. Este mecanismo compara la movilidad diaria del usuario con el fin de determinar el periodo de tiempo que comprende el patrón de movilidad actual del usuario. Este mecanismo es de suma importancia debido a que permite realizar el entrenamiento de cada uno de los modelos de predicción

al considerar únicamente los datos de localización del patrón de movilidad actual. De esta manera, se evita incluir datos de localización correspondientes a diferentes patrones de movilidad.

- Se utilizaron dos conjuntos de datos de acceso público para evaluar el modelo de predicción propuesto; estos conjuntos presentan datos realistas de la movilidad del usuario. Uno de ellos (Kotz *et al.* (2007a)) contiene registros de las conexiones de los usuarios a puntos de acceso dentro de un campus universitario, el otro proporciona datos de la movilidad de los usuarios en áreas urbanas Zheng *et al.* (2009).

- Los resultados que se obtuvieron por el modelo de predicción propuesto se comparan con aquellos que se obtuvieron con un método basado en NextPlace (Scellato *et al.* (2011)). Con el modelo propuesto se obtuvo una precisión de hasta 85 % para periodos cortos de predicción, y hasta un 70 % cuando el periodo de predicción es de siete horas, ambas precisiones son mayores que las obtenidas con el método basado en NextPlace.

- Con el objetivo de contrarrestar la falta de datos de localización y por consiguiente la identificación parcial de los puntos de interés, se definió un mecanismo basado en la técnica denominada *filtrado colaborativo*. Para un usuario dado, este mecanismo permite agregar lugares que inicialmente no se consideraron como puntos de interés. Para ello se toma como referencia los lugares que fueron visitados por los usuarios que son similares a éste. La similitud se encuentra en función de los lugares visitados. Este mecanismo permitió un incremento de la precisión de hasta 13 %.

1.8. Organización de la tesis

El resto de este trabajo se organiza de la siguiente manera:

En el Capítulo 2 se presenta una descripción de aquellos trabajos que son relevantes a la predicción espacio - temporal. Específicamente, el trabajo relacionado se divide en aquellos modelos de predicción que consideran únicamente el aspecto espacial, y aquellos que consideran tanto el aspecto espacial como el temporal. De igual manera, se

realiza una descripción de aquellos trabajos que utilizan la similitud de los usuarios para realizar la predicción de la movilidad del usuario.

En el Capítulo 3 se analiza la movilidad de los usuarios a fin de caracterizarla, y así se describe como se determinaron los factores para definir el modelo de predicción. Posteriormente, se define el modelo oculto de Markov.

En el Capítulo 4 se presenta la evaluación del modelo de predicción propuesto. Además, se detallan las características de los conjuntos de datos utilizados. Debido a que se realizaron diferentes experimentos, para cada uno de ellos se efectúa una descripción acerca de la configuración, los datos que se utilizaron, y la manera en que se determina la efectividad del modelo de predicción.

En el Capítulo 5 se presentan los resultados que se obtuvieron en cada uno de los experimentos definidos. Los resultados se presentan en función de la precisión obtenida por los modelos de predicción.

En el Capítulo 6 se presenta un conjunto de escenarios de aplicación en los cuales la predicción de la movilidad tiene un rol importante. De manera particular, los escenarios presentados se encuentran enfocados en ofrecer servicios de utilidad para la población, por lo cual se requiere de la participación de la población.

Finalmente, en el Capítulo 7 se presenta una discusión acerca de lo aprendido durante el desarrollo del trabajo de tesis. Así también, se presentan las limitaciones inherentes a este trabajo. Además, se presentan algunos proyectos que se han identificado en el transcurso del trabajo, los cuales dan la pauta para realizar investigación futura.

Capítulo 2. Marco teórico

De acuerdo al enfoque de esta investigación, el trabajo relacionado se cataloga como: modelos de *predicción espacial* y modelos de *predicción espacio-temporal*. Se presenta una descripción de cada uno de los trabajos relacionados, enfatizando en el alcance de éstos, la definición del modelo de predicción, y la información de contexto que utilizan. Además, se presenta la descripción de aquellos trabajos que toman como referencia aspectos sociales para realizar la predicción de la movilidad del usuario.

2.1. Modelos de predicción espacial

La predicción espacial tiene como objetivo predecir la próxima o próximas ubicaciones del usuario. Esto es, considerando un conjunto de lugares identificados previamente ($p = \{p_1, p_2, p_3, ..., p_n\}$, donde p_i representa el i-ésimo lugar en el cual ha estado el usuario), los modelos de predicción espacial consideran la ubicación actual (p_j) del usuario para estimar de manera probabilística el siguiente lugar en el que estará éste (p_i), o bien, la secuencia de lugares que visitará ($seq(p_i), i = 1..N$)(Ashbrook y Starner (2003); Kim *et al.* (2006); Krumm y Horvitz (2006); Liao *et al.* (2006); Nicholson y Noble (2008); Song *et al.* (2006)).

Krumm y Horvitz (2006) presentan un sistema denominado *predestination*, el cual permite predecir hacia dónde se dirige un conductor conforme se moviliza de una región geográfica a otra. Para definir el modelo de predicción, el área geográfica de interés se segmenta en celdas de un kilómetro de longitud; el área comprende una región de 40x40 kilómetros. Posteriormente, al considerar el historial de trayectorias del conductor, se define la matriz de transición que define la probabilidad ($P(j|i)$) de que el conductor se traslade de una celda i a una celda j. Además, el modelo de predicción considera el tipo de destino, la eficiencia y tiempos de manejo.

De manera similar, Ziebart *et al.* (2008) presentan el sistema *PROCAB*, el cual utiliza las trayectorias históricas de 25 taxistas para identificar patrones de manejo. Los autores definen la matriz de transición de los taxistas entre diversas regiones geográficas como un proceso de Markov, en el cual la transición entre los diferentes estados tiene asignado un peso que depende de información acerca del tipo de ruta (autopista, inter-

sección, calle principal, calle secundaria), límite de velocidad y número de carriles. Con esta información el modelo predice el destino del taxi tomando como referencia una ruta parcialmente recorrida. Por su parte, Calabrese *et al.* (2010) definen un modelo probabilístico que considera no sólo las trayectorias históricas de un usuario en específico, sino también las trayectorias de la población, puntos de interés, categorías de los puntos de interés y tiempos de traslado entre éstos, distancia promedio que se moviliza cada uno de los usuarios y el tipo de suelo (i.e. bosque, área comercial, área industrial, etc.) para realizar la predicción de la movilidad de un usuario en específico.

Con el objetivo de utilizar las trayectorias de la población para definir el modelo de predicción de un usuario en particular, se define un área geográfica que es común para todos los involucrados. El área geográfica se divide en N celdas de un tamaño dado. La identificación de los puntos de interés se realiza considerando la información de Yelp [1], luego los puntos de interés son agrupados en 22 categorías.

La predicción espacial también es útil para la comunicación en redes celulares. Específicamente, la predicción espacial se utiliza para estimar la carga que tendrán las celdas de comunicación, o bien para transferir el servicio de comunicación de una celda a otra sin problema alguno (*smooth hand-off*)(Cheng *et al.* (2003)). En este ámbito, en Yavas *et al.* (2004) el objetivo es predecir en cuál celda de una red celular estará el usuario, teniendo como restricción la topología de la red; existe un número limitado de celdas a las cuales el usuario puede llegar desde su ubicación actual. El problema se modela como un proceso de Markov en el que cada celda representa un estado del proceso, y la movilidad histórica del usuario entre las celdas define la matriz de transición.

Un trabajo similar es *BreadCrumbs* (Nicholson y Noble (2008)), en el cual se combina la información de la movilidad de los usuarios con la información de las conexiones a puntos de acceso para predecir la próxima ubicación donde el usuario tendrá conectividad. Al igual, en BreadCrumbs se utilizan las cadenas de Markov para modelar el problema.

Song *et al.* (2003) utilizan datos de conexiones a puntos de acceso para predecir el próximo punto de acceso donde se conectará el usuario. Cada punto de acceso se re-

[1] http://www.yelp.com

presenta mediante un caracter, y los movimientos del usuario se representan mediante una cadena de caracteres. Song *et al.* (2003) utilizan varios predictores para maximizar la precisión, y encontraron que un predictor (de bajo orden) basado en Markov es tan eficiente como los predictores de mayor orden.

Con estos resultados, los autores confirman que los movimientos recientes del usuario resultan ser un mejor predictor en comparación con las probabilidades que se obtienen a partir de las trayectorias o movimientos históricos del usuario. Por su parte, Ashbrook y Starner (2003) utilizan datos GPS para, en una primera instancia identificar lugares significativos, esto es, lugares que son importantes en las actividades cotidianas del usuario. Posteriormente, consideran las probabilidades de transición entre estos lugares significativos para definir un modelo de predicción que se basa en un cadena de Markov de segundo orden.

Eagle y Pentland (2006, 2009) utilizan los datos recolectados durante el proyecto Reality Mining del MIT para definir un modelo de predicción de la movilidad de los usuarios. Los registros de localización se obtienen a partir de las conexiones de los dispositivos móviles a las torres celulares (GSM), por lo que cuentan con datos de localización a un nivel de área geográfica. Los autores utilizan una red bayesiana dinámica para predecir la próxima ubicación del usuario.

Por su parte, Nguyen y Giordano (2012) presentan un enfoque que mezcla los resultados de diferentes modelos de predicción para maximizar la precisión al estimar la próxima ubicación del usuario. Los autores evalúan el modelo de los k vecinos, redes bayesianas dinámicas, máquina de soporte vectorial y árboles de decisión, posteriormente los autores asignan un peso a los resultados de cada modelo a fin de seleccionar de manera precisa el próximo lugar a visitar. Además de considerar aspectos espaciales y temporales, los autores utilizan aspectos sociales, por ejemplo, los registros de llamadas realizadas/recibidas, registros del acelerómetro, tono del dispositivo (volumen), entre otros. El enfoque de fusión se evaluó con los datos del Nokia Mobile data Challenge (Laurila *et al.* (2012)), y se obtuvo una precisión promedio de 61 %.

La predicción espacial no sólo se ha aplicado en exteriores, como es el caso de los tra-

bajos mencionados. Petzold *et al.* (2005a,b) presentan algoritmos para predecir la próxima ubicación dentro de un edificio. Los autores presentan varios algoritmos: dos redes neuronales, una red bayesiana, un predictor de estado, y un predictor basado en Markov; no encontraron resultados concluyentes para elegir el mejor algoritmo de predicción.

En el proyecto MavHome (Managing and Adaptive Versatile Home) de la Universidad de Texas, se utiliza la predicción de la ubicación del usuario para aumentar la comodidad de éste, y minimizar costos de operación. La predicción se utiliza para ajustar la iluminación y la temperatura de las habitaciones en donde estará el usuario (Das *et al.* (2002); Cook *et al.* (2003)). De igual manera, Gellert y Vintan (2006) presentan un modelo basado en los modelos ocultos de Markov, en donde los estados ocultos se definen por oficinas dentro de un edificio. El propósito de este enfoque es predecir la próxima oficina que se visitará. Los autores utilizan diversos valores para el orden del HMM, así como el número de estados ocultos a fin de maximizar la precisión; obtienen una precisión de hasta 84.81 % cuando se utiliza un HMM de orden 1 y 4 estados ocultos. Bhattacharya y Das (2002), Liu y Maguire (1996), y Ashbrook y Starner (2003) han presentado otros modelos de predicción que se basan en Markov.

Al considerar el alcance de los trabajos mencionados, no es posible conocer la secuencia de lugares en los que estará el usuario porque el objetivo de éstos es predecir sólo el próximo lugar que visitará el usuario; la mayoría de estos trabajos se basan en modelos Markovianos o Bayesianos. En Song *et al.* (2003) se puede encontrar la evaluación de varias técnicas para realizar la predicción de la próxima ubicación del usuario.

2.2. Predicción espacio-temporal

Antes de presentar los trabajos relacionados, es necesario realizar una aclaración importante ya que el término predicción espacio-temporal puede ser ambiguo y causar confusión. La predicción espacio-temporal se puede considerar como dos aspectos; el primero, predecir dónde (lugar p_i) estará un usuario en un tiempo dado (tiempo t); el segundo, predecir a qué hora (tiempo t) un usuario estará en un lugar determinado (lugar p_i). A la fecha, sólo el trabajo de Burbey y Martin (2012a) contempla la predicción espacio-temporal como dos aspectos. Sin embargo, en los trabajos relacionados, y en el

presente trabajo, el término hace referencia a la predicción de la ubicación del usuario en un periodo de tiempo. Esto es, tomando como referencia el tiempo actual T, el objetivo es conocer en qué lugar p_i o secuencia de lugares $seq(p_i), i = 1...N$ estará el usuario en un periodo de tiempo dado $[T, T + \Delta T]$, donde ΔT representa el periodo de predicción (e.g., cinco horas).

En el trabajo de Burbey y Martin (2012a), se utilizan dos modelos probabilísticos basados en las propiedades periódicas y no periódicas de la movilidad. El modelo periódico se basa en la premisa de que los usuarios visitan ciertos lugares de manera periódica (i.e. cada 3 horas, diariamente, cada mes); el modelo no periódico extrae patrones de movimientos repetitivos, la premisa consiste en que los patrones tienden a repetirse en un determinado periodo de tiempo en el futuro. Este modelo obtiene una precisión promedio de 52 % al considerar un ΔT de 30 días.

El periodo de predicción (ΔT) considerado por cada modelo de predicción varía de acuerdo al enfoque del trabajo. Al analizar el estado del arte es posible diferenciar entre modelos que predicen la ubicación del usuario en el corto (e.g., 1 hora) y en el largo plazo (e.g., 1 año).

2.2.1. Predicción a largo plazo

Sadilek y Krumm (2012b) presentan un modelo de predicción para estimar la movilidad de usuarios en el largo plazo. Ellos consideran tanto características espaciales (latitud, longitud y áreas geográficas) como temporales (día de la semana, días festivos y hora del día). Definen lo que llaman *eigendays*, que representan patrones periódicos de movilidad, los cuales se utilizan para predecir la ubicación del usuario. El modelo de predicción obtiene una buena precisión aún cuando el periodo de predicción ΔT es de 80 semanas. Por su parte, Sodkomkham *et al.* (2013) presentan el modelo *APP* para predecir la ubicación considerando un ΔT de hasta un mes.

A diferencia del trabajo de Sadilek y Krumm (2012b), APP se emplea en interiores por lo que se utilizan diversos sensores para conocer la movilidad de los usuarios. En APP los patrones de movilidad se definen mediante un arreglo (1). Este arreglo consiste de 32 índices, los primeros 24 índices definen las visitas al lugar x durante un periodo

específico del día, el cual se define a una hora. Los siguientes índices corresponden al día de la semana, y el último corresponde a un día festivo. De esta manera, se modela un comportamiento en particular en el cual se visitan ciertos lugares en determinados periodos de tiempos en varios días de la semana. *APP* obtiene una precisión promedio de 55 % al considerar un ΔT de 30 días.

$$d_x = [v_x, dayweek, holiday] = [v_{to}, ..., v_{t23}, Sun, Mon, ..., Sat, Hol] \quad (1)$$

2.2.2. Predicción a corto plazo

A diferencia de los dos trabajos anteriores, el resto de los trabajos se enfocan en estimar la movilidad del usuario considerando periodos cortos de predicción (ΔT). Vu *et al.* (2011b) presentan el modelo de predicción *Jyotish*, el cual utiliza datos de conexiones a puntos de acceso y registros de Bluetooth con el fin de analizar la movilidad e identificar patrones de comportamiento. Los datos se agrupan de acuerdo al tipo de día, ya sea día laboral o fin de semana. Posteriormente, los datos se reagrupan de acuerdo a diferentes periodos de tiempo (1-8 horas). De esta manera, el objetivo es conocer en cuáles lugares estuvo el usuario en un determinado día y periodo de tiempo.

Finalmente, cada periodo de tiempo se aumenta con datos de contactos Bluetooth; es decir, se conoce la interacción con otras personas durante dicho periodo de tiempo. A partir de ello, *Jyotish* predice dónde estará el usuario considerando el tipo de día, y el periodo de tiempo. Además, predice la duración de la estadía en dicho lugar e infiere con cuál persona o personas se encontrará el usuario en dicho periodo. La evaluación de este modelo de predicción se lleva a cabo considerando datos realistas de 50 personas, los cuales se recolectaron en el Campus de la Universidad de Illinois. Con respecto a la predicción de la ubicación, *Jyotish* obtiene una precisión de hasta 80 %.

Otro modelo de predicción es *WhereNext* (Monreale *et al.* (2009)), a diferencia de *Jyotish*, este modelo tiene como objetivo el estimar la ubicación futura de un vehículo. Para ello, *WhereNext* utiliza datos espaciales y temporales para predecir la próxima región en la que estará un vehículo, y así también la hora del arribo a dicha región. Los autores

utilizan datos GPS, los cuales se agrupan en regiones, las transiciones entre estas regiones se representan mediante un árbol de decisión. Cada hijo representa la transición desde la región padre a la región hijo, además incluye información del intervalo de tiempo en la que se realizó la transición del nodo padre al hijo. Posteriormente, las trayectorias parciales se comparan con las rutas en el árbol para predecir regiones futuras.

La evaluación de *WhereNext* se lleva a cabo utilizando un conjunto de datos que contiene trayectorias GPS de 17000 vehículos. Este modelo de predicción alcanza una precisión de hasta 54 %.

En el trabajo de tesis doctoral de Ingrid Burbey (Burbey (2011)), se presenta un modelo de predicción basado en Markov, el cual utiliza datos de conexiones a puntos de acceso para modelar la movilidad de los usuarios. Cada registro de conexión contiene la ubicación (punto de acceso), fecha y hora de conexión. Con estos datos Burbey (2011) predice las próximas ubicaciones del usuario. Los registros de conexión corresponden a 275 estudiantes que se movilizan en un campus universitario. Además, en este trabajo Burbey (2011) presenta otro método de predicción cuyo objetivo es predecir el tiempo en que un usuario estará en un lugar dado. La precisión de la predicción alcanza hasta un 91 %.

Mathew *et al.* (2012) presentan otro modelo de predicción basado en Markov, el cual realiza la predicción de la próxima región geográfica en la que estará el usuario; utilizan los modelos ocultos de Markov para realizar el modelado de la movilidad. Los autores utilizan registros de GPS (latitud, longitud, hora del día) que posteriormente se asocian a diferentes regiones geográficas de un tamaño dado. Luego, las visitas a las regiones se catalogan de acuerdo al periodo de tiempo en que éstas se realizaron: visitas en los días laborales de 7:00 AM a 7:00 PM, visitas en los días laborales de 7:00 PM a 7:00 AM, y visitas en los días de fin de semana. Este modelo de predicción obtiene una precisión de hasta 13.85 %; la precisión es baja debido a que consideran todos los registros del conjunto de datos Geolife, y ya que éste contiene registros de localización de 4 años, los usuarios cuentan con varios patrones de movilidad.

Los modelos de predicción no sólo utilizan datos de los sensores incorporados en

los dispositivos móviles, algunos proyectos utilizan otro tipo de datos contextuales. Por ejemplo, Krumm y Brush (2011) presentan un algoritmo que utiliza los datos de GPS y la información de cuestionarios para predecir cuándo un usuario estará en su casa o fuera de ella. En los cuestionarios los usuarios anotan los tiempos de arribo y partida de su hogar; sin embargo, los autores demuestran que los registros de llegada y salida del hogar no son precisos, por lo que definen un modelo probabilístico a partir de las lecturas GPS.

El modelo de predicción resulta más preciso que los registros definidos por los usuarios; obtiene hasta un 65 % de precisión. Partiendo del enfoque de este trabajo, Scott *et al.* (2011) presentan el sistema *PreHeat*, el cual toma como referencia datos de RFID y datos de sensores de movimiento (colocados en el hogar) para definir un modelo cuyo objetivo es predecir cuándo los usuarios estarán en casa y fuera de ella. Para ello, cada día de la semana se define como un arreglo de 96 posiciones, donde cada posición representa un periodo de 15 minutos. El valor de cada índice es binario, define si el usuario estuvo o no en el hogar. Para realizar la predicción, el modelo considera las observaciones parciales del día en cuestión, y compara con los arreglos históricos a fin de encontrar un comportamiento similar y así predecir cuándo el usuario estará en el hogar.

Este modelo tiene como escenario de aplicación el ahorro de energía en el hogar; al estimar el tiempo en que el usuario estará en su hogar, se evita el gasto de energía innecesaria por parte del sistema de calefacción, ahorrando gas y minimizando el tiempo en que la casa está ocupada, pero fría. El sistema se evaluó en 5 hogares, obteniendo una precisión de hasta 85 %. Posteriormente, los autores presentan *EarlyOff*, el cual es una modificación de *PreHeat*. La diferencia radica en que *EarlyOff* controla (apaga) la calefacción anticipando la salida del hogar por parte de los residentes.

Otro trabajo que resulta de interés es *NextPlace* (Scellato *et al.* (2011)), el cual predice el lugar o lugares en donde estará el usuario en un periodo de tiempo dado. Además, *NextPlace* predice la hora de llegada a dichos lugares, y el tiempo de estadía en éstos. Para ello, en una primera instancia se identifican los lugares que son significativos para cada usuario. Luego, para un instante de tiempo actual t y un periodo de predicción ($\Delta(T)$), los autores toman como referencia la secuencia de los últimos lugares visitados

por el usuario, y utilizan los registros históricos para encontrar un patrón similar, y así predecir el siguiente lugar que visitará el usuario. Además, se predice el tiempo de arribo y el tiempo de estadía; la mejor precisión (alrededor de 90 %) se obtiene cuando se considera un periodo de predicción ($\Delta(T)$) de 5 minutos, y los últimos 3 lugares visitados por el usuario. Si el periodo de predicción se incrementa a 60 minutos, la precisión disminuye a un 70 %.

NextPlace se evaluó con cuatro conjuntos de datos que se encuentran disponibles de manera pública. Dos de estos conjuntos contienen registros de GPS; uno de ellos contiene registros de los usuarios que utilizaron la aplicación CenceMe (Miluzzo *et al.* (2008)), y el otro contiene registros de la movilidad de taxis en la ciudad de San Francisco (Piorkowski *et al.* (2009)). Los dos conjuntos de datos restantes contienen registros de conexiones a puntos de acceso del campus de Dartmouth (Kotz *et al.* (2007b)), y de la red inalámbrica Ile San Fils en Montreal, Canadá (Lenczner *et al.* (2007)).

Aunque los trabajos mencionados son útiles en varios dominios de aplicación, éstos no ofrecen la funcionalidad que se requiere para satisfacer el propósito de este trabajo. A pesar de que algunos de los modelos de predicción consideran un periodo de predicción ΔT de varias horas, la precisión que éstos obtienen es baja (40 % para un ΔT de 8 horas). Además, estos modelos de predicción no consideran la dinamicidad de la movilidad del usuario a lo largo del tiempo. Esto es, el entrenamiento de cada modelo de predicción se realiza con los datos de diversos patrones de movilidad, dando la pauta para definir un modelo inadecuado e impreciso. Así también, estos modelos de predicción contemplan la movilidad del usuario como una sola entidad; sin embargo, en trabajos previos se identificaron diferencias en la movilidad del usuario con respecto al día de la semana (Chon *et al.* (2012); Farrahi y Gatica-Pérez (2011); Hsu *et al.* (2007a); Motahari *et al.* (2012)), por lo que es necesario considerar diversos modelos de predicción.

Además de encontrar propuestas de modelos de predicción, en la literatura hay varios artículos que se han enfocado en hacer un análisis, o bien una comparación de los modelos de predicción actuales. Por ejemplo, Burbey y Martin (2012b) en una primera instancia, presentan una taxonomía de acuerdo a las técnicas de aprendizaje de máquina (*machine learning*): modelos supervisados y no supervisados. Posteriormente, presentan

una clasificación de acuerdo a la minería de datos en cuestión temporal: uso del tiempo, y tipo de datos a utilizar.

Por su parte, Chon *et al.* (2012) presentan la evaluación empírica de varios modelos de predicción tomando como referencia tres enfoques, aquellos modelos que son dependientes de la ubicación, independientes de la ubicación, y aquellos que consideran otras características. Después de evaluar nueve modelos de predicción, entre los que destacan los modelos basados en Markov y algunos basados en *NextPlace* (Scellato *et al.* (2011)), Chon *et al.* (2012) presentan 3 aspectos a considerar: 1) encontraron una alta regularidad en los aspectos espacial y temporal, 2) en sus experimentos los modelos dependientes de la ubicación obtuvieron mejores resultados; sin embargo, la secuencia de lugares visitados es inadecuada cuando se desea predecir el tiempo de estadía , y 3) los métodos que utilizan información temporal o aquellos que utilizan la probabilidad de regreso resultaron ser efectivos para extraer patrones importantes. Los resultados se obtuvieron al analizar los datos granulares de 10 estudiantes.

2.3. Modelos de predicción basados en aspectos sociales

En años recientes, diversos trabajos investigan la relación entre los patrones de movilidad y las interacciones sociales para predecir la próxima ubicación del usuario y los vínculos sociales, tomando como referencia la similitud de los patrones de movilidad (McGee *et al.* (2013); Terry *et al.* (2002); Backstrom *et al.* (2010); Crandall *et al.* (2010); Xiong *et al.* (2012); Domenico *et al.* (2013)).

Por ejemplo, Lian *et al.* (2013) presentan un método basado en la técnica de filtrado colaborativo (Adomavicius y Tuzhilin (2005a)) para predecir dónde un usuario registrará su próxima presencia (*check-in*) al considerar los patrones de comportamiento de los usuarios similares a éste. De igual manera, Noulas *et al.* (2012) utilizan los datos de registros de presencia (*check-in*) de Foursquare para predecir la movilidad de los usuarios. El modelo de predicción utiliza el tipo de lugar, y las características espaciales y temporales de los registros de presencia.

Por su parte, Calabrese *et al.* (2010) presentan un algoritmo que combina el comportamiento individual y colectivo. El algoritmo combina las trayectorias históricas de un

usuario específico y de la población al considerar características geográficas y puntos de interés. El comportamiento individual se define como una cadena de Markov de primer orden, donde los estados ocultos son los puntos de estadía que visitó el usuario, y la probabilidad de transición entre el estado i al estado j se define al considerar la movilidad histórica del usuario.

El comportamiento colectivo se modela como el promedio ponderado entre la influencia de la distancia entre los puntos de estadía, los puntos de estadía, y el uso del área geográfica. Un aspecto importante de este algoritmo, es la selección de la colectividad de manera precisa.

Como mencionan Calabrese *et al.* (2010), si se toma como referencia un empresario, es necesario elegir una colectividad que tenga hábitos similares a éste. Sin embargo, en este trabajo los autores no definen una colectividad específica para cada usuario considerado. El modelo propuesto toma como referencia los registros de 2000 usuarios, y obtienen una precisión del 60 % al predecir la próxima ubicación del usuario.

Gong *et al.* (2011) hacen uso de la información de las redes sociales para predecir la próxima ubicación de un usuario dado al considerar la última ubicación de sus amigos cercanos. Cabe mencionar que este enfoque no considera el historial de los lugares que el usuario visitó. Los autores definen una red social entre los usuarios; el peso del enlace entre cada par de usuarios se define de acuerdo a la cantidad de tiempo que los usuarios pasan juntos en los mismos lugares. Los autores utilizan un modelo oculto de Markov de segundo orden.

Cho *et al.* (2011) proponen el modelo PSMM (*Periodic and Social Mobility Model*) que considera los movimientos del usuario como distribuciones Gausianas independientes del tiempo. Los autores demuestran que los usuarios exhiben un fuerte comportamiento periódico a través de ciertos periodos de tiempo del día, alternando entre ubicaciones primarias (hogar) y secundarias (trabajo) en días laborales, y el hogar y ubicaciones relacionadas a la red social de los usuarios en los fines de semana. Los autores argumentan que la estructura de la red social no afecta los movimientos de corto alcance de los usuarios, en cambio sí influye en los movimientos de largo alcance. PSMM predice la ubicación

del usuario a cualquier hora del día con un 40 % de precisión.

Posteriormente, Tarasov *et al.* (2013) presentan un método para predecir la ubicación del usuario, este método se basa en los aspectos espaciales, temporales y sociales. A diferencia de PSMM, Tarasov *et al.* (2013) utilizan un modelo de radiación. Sadilek *et al.* (2012a) proponen *FLAP* para inferir la amistad y la ubicación más probable de un usuario dado a cualquier hora del día. Los traslapes de mensajes y de las listas de amistades se utilizan en un modelo basado en Markov para definir el grafo de amistad entre los usuarios. Después de definir el grafo de amistades, se define una red bayesiana dinámica que utiliza los lugares que el usuario visitó, los lugares visitados por los amigos del usuario considerado, la hora y el tipo de día (laboral, fin de semana) para predecir la ubicación del usuario. Los autores probaron el método con usuarios de Twitter de la ciudad de Los Angeles y Nueva York, y mostraron que el modelo predice la ubicación de un usuario con un 54 % de precisión, aun para los usuarios que no cuentan con datos explícitos de localización.

Los trabajos anteriores utilizan diversos datos contextuales para inferir los lazos sociales o la similitud entre un par de usuarios; información de la lista de contactos, llamadas telefónicas realizadas, mensajes de texto enviados, estructura de la red social, colocación, información geográfica (tipo de área geográfica, significado semántico de los puntos de interés), entre otros. Una vez que se conoce la amistad o la similitud, resulta viable considerar la movilidad de los k amigos o usuarios similares para inferir la ubicación de un usuario dado. En cambio, en el enfoque de este trabajo, sólo se cuenta con datos de localización, y por lo tanto la similitud se infiere a partir de las regiones o áreas geográficas que los usuarios visitaron.

2.4. NextPlace

A la fecha, el trabajo de Scellato *et al.* (2011) es el más similar al enfoque que se presenta en este trabajo. Debido a ello, la evaluación del modelo de predicción en diversos experimentos se realiza comparando los resultados de éste con aquellos que se obtienen al utilizar NextPlace. A continuación se presenta el funcionamiento de *NextPlace* a detalle.

Para cada usuario, se cuenta con registros de los lugares que visitó, la hora en que

se realizó la visita, y la duración de cada una de las visitas. De esta manera, el algoritmo predice la o las próximas visitas a un lugar dado considerando las visitas históricas a éste: $((t_1, d_1), (t_2, d_2), \ldots, (t_n, d_n))$, donde t_i es el tiempo de arribo del usuario, y d_i es la duración de la visita. El índice corresponde al número de visita al lugar considerado.

- Se crean dos series de tiempo a partir de la secuencia de las visitas previas: la serie de tiempo que define los tiempos de llegada C, y la serie de tiempo de la duración de las visitas D, definidas como:

  $C = (c_1, c_2, \ldots, c_n)$

  $D = (d_1, d_2, \ldots, d_n)$

  donde c_i representa el tiempo del día en segundos (i.e. c_i está en el intervalo [0,86400]);

- se busca en la serie de tiempo C las secuencias de m valores consecutivos $(c_{i-m+1}, \ldots, c_i)$ que sean similares a los últimos m valores $(c_{n-m+1}, \ldots, c_n)$;

- se predice el próximo valor de la serie de tiempo C al promediar todos los valores c_{i+1} que siguen en cada secuencia identificada;

- a la vez, en la serie de tiempo D, se seleccionan las secuencias correspondientes $(d_{i-m+1}, \ldots, d_i)$; las secuencias necesitan estar en los mismos índices que las secuencias en C;

- el próximo valor de la serie de tiempo D se estima al promediar todos los valores d_{i+1} que siguen en cada secuencia identificada.

De esta manera, si las últimas tres visitas a un cierto lugar son el día lunes a las 18:30, lunes a las 22:00 y martes a las 8:15, se analizan las visitas históricas para encontrar secuencias que numéricamente sean cercanas a (18:30, 22:00, 8:15), por ejemplo (18:10, 21:50, 8:35) y (18:35, 22:10, 8:00), y suponiendo que las próximas visitas que siguen estas secuencias son a las 13:10 y 12:40 y la duración en estas visitas fue de 40 y 30 minutos, respectivamente, se predice que la próxima visita será a las 12:55 con una duración 35 minutos; promediando los tiempo de arribo y de estadía. El parámetro m tiene

un impacto directo en la precisión de la predicción; ésta se puede mejorar al considerar más visitas, con lo cual se identifican patrones particulares que pueden estar presentes en periodos específicos de tiempo.

El algoritmo anterior se generaliza para predecir no sólo la próxima visita a un lugar significativo dado, sino también las próximas visitas considerando todos los lugares significativos. Si se considera que en un instante T se desea conocer en cuáles lugares significativos estará el usuario i después de ΔT segundos, el proceso para realizar esta predicción es:

- para cada lugar significativo se predice la secuencia de las próximas k visitas (iniciando con $k = 1$), y se crea una secuencia global de todas las visitas que se estimaron $(loc_1, t_1, d_1), \ldots, (loc_n, t_n, d_n)$ para $t_1 \leq \ldots \leq t_n$.
- si hay una predicción (loc_i, t_i, d_i) que satisface $t_1 \leq T + \Delta T \leq t_i + d_i$, loc_i se considera como el lugar significativo que se predijo (se da el caso de que varios lugares significativos cumplen con la condición, en ese caso se realiza una selección aleatoria);
- si ninguna predicción satisface la condición anterior, se tienen dos opciones: si el instante mínimo de llegada t_1 es más pequeño que $T + \Delta T$, la predicción necesita ampliarse a fin de encontrar una visita apropiada, así el parámetro k se incrementa y el algoritmo se repite considerando las nuevas visitas estimadas. De otra manera, aumentar la predicción proporciona visitas que inician después de $T + \Delta T$ y las cuales no se pueden explotar para la predicción; el algoritmo termina y da como resultado que el usuario no estará en ningún lugar significativo.

2.5. Comparativa del estado del arte

En las Tablas 3 y 4 se presentan una descripciones de los trabajos relacionados en función del dominio de aplicación, datos que toman en consideración, método de predicción, precisión, y en el caso que el modelo permita realizar la predicción espacio temporal, se presenta el máximo periodo de tiempo considerado (ΔT).

Tabla 3: Comparación de los trabajos relacionados

Proyecto	Aplicación	Ubicación a través de	Factores	Área geográfica	Método de predicción	ΔT	Precisión alcanzada
Jyotish	Predicción de la siguiente ubicación	Wi-Fi y Bluetooth	Tipo de día, periodo de tiempo, fecha, hora, y ubicación	Campus universita-rio	Bayes	De 1 a 8 horas	80 %
WhereNext	Predecir la próxima ubicación	GPS	Hora del día, coordenadas cartesianas	Área urbana	Árbol de decisión	No definido	54 %
Nguyen y Giordano (2012)	Predecir la próxima ubicación del usuario	GPS, GSM	Registros de llamadas, modo del tono y estado de carga del dispositivo; Movimiento, tiempo, lugar, tipo de día	Ciudad	Fusión de KNN, SVM, árboles de decisión, y redes bayesianas	No definido	61 %
Gellert y Vintan (2006)	Predecir la próxima habitación a ser visitada	No especifica	Fecha, hora, oficina	Edificio	HMM	No definido	Hasta 84.81 %
Mathew *et al.* (2012)	Predecir el área geográfica dónde estará el usuario en un tiempo dado	GPS	Hora del día, tipo de día, área geográfica	Urbana	HMM	No definido	Hasta 13.85 %

Tabla 4: Comparación de los trabajos relacionados

Proyecto	Aplicación	Ubicación a través de	Factores	Área geográfica	Método de predicción	ΔT	Precisión alcanzada
PreHeat	Controlar la calefacción del hogar	Sensores de movimiento y RFID	Hora del día, presencia	Hogar	Modelo probabilístico	Hogar	Hasta de 85 %
Burbey y Martin (2008)		Wi-Fi	Hora del día, ubicación	Campus universitario	Algoritmo PPM (Prediction by Partial Match)	No definido	Hasta 92 %
Eagle y Pentland (2009)	Predecir el comportamiento del usuario	GSM	Hora del día, etiquetas para 4 lugares predefinidos	Área urbana	Eigenbehavior	No definido	No definido
Krumm y Brush (2011)	Estimar cuando una persona estará en casa, o fuera de ella	GPS y registros de llegada y salida de casa	Día de la semana, hora,	Área urbana	Modelo probabilístico	No definido	Hasta un 65 %
Burbey y Martin (2012a)	Próxima ubicación del usuario	Wi-Fi	Lugar, tiempo, ubicación	Campus universitario	Algoritmo PPM	No Definido	Hasta 55 %
Sadilek y Krumm (2012b)	Predecir en dónde estará una persona en el futuro lejano (i.e. un año)	GPS	datos continuos (Latitud, longitud), datos discretos (celdas de 400 metros de lado), Día de la semana, hora, tipo de día	Área urbana	Fourier y análisis de componente principal (PCA)	No definido	Hasta un 93 %

2.6. Resumen

En este capítulo se presentó una descripción de los trabajos relacionados, especialmente de aquellos que se enfocan en la predicción espacial y en aquellos que realizan la predicción espacio-temporal. En las Tablas 3 y 4 se comparan algunos de los trabajos relacionados considerando algunos factores como: el área geográfica que contemplan, la precisión que éstos obtienen, el tipo de tecnología que utilizan para capturar datos, entre otros factores.

Al analizar los trabajos que permiten realizar la predicción espacio-temporal, se encontró que éstos se han enfocado en realizar la predicción en el corto y largo plazo. Es decir, estos trabajos consideran tiempos de predicción de algunos minutos (e.g., cinco minutos) o un par de años. Por lo tanto, estos enfoques no resultan adecuados para predecir de manera precisa la ubicación del usuario en las próximas horas (e.g., mediano plazo). Además, al analizar los modelos de predicción actuales, se identificaron varios aspectos que no se han considerado en dichos modelos, pero que resultan de suma importancia si se desea incrementar la precisión de la predicción.

Después de analizar los modelos de predicción, se observó que el enfoque y el dominio de aplicación de cada uno de éstos es diferente. Mientras algunos utilizan datos de localización de un periodo prolongado de tiempo para así identificar patrones que prevalecen en un periodo extenso de tiempo, otros utilizan una cantidad limitada de datos para conocer el comportamiento más reciente del usuario. Debido a ello, no es posible realizar una comparación directa entre los modelos de predicción, por consiguiente, para evaluar el enfoque propuesto se define un método basado en *NextPlace* a fin de realizar una comparación directa.

Debido a lo anterior, en el siguiente capítulo se propone y describe un modelo de predicción espacio-temporal que permite estimar la ubicación del usuario en el mediano plazo. Además de definir el modelo de predicción, en el siguiente capítulo se discuten varios aspectos que son cruciales para obtener una predicción precisa. Estos aspectos se consideran en el modelo de predicción con el objetivo de maximizar la precisión de la predicción. Entre estos aspectos destaca el entrenamiento del modelo de predicción, y la

actualización del modelo conforme cambia la movilidad del usuario.

Capítulo 3. Modelo de predicción espacio temporal

Con el propósito de definir el modelo de predicción espacio-temporal, en la Subsección 3.1 se describen las características de la movilidad del usuario que se identificaron en trabajos relacionados, y se tomaron como referencia para definir diversos modelos de predicción. Luego, en la Subsección 3.2, se presenta el modelo de predicción propuesto, el cual se basa en varias de las características de la movilidad y la relación que existe entre éstas.

3.1. Movilidad del usuario

De cierta manera pareciera ser que la movilidad de los usuarios es dinámica; sin embargo, los resultados de trabajos relacionados demuestran que la mayoría de los usuarios tienen patrones de movilidad definidos (Eagle y Pentland (2006); Farrahi y Gatica-Pérez (2011); Gonzalez *et al.* (2008)); sólo en casos extraordinarios el usuario tiene un comportamiento dinámico durante un periodo de tiempo (e.g., políticos en periodo de campaña).

La movilidad de los usuarios se encuentra definida por las actividades cotidianas que realizan, o bien por sus hábitos, tales como las actividades laborales, escolares, de recreación, u otras que varían a lo largo del tiempo. De esta manera, en los trabajos previos se identifican patrones de movilidad de duración variante: diarios, semanales, mensuales, anuales, entre otros. Por ejemplo, en el caso de un estudiante es posible identificar varios patrones que corresponden a los ciclos escolares y a los periodos vacacionales; en el caso de un trabajador que labora por turnos, sus patrones de movilidad dependen del cambio de turno, días de asueto, y al igual, de los periodos vacacionales.

Asimismo, otros autores (Djordjevic *et al.* (2011); Cao *et al.* (2007)) identifican periodicidad en los patrones de movilidad de los usuarios. Frecuentemente, las actividades de los usuarios en los días laborales son similares; los usuarios tienden a organizar sus agendas de acuerdo a sus actividades laborales o escolares. Así, también, la periodicidad en el comportamiento de los usuarios se observa en los fines de semana.

Otro aspecto que se ha identificado en trabajos relacionados, es que la movilidad del usuario es diferente para cada día de la semana, de esta manera cada día de la semana

cuenta con una patrón de movilidad repetitivo (Chon *et al.* (2012); Farrahi y Gatica-Pérez (2011); Hsu *et al.* (2007a); Motahari *et al.* (2012)). Por consiguiente, los lugares que un usuario visita en un día determinado de la semana (e.g., lunes), tenderán a ser los mismos lugares que el usuario visitará en los días siguientes (e.g., próximos lunes).

Otra característica importante de la movilidad es el hecho de que la próxima ubicación del usuario se encuentra definida o relacionada con la ubicación actual del mismo (Song *et al.* (2003); Mathew *et al.* (2012)). De esta manera, si un usuario dado se encuentra en su hogar existe cierta probabilidad de que su siguiente ubicación sea su trabajo, o bien una cafetería.

Finalmente, otro aspecto que se ha identificado en trabajos previos (Vu *et al.* (2011b); Krumm y Brush (2011); Scott *et al.* (2011)), es que la estadía del usuario en un lugar determinado se encuentra relacionada a la hora del día. De esta manera, existe una mayor probabilidad de que el usuario se encuentre en su hogar desde la noche hasta el amanecer, o bien que se encuentre en su lugar de trabajo durante la jornada regular (e.g., 9:00-18:00).

3.2. Propuesta del modelo de predicción espacio temporal

Debido a las características que se presentan, y a fin de definir el modelo de predicción, en una primera instancia el reto se encuentra en analizar la movilidad de los usuarios para entender, identificar, y caracterizar los patrones de comportamiento. Posteriormente, al tener conocimiento de los patrones de movilidad, resulta factible realizar la predicción de la movilidad del usuario tanto en el aspecto espacial como en el aspecto temporal.

En este trabajo se han tomado como referencia las características (*features*) anteriores y la relación entre éstas para definir el modelo de predicción. Debido a ello, en este trabajo se postula que la movilidad del usuario es un proceso estocástico, y que ésta se puede definir como una cadena de Markov. La propiedad Markoviana (Rabiner (1989)) estipula que el estado actual es función del estado anterior.

Considerando que la movilidad del usuario es un proceso estocástico, y que éste se puede representar como una cadena de Markov, resulta factible definir un modelo de

predicción que infiera la próxima ubicación del usuario. Para ello se toma como referencia únicamente la ubicación actual del usuario (*característica: ubicación actual*). Esto es, si se considera que la ubicación actual del usuario es su casa, sólo se puede inferir que la siguiente ubicación más probable es su oficina. Sin embargo, el hecho de considerar sólo la propiedad Markoviana, no es suficiente para el propósito de este trabajo. Por ejemplo, si para un día lunes la ubicación actual del usuario es su casa, es más probable que la próxima ubicación sea su oficina; en cambio, este comportamiento es diferente para un día sábado o domingo ya que no son días laborales. Para estos días, la próxima ubicación más probable es el supermercado o bien un lugar de recreación. Por lo tanto, considerar únicamente la ubicación actual no es suficiente, ya que varios lugares pueden ser las siguientes ubicaciones del usuario (Figura 6(a)).

Debido a ello, el modelo de predicción propuesto considera el aspecto temporal: *día de la semana*. Ya que la movilidad del usuario es diferente para cada día de la semana, es necesario considerar tanto la ubicación actual del usuario como el día de la semana, con el fin de realizar una mejor predicción (Figura 6(b)). No obstante, estos dos factores no son suficientes para realizar la predicción espacio-temporal de manera precisa. La precisión se puede maximizar al considerar el siguiente aspecto.

Con el objetivo de realizar la predicción espacio temporal, el modelo propuesto también considera la hora del día (*característica: hora del día*). De esta manera, el modelo toma en cuenta el hecho de que para un día dado, la próxima ubicación más probable difiere al considerar la hora actual. Por ejemplo, si en un día lunes el usuario se encuentra en su casa a las 8:00 AM, su próxima ubicación más probable es su oficina, pero no el cine o un restaurante bar; sin embargo, si el usuario se encuentra en su casa a las 20:00, su siguiente ubicación más probable es el cine (Figura 6(c)).

Así, al considerar la relación de la ubicación actual del usuario y la hora del día es posible inferir la siguiente ubicación del usuario. Por lo tanto, en este trabajo se considera la relación entre la ubicación actual del usuario, la hora del día, y el día de la semana para caracterizar de una mejor manera la movilidad del usuario, y como consecuencia definir el modelo de predicción espacio-temporal.

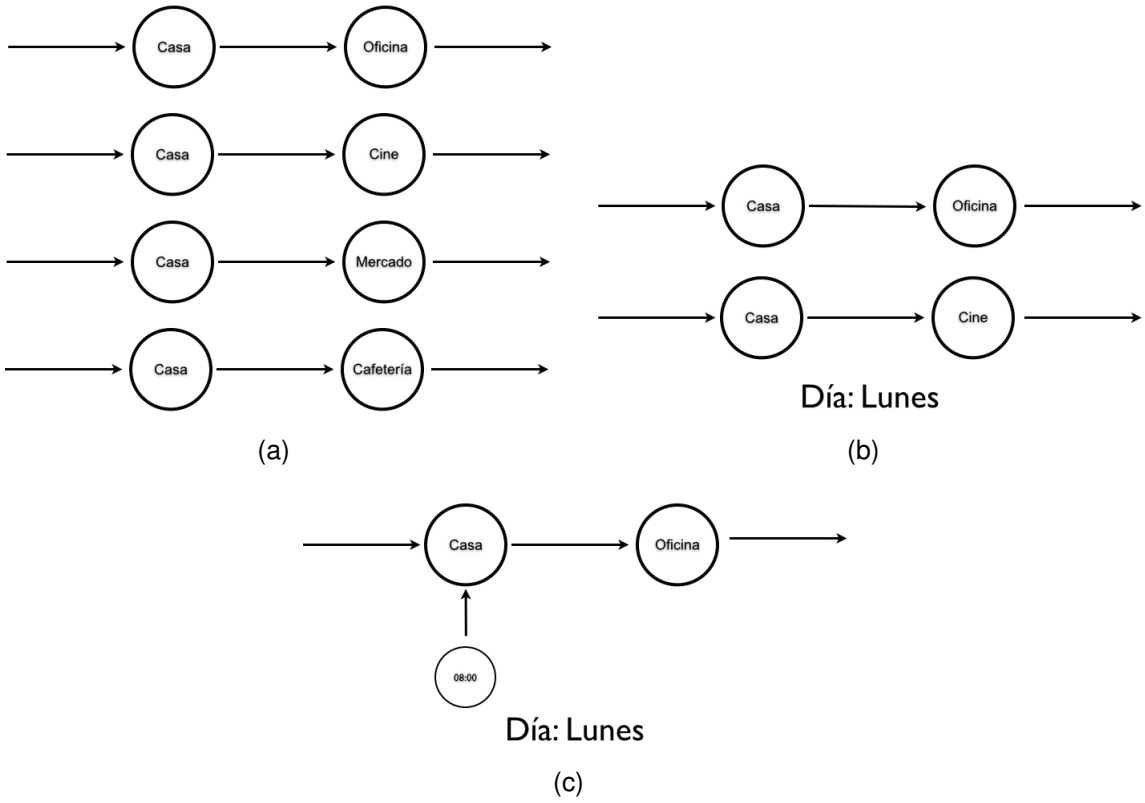


Figura 6: Uso del contexto en la predicción de la movilidad; a) Considerando sólo la ubicación actual, se tienen varias opciones como próxima ubicación. b) Considerando la ubicación actual y el día de la semana las opciones se acotan. c) Considerando la ubicación actual, el día de la semana, y la hora del día se obtiene una predicción más precisa.

Por lo tanto, la ***hipótesis de este trabajo es que una vez que los datos de localización se agrupan de acuerdo al día de la semana, la secuencia de los lugares visitados por el usuario forma una cadena de Markov*** (Figura 7). Al considerar esta hipótesis, es necesario aclarar que esta premisa es válida sólo cuando el usuario se traslada entre ciertos lugares. Estos lugares son aquellos en donde el usuario pasa un cantidad mínima de tiempo (i.e. umbral de tiempo t) en cada estadía, y además los visita frecuentemente. En los trabajos relacionados, estos lugares se conocen como lugares significativos, lugares de interés y puntos de interés (POI). En este trabajo, se utiliza el término *puntos de interés* para hacer referencia a dichos lugares.

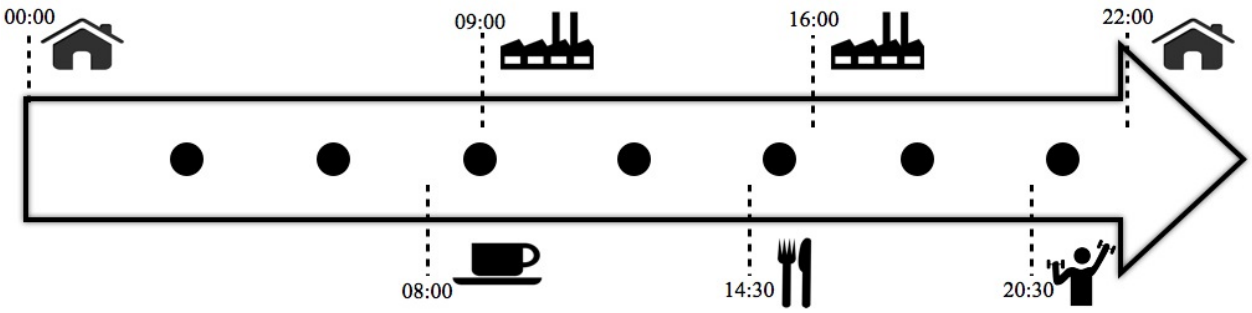


Figura 7: La próxima ubicación (POI) del usuario depende de la ubicación actual (POI) del mismo y la hora del día.

3.2.1. Puntos de interés en las actividades de los usuarios

Los puntos de interés son de suma importancia en el modelo de predicción espacio-temporal, ya que éstos se toman como referencia para realizar la predicción; no resulta factible considerar todos los lugares que el usuario visita, debido a que algunos de ellos sólo los visita de manera ocasional.

Al analizar las actividades cotidianas de los usuarios, se identifican aquellos lugares que son significativos o de interés para el usuario (Montoliu *et al.* (2013); Montoliu y Gatica-Pérez (2010); Ashbrook (2002); Ashbrook y Starner (2003); Kang *et al.* (2005)). Por ejemplo, considere el siguiente escenario: las actividades de Juan en un lunes determinado inician en su hogar. Posteriormente, Juan va a la cafetería y pasa ahí varios minutos; luego, va al supermercado y finalmente llega a su trabajo alrededor de las 9:00 AM. Después de algunas horas, Juan sale a comer a un restaurante cercano, para luego regresar al lugar de trabajo. Al término de su jornada laboral, Juan va a ejercitarse a un

gimnasio, y posteriormente se traslada a su casa. Al considerar la movilidad de Juan para el siguiente lunes, sus actividades no varían mucho; después de salir de su casa, Juan va nuevamente a la cafetería. Luego, antes de llegar a su lugar de trabajo llega a la lavandería. Alrededor de las 2:00 PM, Juan va al restaurante para luego regresar al trabajo. Varias horas después, Juan va al gimnasio, y finalmente llega a su casa.

Al considerar el escenario anterior, es posible identificar aquellos lugares que son POIs (Figura 8, recuadros negros), y aquellos lugares que el usuario sólo visita de manera ocasional (Figura 8, recuadros punteados). Además, se puede observar que la movilidad del usuario entre los POIs no cambia (casa - cafetería - trabajo - restaurante - trabajo - gimnasio - casa); sólo cambian los lugares que se visitan en la transición entre los POIs (lavandería y supermercado) (Figura 8).

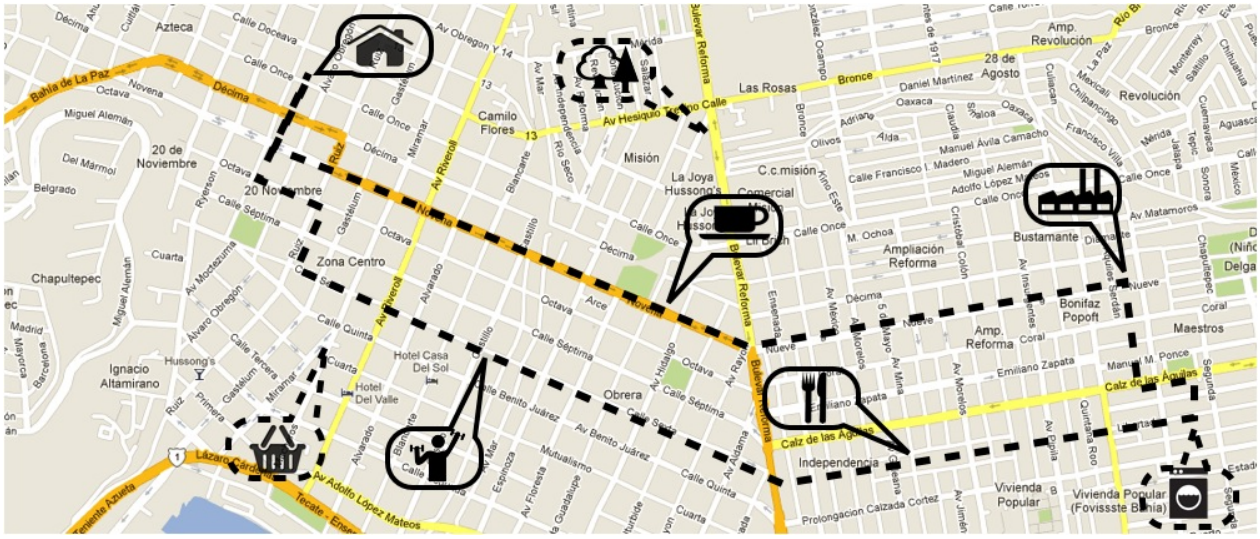


Figura 8: Varios puntos de interés se identifican en las actividades cotidianas del usuario.

3.2.2. Definiendo la movilidad del usuario como un modelo oculto de Markov

Al considerar que la movilidad del usuario entre los puntos de interés se puede describir como una cadena de Markov, y además que la estadía del usuario en cada uno de los puntos de interés se encuentra relacionada a la hora del día, y con el objetivo de definir el modelo de predicción espacio-temporal, resulta viable definir la movilidad del usuario como un modelo oculto de Markov (Rabiner (1989)) (Figura 9).

3.2.2.1. Modelos ocultos de Markov

Los modelos ocultos de Markov (HMM) son un modelo estadístico en el que se supone que el sistema a modelar es un proceso de Markov. Los HMM consisten de un conjunto de estados ocultos (Q), un conjunto de observaciones (O), probabilidades de transición (A), probabilidades de emisión (B), y probabilidades iniciales para cada estado oculto (π).

Los modelos ocultos de Markov adquieren el nombre de dos propiedades. Primero, se supone que la observación en el tiempo t fue generada por algún proceso Q_t, el cual se encuentra oculto para el observador. Segundo, se supone que el estado de este proceso oculto satisface la propiedad Markoviana: esto es, dado el valor de Q_{t-1} , el estado actual Q_t es independiente de todos los estados anteriores a Q_{t-1}. Cualquier estado en algún tiempo t contiene la información necesaria acerca de la historia del proceso para predecir el futuro del proceso. Una tercera suposición de los modelos ocultos de Markov es que la variable de estados ocultos es discreta; (Q_t) puede tomar K valores los cuales se denotan como $1, ..., K$.

La definición formal de un HMM se representa de la siguiente manera:

$$\lambda = (A, B, \pi) \tag{2}$$

Se define Q como la secuencia de estados de longitud T y O como las observaciones correspondientes:

$$Q = (q_1, q_2, ..., q_T) \tag{3}$$

$$O = (o_1, o_2, ..., o_T) \tag{4}$$

Se define la matriz de transición A, la cual define la probabilidad de que el estado j proceda al estado i. Estas transiciones son independientes del tiempo:

$$A = [a_{ij}], a_{ij} = P(q_t = s_j | q_{t-1} = s_i) \tag{5}$$

La matriz de observaciones B define la probabilidad de la observación k fue producida

por el estado j:

$$B[b_i(k)], b_i(k) = P(x_t = v_k | q_t = s_i) \tag{6}$$

Finalmente π representa el vector de probabilidades iniciales.

$$\pi = [\pi_i], \pi_i = P(q_1 = s_i) \tag{7}$$

Como se mencionó, en los HMM se realizan dos suposiciones. La primera, la propiedad Markoviana estipula que el estado actual depende únicamente del estado anterior, esto representa la memoria del modelo:

$$P(q_t | q_1^{t-1}) = P(q_t | q_{t-1}) \tag{8}$$

La premisa de independencia estipula que la salida de la observación en el tiempo t es dependiente sólo del estado actual, y por lo tanto es independiente de las observaciones y estados previos:

$$P(o_t | o_1^{t-1}, q_1^t) = P(o_t | q_t) \tag{9}$$

Un HMM es un proceso estocástico doble, compuesto de un proceso no observable, ***oculto*** (que representa un proceso de Markov), y un proceso observable. Se tiene por una parte una serie de coeficientes de probabilidad de transición que determinan la secuencia de estados que seguirá el modelo, y por otra parte, unas funciones de probabilidad asociadas a cada estado en particular que determinan la salida que se observará en ese estado. El primer proceso que no es directamente visible se puede observar/deducir a través del segundo proceso estocástico. Los modelos ocultos de Markov resultan útiles cuando el patrón que se desea encontrar no se puede describir por un proceso de Markov.

Una vez que un sistema se ha definido como un HMM, se pueden solucionar tres problemáticas:

- *Evaluación (Evaluation)*. Este enfoque se utiliza cuando se tienen diversos HMM describiendo diferentes sistemas. La idea es identificar el HMM que con mayor probabilidad generó una secuencia de observaciones. Este enfoque se utiliza principalmente en sistemas de reconocimiento del habla, donde se utilizan diversos modelos de Markov para modelar una palabra en particular.

- *Decodificación (Decoding)*. Este enfoque permite conocer la secuencia de estados ocultos que con mayor probabilidad generó una secuencia de observaciones dada. En muchos casos el interés radica en conocer los estados ocultos, ya que representan algo de valor que no es directamente observable.

- *Aprendizaje (Learning)*. Este enfoque permite generar un HMM a partir de una secuencia de observaciones; esto es, determinar el vector π, y las matrices A y B con mayor probabilidad de describir la secuencia de observaciones dada.

Una vez que se han descrito los modelos ocultos de Markov, la movilidad del usuario se define como un HMM de la siguiente manera: los estados ocultos se representan por el conjunto de puntos de interés. Por su parte, las observaciones las definen las diferentes horas del día o periodos de tiempo; la matriz de transición la define la movilidad del usuario entre los POIs (probabilidad de transición de un POI a otro), y la matriz de confusión la define la probabilidad de que el usuario se encuentre en cada uno de los POIs a determinadas horas del día (observaciones). De esta manera, el proceso oculto se encuentra definido por la transición del usuario entre los POIs.

3.2.3. Identificando la secuencia de lugares a visitar

Ya que el objetivo de este trabajo es predecir los lugares en los que estará el usuario en las próximas horas, y además predecir la hora en que estará en dichos lugares, resulta de particular interés el enfoque de *decodificación*. De esta manera, resulta factible considerar un conjunto de observaciones (e.g., 8:00, 9:00. 10:00, etc.) que representa un periodo de tiempo, y así determinar con una cierta probabilidad la secuencia de estados ocultos (secuencia de lugares) que corresponda a la secuencia de observaciones que se consideró. Por consiguiente, se conocen los lugares en los que estará el usuario (estados ocultos) y los tiempos de arribo a éstos (observaciones) (Figura 9).

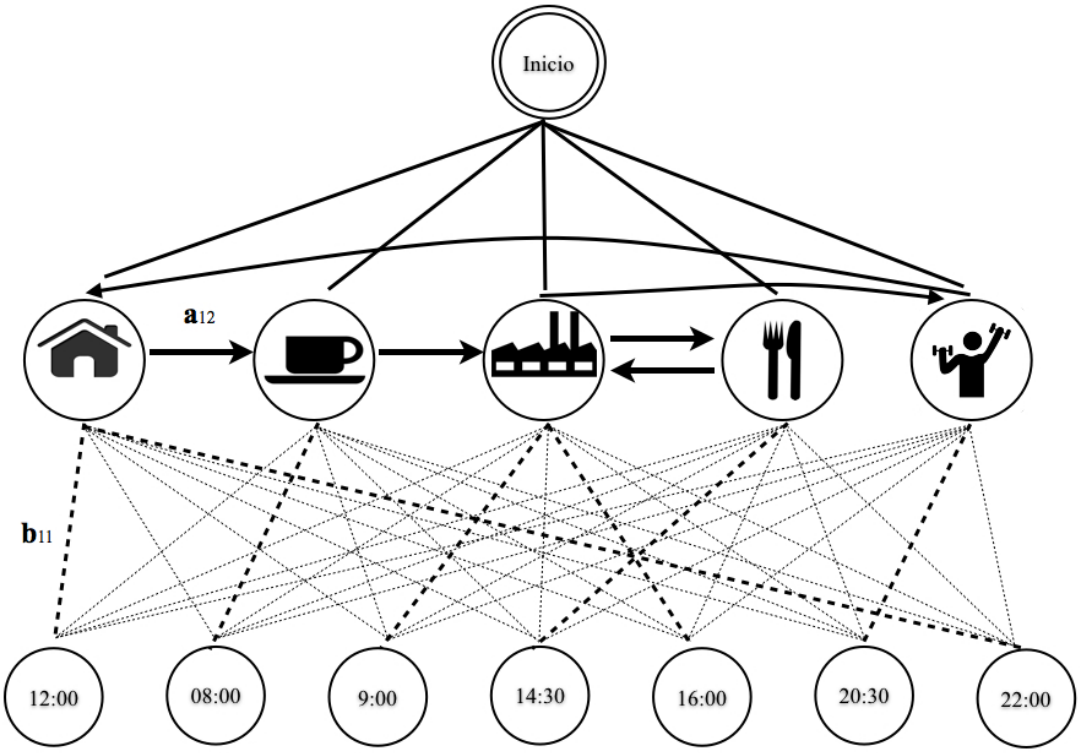


Figura 9: Definiendo la movilidad del usuario como un HMM para realizar la predicción espacio-temporal.

A fin de facilitar la tarea de identificar la secuencia de estados ocultos que con mayor probabilidad generó la secuencia de observaciones que se consideró, se utiliza el algoritmo de Viterbi. El algoritmo de Viterbi (Viterbi (2006)) funciona de la siguiente manera, primero se define:

$$\delta_t(i) = max_{q_1,q_2,\ldots,q_{t-1}} P(q_1 q_2 \ldots q_t = s_i, o_i, o_2 \ldots o_t | \lambda) \tag{10}$$

como la probabilidad de la trayectoria de estados más probable para la secuencia de observación parcial. Luego, el algoritmo de Viterbi funciona de la siguiente manera:

Inicialización:

$$\delta_1(i) = \pi_i b_i(o_1), 1 \leq i \leq N, \psi_{1(i)} = 0 \tag{11}$$

Recursion:

$$\delta_t s(i) = max_{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t), 2 \leq t \leq T, 1 \leq j \leq N \tag{12}$$

$$\psi_t(j) = argmax_{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right], 2 \leq t \leq T, 1 \leq j \leq N \tag{13}$$

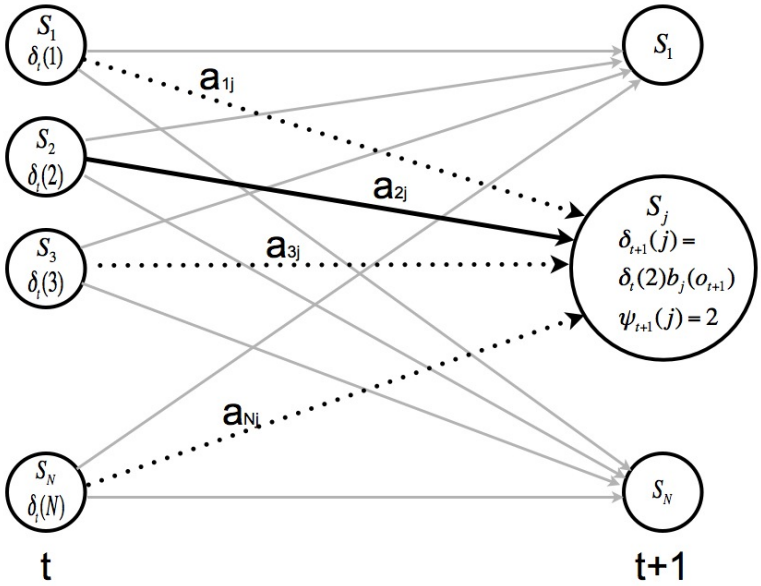


Figura 10: La recursión en el algoritmo de Viterbi.

Terminación:

$$P^* = max_{1 \leq i \leq N} \left[\delta_T(i)\right] \tag{14}$$

$$q_T^* = argmax_{1 \leq i \leq N} \left[\delta_T(i)\right] \tag{15}$$

Secuencia óptima de retroceso de estados:

$$q_T^* = \psi_{t+1} q_{T+1}^*, t = T-1, T-2, ..., 1 \tag{16}$$

El proceso de recursión se presenta en la Figura 10. Las probabilidades se maximizan y se almacena el estado que se eligió como el máximo para utilizarse como punto de retorno (*backpointer*). El retroceso (*backtracking*) permite encontrar la mejor secuencia de estados ocultos desde los puntos de retorno (*backpointers*) que se almacenaron en el proceso de recursión (Figura 11).

Después de representar la movilidad del usuario como un HMM y utilizar el algoritmo de Viterbi, resulta viable realizar la predicción espacio-temporal. Retomando el escenario

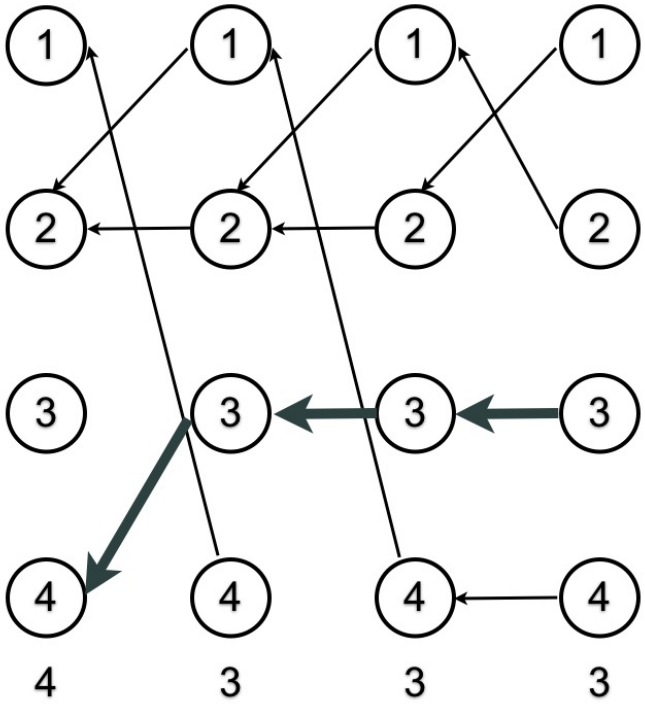


Figura 11: Proceso de retroceso en el algoritmo de Viterbi.

presentado, si el día es un lunes y la hora presente es 11:00 AM, y se desea conocer dónde estará Juan de las 11:00 a las 18:00 horas, se toma como referencia el vector π, la matriz de transición (A), y la matriz de confusión (B), para identificar la secuencia de lugares dónde Juan estará en el periodo de predicción considerado. En este caso, si el usuario cuenta con seis puntos de interés, y las observaciones se encuentran definidas cada 1:30 horas, el periodo de predicción considerado abarca las observaciones: 11:30, 13:00, 14:30, 16:00 y 17:30. El resultado de la predicción para dicho periodo sería que a las 11:30 Juan estará en la oficina, de igual manera a las 13:00; a las 14:30 Juan estará en el restaurante, y a las 16:00, como a las 17:30, estará nuevamente en la oficina (Figura 12).

3.3. Aspectos a considerar en el modelo de predicción

Con lo que se discute hasta el momento, se tiene conocimiento de cómo caracterizar la movilidad a fin de definir el modelo de predicción espacio-temporal utilizando los modelos ocultos de Markov. Sin embargo, más allá de conocer el método a utilizar, la definición del modelo de predicción conlleva otros factores que son determinantes para obtener una predicción precisa, y definir un modelo de predicción robusto.

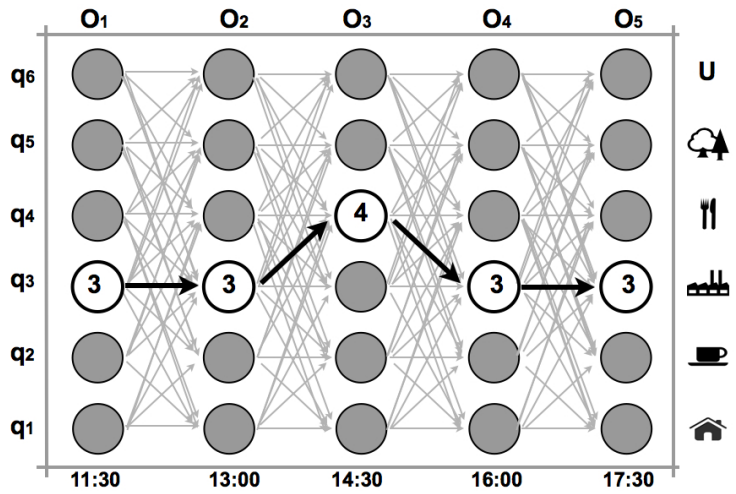


Figura 12: Uso del algoritmo de Viterbi para identificar la secuencia de POIs donde estará el usuario en un periodo de tiempo dado.

Entre estos factores destacan la identificación de puntos de interés, definición de las observaciones, entrenamiento del modelo de predicción, la predecibilidad y dinamicidad de la movilidad del usuario, y la cantidad necesaria de datos para definir el modelo de predicción.

Al analizar los trabajos relacionados (Subsección 2.2), se encontró que éstos no contemplan los aspectos mencionados; estos aspectos son determinantes para obtener una precisión mayor a la que se presenta en los trabajos relacionados. En la Figura 13 se presentan los aspectos que se consideran en este trabajo. Cada uno de ellos ha representado un reto en diversos trabajos relacionados y, actualmente, son motivo de estudio. A continuación se realiza una descripción de cada uno de ellos.

3.3.1. Puntos de interés

3.3.1.1. ¿Qué es un punto de interés?

Con el objetivo de evitar una confusión futura con el uso del término punto de interés, se da una explicación al respecto. En este trabajo, y en los trabajos relacionados se utilizan varios términos (e.g., punto de interés, lugar significativo, punto de estadía) para

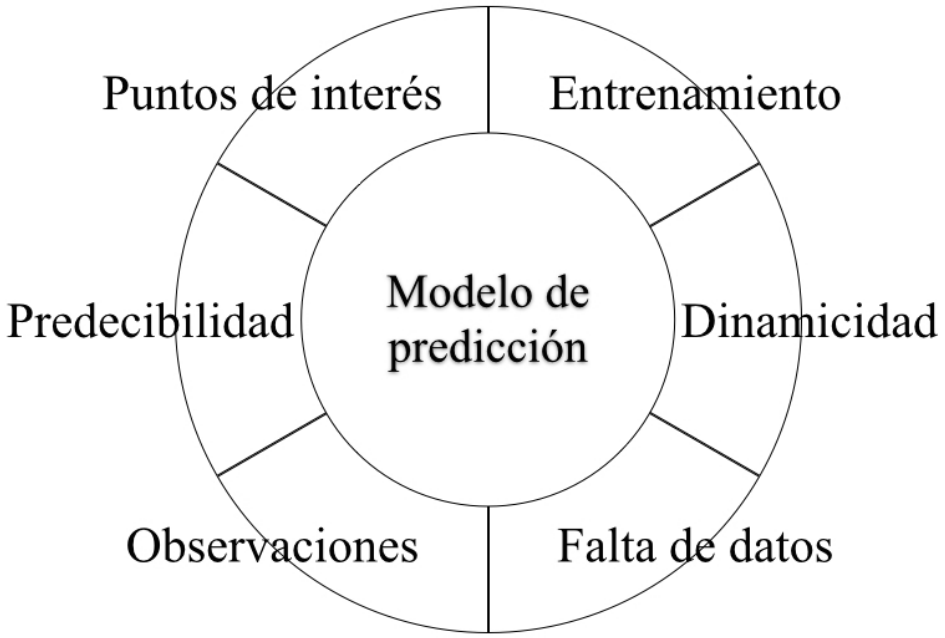


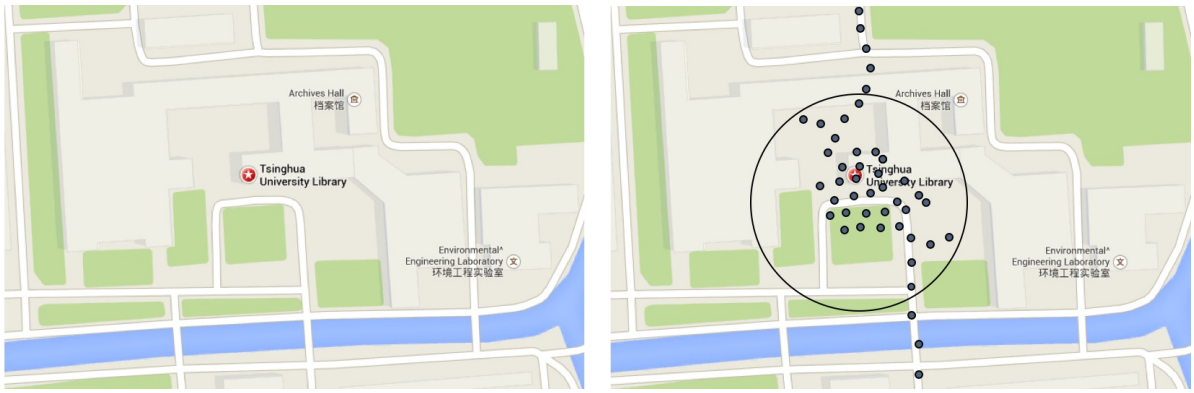
Figura 13: Aspectos a considerar en la definición del modelo de predicción.

describir el área o ubicación geográfica que es importante en las actividades cotidianas de los usuarios.

Para realizar la descripción de los puntos de interés se utiliza el significado semántico de éstos. Por lo tanto, se hace referencia al hogar, oficina, entre otros, para hacer alusión a los lugares que son importantes para el usuario. Sin embargo, de manera práctica, los puntos de interés representan un área geográfica que abarca o incluye el lugar que es significativo para el usuario.

Por ejemplo, al considerar la Figura 14(a), un lugar significativo para un usuario dado sería la biblioteca de la universidad; sin embargo, los algoritmos consideran el área geográfica circundante donde el usuario estuvo en un periodo de tiempo (Figura 14(b), área definida por el circulo). La granularidad del área geográfica depende del tamaño del clúster que se utilice para identificar dichos puntos de interés. O bien, cuando se utilizan los registros de conexiones a puntos de acceso, el punto de interés lo define el identificador del punto de acceso (*SSID*); el punto de acceso se encuentra en la periferia del lugar que es de interés para el usuario.

Sólo en aquellos casos en los cuales se utilizan datos de los servicios basados en



(a) (b)

Figura 14: Significado del puntos de interés; a) punto de interés: biblioteca; b) área geográfica que abarca el punto de interés.

localización, los puntos de interés se identifican y se etiquetan de acuerdo al significado semántico de éstos; sin embargo, el aspecto semántico no se considera en este trabajo debido a la falta de datos de esta índole.

El primer aspecto a considerar en el modelo de predicción es el identificar los puntos de interés. Estos lugares se tomarán como referencia para realizar la predicción espacio-temporal. Los lugares significativos o puntos de interés (POI) tienen un rol fundamental en el modelo de predicción propuesto. La identificación correcta de estos lugares define en gran medida la eficiencia del modelo. De manera general, un POI es aquel lugar en el que el usuario pasa un tiempo mínimo determinado y visita frecuentemente.

3.3.1.2. Identificación de puntos de interés

Identificar los puntos de interés de un usuario es un tema de investigación importante en diferentes áreas, como cómputo ubicuo, planeación urbana, salud, entre otras.

Actualmente, los dispositivos móviles producen una gran cantidad de datos de ubicación que son útiles para descubrir los lugares donde el usuario pasa su tiempo (Chen y Kotz (2000); Korpipaa *et al.* (2003)). Por ejemplo, al considerar la Figura 15 se observa que actualmente es posible obtener datos de localización a partir de las redes sociales, los sistemas basados en localización (LBS, por sus siglas en inglés), de las interfaces de comunicación (GSM, Wi-Fi, Bluetooth), y del GPS. De esta manera, existe la necesidad de definir algoritmos que permitan convertir la gran cantidad de datos crudos en

información útil, y así identificar los lugares que son significativos para el usuario.

Al momento, diversos trabajos se han enfocado en identificar estos lugares utilizando diversos enfoques de acuerdo al propósito de la aplicación. De manera general estos trabajos se pueden clasificar como:

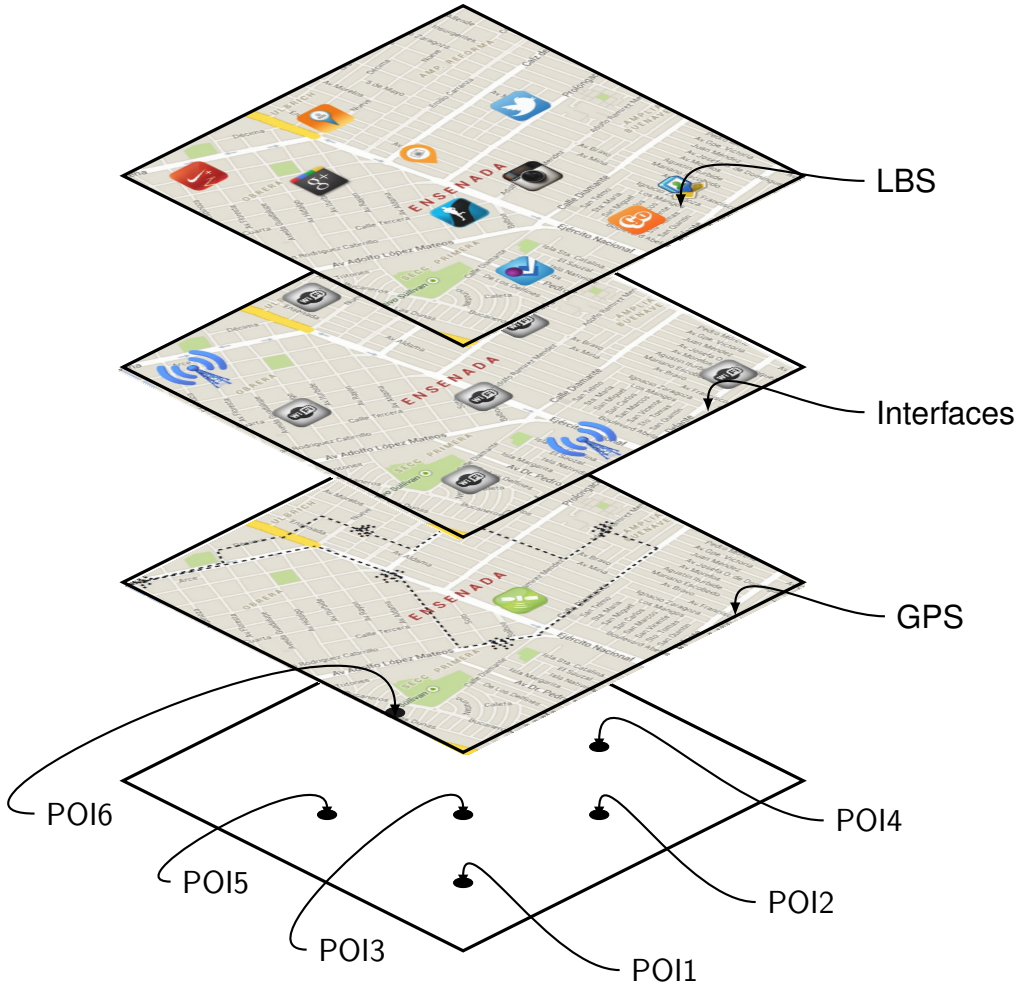


Figura 15: Identificación de puntos de interés.

- Algoritmos basados en el tiempo de estadía. Este enfoque se basa en la premisa de que la importancia de un lugar es directamente proporcional al tiempo de estadía en él (Kang *et al.* (2005); Kim *et al.* (2006); Scellato *et al.* (2011); Ye *et al.* (2009)).
- Algoritmos basados en densidad. Estos algoritmos se basan en la densidad de registros GPS, o bien de registros de localización dentro de un área determinada (Ram *et al.* (2010); Zhou *et al.* (2007)).

- Algoritmos basados en la pérdida de la señal. Debido a la pérdida de la recepción de la señal GPS cuando el usuario se encuentra en interiores, este tipo de algoritmos toman como referencia la pérdida y la reaparición de la señal GPS para identificar potenciales puntos de interés (Ashbrook y Starner (2003); Marmasse y Schmandt (2000)).

Cada uno de los enfoques mencionados es útil en aplicaciones específicas; sin embargo, la funcionalidad de éstos no satisface el propósito del presente trabajo; no resulta factible considerar sólo uno de estos enfoques para identificar los puntos de interés. Por ejemplo, un lugar que fue visitado por un usuario en varias ocasiones no se puede considerar como punto de interés, ya que las visitas fueron realizadas hace varios meses/semanas anteriores a la fecha de interés. De igual manera, un lugar en el cual un usuario pasa un tiempo prolongado no se puede considerar como de interés, si no se visita una cantidad mínima de ocasiones.

Debido a lo anterior, en este trabajo se toman como referencia los *datos de localización y tiempo (aspectos espacial y temporal), la frecuencia de visitas, y el periodo de tiempo que abarca el patrón de movilidad actual del usuario*. La identificación de los puntos de interés considera los siguientes factores:

- Tiempo de estadía.
- Frecuencia de visitas.
- Tamaño del clúster. Este aspecto define el área geográfica que comprende el punto de interés.
- Periodo del patrón de movilidad (tamaño de la ventana). Este aspecto define la cantidad de datos que serán utilizados para identificar el punto de interés.

Al considerar los aspectos anteriores, se identifican aquellos lugares que actualmente (a la fecha de interés) son significativos en las actividades cotidianas del usuario. Ya que los usuarios realizan sus actividades cotidianas tanto en lugares en interiores como en

exteriores, se aplican los cuatro aspectos anteriores a los algoritmos de dos trabajos relacionados para identificar puntos de interés en interiores (Ashbrook y Starner (2003)) y exteriores (Kang *et al.* (2005)).

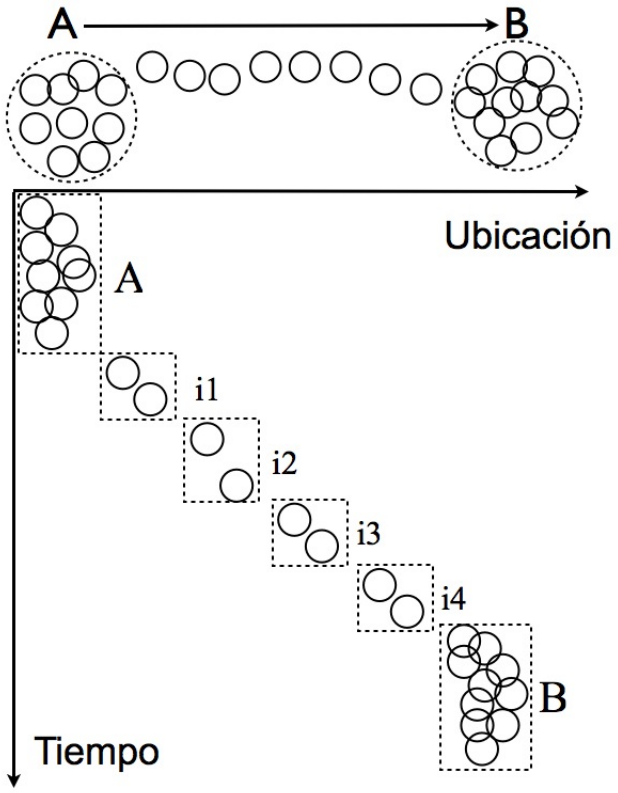


Figura 16: Identificación de puntos de interés mediante el algoritmo de Kang *et al.* (2005).

El algoritmo que presenta Ashbrook y Starner (Ashbrook y Starner (2003)) permite identificar lugares significativos en interiores. Para ello, el algoritmo se basa en la pérdida de la señal GPS dentro de un radio dado r y toma como referencia un umbral de tiempo t para la pérdida de la señal.

Por su parte, Kang *et al.* (2005) proponen un algoritmo basado en el tiempo de estadía para identificar los lugares significativos en exteriores. Este algoritmo compara cada registro de GPS con los registros previos en un clúster dado; si los registros GPS se alejan del clúster actual, se forma uno nuevo. Kang *et al.* (2005) definen dos umbrales t y r para el tiempo de estadía mínimo y el radio del clúster, respectivamente. Por ejemplo, considerando la Figura 16, un usuario se traslada desde un lugar A a un lugar B. Mientras el usuario se encuentra en el lugar A durante al menos t minutos, y los registros GPS se encuentran cercanos (a una distancia d entre ellos), se forma el clúster A. Cuando el

usuario se traslada hacia el lugar B, los registros de localización se alejan del clúster A, y se generan varios clústers intermedios (i1, i2, i3, i4, i5); éstos se descartan si el usuario pasa menos de t minutos en ellos. Posteriormente, al llegar al lugar B, el usuario pasa el tiempo suficiente en este lugar, y se forma el clúster B. El pseudo código del algoritmo se presenta a continuación:

Algoritmo 1 Obtener puntos de interés

```
Input: measured location loc
current cluster cl
pending locations plocs
significant places Places
si distance(cl, loc) < d entonces
    add loc to cl
    clear plocs
si no
    si plocs.length > l entonces
        si duration(cl) > t entonces
            add cl to Places
        fin si
        clear cl
        add plocs.end to cl
        clear plocs
        si distance(cl, loc) < d entonces
            add loc to cl
            clear plocs
        si no
            add loc to plocs
        fin si
    si no
        add loc to plocs
    fin si
fin si
```

Cuando un clúster se agrega al conjunto de lugares significativos (Places), el algoritmo verifica la condición de fusión: si el centroide del clúster se encuentra a menos de una distancia $d/3$ de un lugar existente (Place), el clúster se fusiona con este lugar; en caso contrario, se agrega un nuevo lugar (Place).

Los cuatro factores (tiempo de estadía, frecuencia de visitas, tamaño de cluster y periodo del patrón de movilidad) permiten identificar aquellos lugares donde el usuario ha estado un tiempo determinado y visita frecuentemente. Así también, al considerar el ta-

maño del radio del clúster se identifican puntos de interés a diferente granularidad. Por ejemplo, la casa del usuario se puede considerar como un POI granular, y el centro comercial como un POI a nivel zona. De esta manera, es posible realizar la predicción de la ubicación del usuario a varios niveles de granularidad. Finalmente, la última característica es muy importante debido a que permite identificar aquellos lugares que son significativos para el usuario en una fecha determinada.

3.3.2. Definir las observaciones

Además de identificar de manera adecuada los puntos de interés, otro aspecto importante al realizar la predicción espacio-temporal reside en definir el conjunto de observaciones, esto es, definir los tiempos específicos que se tomarán como referencia para realizar la predicción. Por ejemplo, para definir las observaciones del modelo de predicción de un estudiante para los días lunes, éstas pueden tomar los siguientes valores: 8:00, 14:00, 17:00, 20:00, 22:00, considerando que el estudiante regularmente llega a la escuela a las 8:00 AM, a las 14:00 tiene un receso y sale de la escuela. Luego, a las 17:00 el usuario regresa a la escuela, sale a las 20:00 horas, y finalmente llega a su casa a las 22:00.

En cada uno de los experimentos se describe cómo se definieron las observaciones.

3.3.3. Entrenamiento del modelo de predicción

Después de tener conocimiento de cómo definir el modelo de predicción, identificar los puntos de interés, y definir las observaciones, otro aspecto de suma importancia es cómo realizar el entrenamiento del modelo de predicción. Esto es, determinar la cantidad de datos necesaria para definir/entrenar cada uno de los componentes del modelo de predicción. Por ejemplo, tomando en cuenta que el usuario exhibe un comportamiento diferente en cada día de la semana, si el día actual es un lunes i, el reto se encuentra en determinar la cantidad de lunes previos que son necesarios para definir el modelo de predicción sin considerar datos asociados a otro patrón de movilidad.

Para el modelo de predicción propuesto no resulta adecuado considerar una cantidad arbitraria de datos como lo hacen algunos trabajos relacionados (Mathew *et al.* (2012); Scellato *et al.* (2011)). En el caso de tomar como referencia los datos asociados a un par

de días previos (e.g., 2 últimos lunes), es posible que el modelo no refleje el comportamiento del usuario, en cambio al considerar los datos asociados a un periodo prolongado de tiempo (e.g., 40 últimos lunes), se tiene el problema de considerar los datos asociados a diferentes patrones de movilidad, y así, la definición del modelo de predicción tiene un grado de error.

Debido a lo anterior, la opción que se contempló en este trabajo es utilizar los datos de localización asociados al patrón de movilidad actual del usuario (a la fecha de interés). De esta manera, tanto la identificación de los puntos de interés, como la definición de los componentes de HMM, se realiza con los datos del periodo de tiempo en el cual el usuario exhibe un comportamiento similar. Por lo tanto, el reto se encuentra en identificar el periodo de tiempo que comprende el patrón de movilidad más reciente del usuario.

A fin de identificar el patrón de movilidad del usuario, se compara la movilidad del usuario día con día. La idea es identificar día con día los lugares que visita el usuario y los tiempos de estadía en éstos, a fin de conocer el periodo de tiempo en el cual la movilidad del usuario es similar. Para ello, la movilidad diaria del usuario se representa como un arreglo de N posiciones ($V = [i_1, i_2, I_3, ..., i_N]$) donde cada posición i representa un periodo de tiempo determinado en el día (e.g., 1 hora), y el valor de cada índice i se encuentra determinado por un identificador asociado al lugar en el que estuvo el usuario en dicho periodo de tiempo.

Por ejemplo, al considerar la Figura 17, se observa que en la semana 1 (e.g., lunes 1), el usuario estuvo en un lugar definido por el identificador 1 en el periodo de las 7:30 a las 7:59; posteriormente, en el lugar con identificador 2 en el periodo de las 9:00 a las 9:29; finalmente, en el lugar con identificador 1 en el periodo de las 23:30 a las 23:59. El lunes previo, el usuario tuvo un comportamiento similar en cuanto a los lugares y los tiempos de estadía. Así también, para el lunes anterior (semana 3). Por lo tanto, para el usuario en cuestión se tienen los arreglos $V1 = [..., 1, 0, 0, 2, ..., 1], V3 = [..., 0, 0, 0, 2, ..., 1]$ y $V3 = [..., 0, 0, 2, 2, ..., 1]$. Al definir la movilidad diaria del usuario como un arreglo V, se utiliza la similitud entre estos arreglos para determinar la cantidad de arreglos que son similares ($sim(V, U)$), y así determinar el periodo de tiempo que abarca el patrón de movilidad más reciente del usuario. Una vez que se ha identificado este periodo de

tiempo, se procede a identificar los puntos de interés y definir el modelo de predicción. En el siguiente capítulo se explica a detalle este procedimiento.

Time / Week	...	07:30	08:00	08:30	09:00	...	23:30
1	...	1	0	0	2		1
2	...	0	0	0	2		1
3	...	0	0	2	2		1

Figura 17: La movilidad del usuario como un arreglo.

3.3.4. Predecibilidad de la movilidad del usuario

Una de las premisas de este trabajo reside en el hecho de que la movilidad del usuario entre los POIs se puede representar como una cadena de Markov. Al considerar esta premisa, la movilidad del usuario se representó como un modelo oculto de Markov para definir el modelo de predicción; la eficiencia de los modelos ocultos de Markov depende de la existencia de la propiedad Markoviana. En el caso de que la movilidad del usuario entre los puntos de interés no exhiba la propiedad Markoviana, los HMM no resultan adecuados para realizar la predicción espacio-temporal, y por consiguiente se deberá actuar en consecuencia.

Debido a ello, es necesario corroborar la existencia de la propiedad Markoviana para así definir la movilidad del usuario como un HMM, y por consiguiente obtener una predicción precisa. Para identificar la existencia de la propiedad Markoviana, se utiliza una prueba empírica definida por Zhang *et al.* (2010). En el capítulo cuatro se describe el funcionamiento del método.

3.3.5. Dinamicidad de la movilidad del usuario

Hasta este punto se tiene la suposición de que la movilidad del usuario es estacionaria a lo largo del tiempo; sin embargo, de manera realista, los patrones de movilidad del usuario varían a lo largo del tiempo, y con ello los lugares que son significativos para el

usuario, y/o los tiempos de estadía en los lugares que el usuario visita (Li *et al.* (2010); Farrahi y Gatica-Pérez (2011)).

A lo largo del tiempo, sólo ciertos lugares como el hogar o el trabajo son significativos; no es común que los usuarios cambien su hogar o trabajo en periodos cortos de tiempo. En cambio, las preferencias de los usuarios varían con el tiempo, y así, otros lugares como restaurantes, cines, lugares turísticos, etc., son significativos sólo en determinados periodos de tiempo.

Por ejemplo, considérese que un usuario dado está iniciando clases de natación los días lunes en una alberca pública. Por varias semanas, el usuario toma sus clases los días lunes. Sin embargo, el curso de natación tiene una duración de tan sólo 8 semanas. Así, después de este periodo, la movilidad del usuario cambiará. Otros ejemplos realistas incluyen las actividades de un estudiante; éste tiene varios patrones de movilidad que corresponden a diferentes periodos escolares (semestral, cursos de verano, cuatrimestral), o bien a periodos vacacionales (e.g., verano, navidad, pascua). O en el caso de un trabajador que labora por turnos, sus actividades cambian cada determinado tiempo.

Al considerar lo anterior, no es adecuado contar con un modelo de predicción estático para realizar la predicción de la ubicación del usuario a lo largo del tiempo, ya que, conforme la movilidad cambia, el modelo de predicción no reflejaría dicho cambio. Por lo tanto, es necesario actualizar los parámetros del modelo de predicción a fin de caracterizar de manera precisa la movilidad del usuario. Así, para diseñar el modelo de predicción se debe de responder a la siguiente incógnita: *¿cómo puede el modelo de predicción propuesto adaptarse a los cambios en la movilidad del usuario?*.

Para estudiar este aspecto, una opción viable es utilizar el mecanismo que se discute en la Subsección 3.3.3. De esta manera, se identifica hasta qué punto el usuario exhibe un comportamiento similar, y así se define el modelo de predicción con los datos correspondientes a dicho periodo. Sin embargo, con este enfoque se identifican dos limitantes:

- Al considerar la similitud entre la movilidad del día más reciente (e.g., último lunes) con la movilidad del usuario en los días anteriores (e.g., n lunes), el modelo de

predicción se actualiza si la similitud es mayor a un cierto umbral. Sin embargo, si el patrón de movilidad del usuario cambia de manera gradual, los datos más antiguos no necesitan incorporarse. De esta manera, la actualización del modelo de predicción no sólo contempla la incorporación de datos, sino también eliminar datos innecesarios.

- Al cambiar el patrón de movilidad, es necesario definir un nuevo modelo de predicción, y por consiguiente se requiere contar con una cantidad determinada de datos a fin de definir de manera adecuada el nuevo modelo de predicción. Por ejemplo, al considerar el hecho de que un usuario cambia de ciudad de residencia, o de lugares de interés. Por lo tanto, es necesario recolectar una cierta cantidad de datos antes de definir el modelo de predicción.

Debido a lo anterior, en este trabajo se propone utilizar una ventana deslizante. A partir del uso de esta ventana, el modelo de predicción toma como referencia la movilidad más reciente del usuario. Esto es, considerando que se desea realizar la predicción para un día (d_i), el modelo de predicción se entrena con los datos correspondientes a una ventana de longitud determinada (n) (e.g., las últimas 4 semanas $[d_{i-1-n}, d_{i-1}]$). Después de que se han realizado las predicciones para el día (d_i), la ventana se desplaza; el modelo de predicción se entrena nuevamente (Figura 18, conjunto de entrenamiento 2) y se realizan las predicciones para el día (d_{i+1}), y así sucesivamente, como se muestra en la Figura 18.

Al considerar la ventana deslizante, se evita el problema de definir diferentes modelos de predicción para cada uno de los patrones de comportamiento; se mantiene un modelo de predicción a lo largo del tiempo, el cual considera el comportamiento más reciente del usuario. También se evita el problema de determinar cuándo es necesario eliminar datos de la movilidad del usuario.

Además del enfoque de la ventana deslizante, una opción alterna es que explícitamente el usuario especifique el cambio de comportamiento o patrón de movilidad, o bien que el usuario seleccione de un conjunto de patrones de movilidad identificados (vacaciones, navidad, periodo escolar, etc.), aquel que sea apropiado a la temporada. La segunda opción representa un reto y es motivo de investigación.

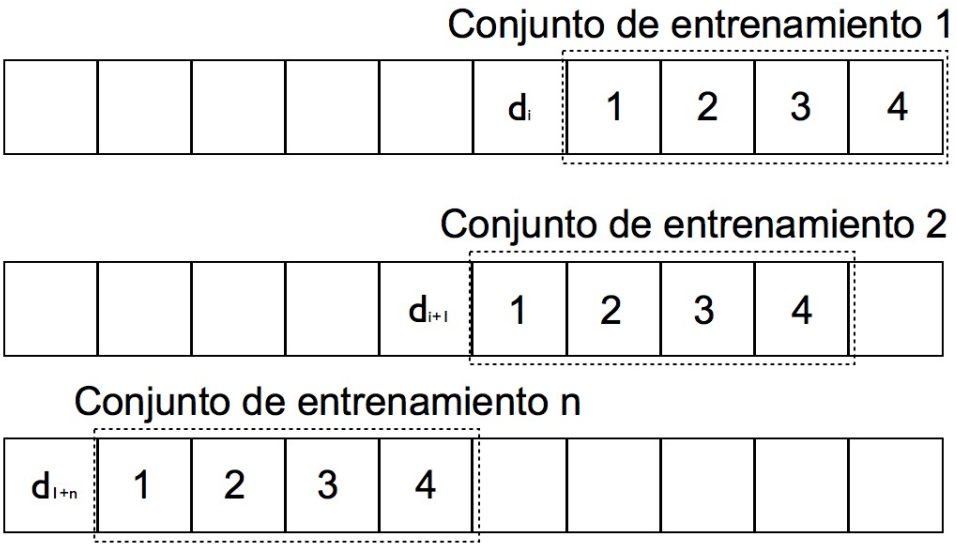


Figura 18: Utilizando una ventana deslizante para actualizar el modelo de predicción.

3.3.6. Complementando la movilidad del usuario con las preferencias colectivas

Hasta el momento, mediante el enfoque propuesto es posible identificar los puntos de interés, modelar la movilidad del usuario entre los puntos de interés como un modelo oculto de Markov, y además el modelo de predicción se actualiza a lo largo del tiempo con los datos correspondientes a la ventana deslizante, así se capturan los cambios en la movilidad del usuario. Sin embargo, para realizar estas acciones se requiere contar con datos de la localización del usuario, lo cual no siempre es posible. Esta limitante se debe principalmente a tres factores:

- Problemas de infraestructura y tecnología. Aunque actualmente una gran cantidad de lugares ofrecen conectividad a internet, los usuarios no cuentan con permiso para conectarse a todos los puntos de acceso. Por lo tanto, si se utilizan datos de conectividad a puntos de acceso para conocer la movilidad del usuario, sólo se tiene conocimiento parcial de la movilidad de éste. Así también, al utilizar la tecnología GPS, ésta conlleva ciertas limitantes; la falta de funcionalidad en interiores, y las condiciones atmosféricas complican la colecta de datos.

- Capacidad del dispositivo móvil. Otro factor a considerar es la energía de los dispositivos móviles. Al recolectar datos de localización a través de múltiples tecnologías, el tiempo de vida de la batería se reduce a un par de horas, y al considerar que los

dispositivos cumplen diferentes funciones para el usuario, éste administra el uso de la batería a lo largo del día. Por lo tanto, al igual que en el caso anterior, sólo es posible recolectar datos de localización por un tiempo limitado.

- Privacidad del usuario. El último aspecto a considerar es la privacidad del usuario. No todos los usuarios están dispuestos a proveer sus datos de localización, explícitamente los usuarios pueden bloquear la colecta, o bien restringirla.

Debido a los aspectos anteriores, la colecta o sensado de los datos de localización no se realiza de manera continua, y como consecuencia se tienen datos escasos para realizar el análisis de la movilidad del usuario. Debido a ello, no es posible tener un entendimiento completo de la movilidad de los usuarios, y por lo tanto, el modelo de predicción resultará inadecuado.

Aunado a lo anterior, el modelo de predicción propuesto toma como referencia un conjunto de puntos de interés para realizar la predicción del usuario. Estos lugares requieren contar con una cantidad mínima de visitas a fin de considerarse puntos de interés; al no contar con datos de localización, estos lugares no cuentan con las visitas requeridas, y por consiguiente, no es posible identificar en su totalidad los puntos de interés e información asociada a éstos (tiempos de llegada y salida a los puntos de interés). Así, la incógnita que el diseñador del modelo de predicción estudia es: ¿cómo compensar la falta de datos de localización a fin de evitar la omisión de puntos de interés y así definir un mejor modelo de predicción?, y por ende realizar una predicción acertada de la movilidad del usuario.

A fin de solucionar este problema, en este trabajo se toman como referencia a los usuarios que son similares en cuanto a los puntos de interés que éstos visitan. Partiendo de la premisa de que hay usuarios que comparten ciertos lugares ya sea por cuestiones de amistad, trabajo, u otros, para un usuario en particular es posible evitar la omisión de puntos de interés al considerar aquellos lugares que visitan los usuarios similares a él. De esta manera, resulta factible agregar como POIs aquellos lugares que no se habían identificado como lugares significativos en una primera instancia.

La falta de datos es un fenómeno común en conjuntos de datos (*dataset*) que se recolectan sin restricción alguna. Ejemplos de la falta de datos se puede encontrar en diversos conjuntos de datos públicos: por ejemplo, el conjunto de datos de Dartmouth (registros de Wi-Fi, (Kotz *et al.* (2007a))), Geolife (registros de GPS)(Zheng *et al.* (2008, 2009, 2010)) y Monarca (registros de acelerómetro, GPS, Wi-Fi, etc.) (Bardram *et al.* (2013)); éstos sólo por citar algunos.

3.3.6.1. El impacto de las preferencias colectivas en las actividades de un usuario

Generalmente, las decisiones de los usuarios se encuentran influenciadas por las preferencias de otros (colectividad); se compra un producto porque algún conocido lo recomendó; se mira una película porque los amigos la miraron. De igual manera, las preferencias del usuario tienen influencia sobre otras personas.

Debido a ello, ciertas personas son similares al considerar los productos que compran, las películas que ven, o los gustos que comparten. Este aspecto también es válido al considerar los lugares que se visitan; por lo regular, se visita un restaurante porque algún amigo invitó; se asiste a un determinado parque porque las amistades realizan ejercicio ahí (Zheng *et al.* (2010); Calabrese *et al.* (2010); Zheng *et al.* (2011); Adomavicius y Tuzhilin (2005a); Karypis (2001); Sarwar *et al.* (2001); Bellotti *et al.* (2008)). Además, existe similitud a nivel espacial entre los usuarios que tienen lugares en común; debido a las actividades escolares, laborales o recreativas (Li *et al.* (2008); Lee y Chung (2011); Ying *et al.* (2010)). Por lo tanto, para conocer qué lugares ha visitado un usuario dado, resulta viable considerar los lugares que han visitado los usuarios que son similares a éste, y por consiguiente evitar la omisión de puntos de interés.

3.3.7. Filtrado colaborativo

Para ello, se utilizan los principios de una técnica que se emplea en los sistemas de recomendación, denominada filtrado colaborativo (*collaborative filtering*). El filtrado colaborativo (CF por sus siglas en Inglés) es un método que surgió en las aplicaciones de comercio electrónico para realizar recomendaciones personalizadas a los usuarios (Schafer *et al.* (1999)).

Este método se basa en la premisa de que, personas que tienen los mismos gustos o preferencias acerca de un conjunto de elementos, es probable que tengan las mismas preferencias sobre otros elementos; si un grupo de usuarios tienen los mismos gustos que Juan, entonces es probable que Juan tendrá los mismos gustos que los otros usuarios en elementos que Juan aún no conoce.

Este método resulta ser eficaz para identificar nuevos productos para los usuarios. CF combina las preferencias de los usuarios que han expresado gustos similares a los de un usuario dado para realizar una predicción para el usuario considerado. Algunos ejemplos conocidos del uso de CF son Amazon, iTunes Netflix, LastFM, StumbleUpon, Reddit, YouTube, Digg y Delicious,

El dominio de información para los sistemas de CF consiste en usuarios que han expresado sus preferencias acerca de varios elementos. La preferencia de un usuario por un elemento se denomina puntuación (*rating*), y se representa como una tupla (*User, Item, Rating*). Estas puntuaciones pueden tomar diferentes formas de acuerdo al dominio de aplicación.

El conjunto de tuplas de puntuación forman una matriz, denominada la matriz de puntuación (*rating matrix*). La matriz de puntuación se define como R, donde $r_{u,i}$ es la puntuación que el usuario u otorgó al elemento i; r_u es el arreglo de todas las puntuaciones que ha otorgado el usuario u, y r_i es el arreglo de todas las puntuaciones otorgadas al elemento i (Figura 19). Generalmente, el filtrado colaborativo se puede categorizar como basado en usuarios, o basado en elementos.

Filtrado colaborativo basado en usuarios. La recomendación de elementos a un usuario dado se basa en los elementos que son de interés para los usuarios similares a éste (Figura 20). Por ejemplo, si un usuario *A* tiene la misma opinión que un usuario *B* acerca de un elemento *x*, es probable que el usuario *A* tenga la misma opinión que el usuario *B* para un elemento *y* (Figura 20). Bajo este enfoque, Netflix utiliza una evaluación de sus películas asignando entre 1 a 5 estrellas; basados en esas evaluaciones, Netflix también ha creado un algoritmo que infiere qué películas le gustarán a un cliente con base en la evaluación de otros clientes y la semejanza que el cliente que recibe la recomendación

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_i	I_N
U_1	$r_{1,1}$	$r_{1,2}$	$r_{1,3}$						...	$r_{1,N}$
U_2									...	
U_3									...	
U_4									...	
U_5									...	
U_6									...	
U_7									...	
U_8	$r_{8,1}$								...	
U_i	...	...	...	...	...	...	...	...	...	...
U_M	$r_{M,1}$								...	$r_{M,N}$

Figura 19: Matriz de puntuación *R* presentando las puntuaciones que *M* usuarios tienen con respecto a *N* elementos.

tiene con ellos.

Filtrado colaborativo basado en elementos. Es la técnica de CF más utilizada en la actualidad (Sarwar *et al.* (2001); Karypis (2001)). Amazon es un buen ejemplo del uso de esta técnica. En lugar de utilizar la similitud entre las preferencias de los usuarios, este enfoque utiliza la similitud entre los patrones de puntuación de elementos. Si dos elementos tienen las mismas puntuaciones positivas y negativas por parte de los usuarios, éstos son similares, y se espera que los usuarios tengan preferencias similares para elementos similares.

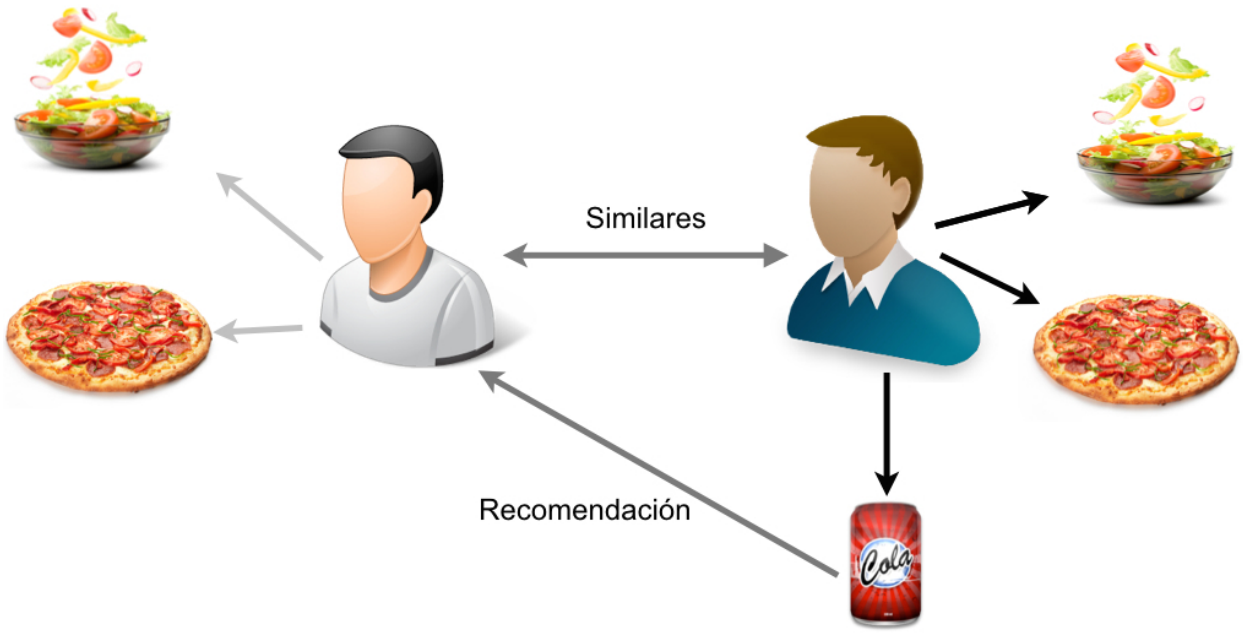


Figura 20: Representación del CF basado en usuarios.

El CF consiste en varias etapas, independientemente del enfoque de interés.

- Definir el perfil de usuario. La primera etapa consiste en definir el perfil de cada usuario a través de las puntuaciones que el usuario ha otorgado a lo largo del tiempo. El perfil del usuario se compone de las puntuaciones numéricas asignadas a cada uno de los elementos. De esta manera, cada usuario u otorga una puntuación $r_{u,i}$ para cada elemento i.

- Obtener la similitud de los usuarios. En el CF basado en usuarios, el objetivo se encuentra en identificar a los usuarios cuyos perfiles sean similares a los de un usuario dado. A estos usuarios, normalmente se conocen como *vecinos*. Esto se realiza al calcular el peso del usuario dado con respecto a los demás y considerando la similitud de las puntuaciones otorgadas a los mismos elementos. En el caso del CF basado en elementos, el objetivo es identificar a aquellos elementos que son similares a un elemento dado.

- Realizar la predicción. Después de que han identificado los K vecinos, es posible combinar las puntuaciones en una predicción al calcular un promedio ponderado de las puntuaciones, utilizando las correlaciones como los pesos. Para el filtrado colaborativo basado en usuarios, el sistema combina las puntuaciones de los N vecinos para generar la preferencia del usuario u con respecto a i (Breese *et al.* (1998)):

$$p_{u,i} = \overline{r}_u + k \sum_{v=1}^{n} (\overline{r}_{v.i} - \overline{r}_v).w_{u,v} \tag{17}$$

donde n es el número vecinos y k es un factor de normalización.

Para el CF basado en elementos, después de identificar el conjunto de S elementos similares a i, $p_{u,i}$ se calcula de la siguiente manera:

$$p_{u,i} = \frac{\sum_{j \in S} s(i,j) r_{u,j}}{\sum_{j \in S} \mid s(i,j) \mid} \tag{18}$$

S representa los k elementos más similares a i que u ha calificado para un vecindario de tamaño k.

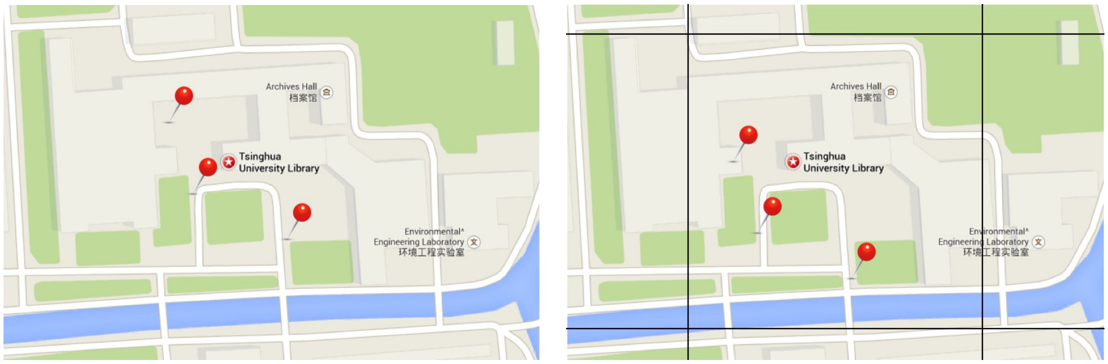
3.3.7.1. El filtrado colaborativo y la ubicación del usuario

A diferencia del enfoque original del CF, en este trabajo, el objetivo se encuentra en evitar la omisión de puntos de interés e información asociada a éstos, al considerar los lugares que fueron visitados por los usuarios que son similares a un usuario dado.

Por ejemplo, considérese que cuatro usuarios, Juan, Marcos, José y Pedro, son compañeros de trabajo y durante la jornada laboral ellos tienen en común varios lugares como la cafetería, el restaurante, la sala de juntas y una oficina. Sin embargo, en algunas ocasiones el dispositivo móvil de Juan se quedó sin energía y no fue posible recolectar datos de localización mientras él se encontraba en el restaurante; el restaurante no cuenta con las visitas necesarias para ser considerado un POI. En cambio, Marcos, José y Pedro no tuvieron problemas con sus dispositivos móviles y lograron recolectar datos de localización mientras los cuatro estaban en el restaurante. Para estos usuarios, el restaurante sí fue considerado como POI.Ya que los cuatro usuarios tienen varios lugares en común, y sólo difieren en un lugar (restaurante), existe cierta probabilidad de que el restaurante debe de considerarse como un punto de interés para Juan.

De esta manera, en lugar de contar con preferencias o puntuaciones de elementos, se tiene preferencias de lugares. Así, en la matriz R, las columnas representan lugares y las filas representan a los usuarios. Sin embargo, obtener la similitud de los usuarios al considerar los lugares que éstos visitan, no es un aspecto trivial.

Debido a que los puntos de interés se identifican para cada usuario y día de la semana, y además, los algoritmos que se utilizan para la identificación de los puntos de interés obtienen el centroide (promedio de latitud y longitud) del área geográfica donde se encuentran dichos lugares, no resulta factible comparar los centroides de los puntos de interés de cada uno de los usuarios para conocer la similitud entre éstos (Figura 21(a)); se requiere considerar áreas geográficas que sean comunes para todos los usuarios con el fin de conocer la similitud. Por ejemplo, considérese el caso en que dos estudiantes asisten a la misma biblioteca; sin embargo, debido al error inherente del GPS los centroides asociados al punto de interés biblioteca difieren. Por lo tanto, en lugar de considerar los centroides de los puntos de interés, se toman como referencia áreas geográficas de



(a) (b)

Figura 21: Similitud del usuario basada en los lugares que visita; a) considerando las coordenadas de los puntos de interés; b) conociendo la similitud de los usuarios tomando como referencia un área o celda geográfica.

un tamaño dado para así obtener la similitud entre los usuarios (Figura 21(b)). De esta manera, un área geográfica o celda incluye/comprende la biblioteca.

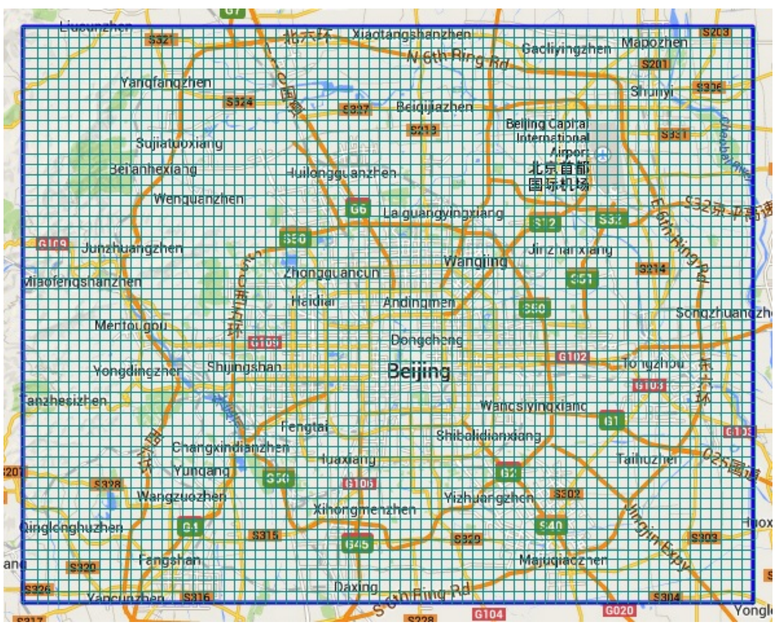


Figura 22: Área geográfica considerada.

Debido a lo mencionado, y con el fin de definir la matriz R, en una primera instancia se requiere seleccionar el área geográfica donde residen o movilizan los usuarios que se desean evaluar. La región que se selecciona se presenta en la Figura 22. Después de que se ha definido esta área, se procede a definir las etapas del CF adaptado al enfoque de este trabajo.

- Construir el perfil del usuario. Después de seleccionar el área geográfica, se proce-

de a dividirla en una cuadrícula de $N \times M$ celdas de un tamaño dado tal como se muestra en la Figura 25(a), cada celda representa una columna en la matriz R. De esta manera, la celda 1,1 representa la columna 1; la celda 1,2 representa la columna 2, y así de manera sucesiva hasta la celda N, M que representa la columna N*M. De esta manera, la matriz R cuenta con $N \times M$ columnas, y U filas que representan la cantidad de usuarios a ser evaluados.

Posteriormente, a fin de definir el arreglo r_u para el usuario i, se toman como referencias las coordenadas de cada uno de los puntos de interés del usuario i. Es importante mencionar que, dado que los POIs se identifican por cada día de la semana y, por consiguiente, se define un modelo de predicción por cada día de la semana, de igual manera se define un arreglo para cada día de la semana. De esta manera, un usuario cuenta con hasta siete arreglos en la matriz R.

Aclarado lo anterior, los índices del arreglo r_u se marcan con un 1 para indicar que en la celda correspondiente el usuario i cuenta con un POI. En caso contrario, los índices del arreglo se marcan con 0. Por ejemplo, considerando la Figura 25, el usuario i cuenta con seis POI (Figura 25(a)). Las celdas que corresponden a la ubicación de dichos POI se marcan, como se muestra en la Figura 25(b), y finalmente los índices correspondientes en el arreglo r_u se etiquetan con 1, como se muestra en la Figura 25(c). De esta manera, se tiene conocimiento de aquellas celdas en las que el usuario pasa su tiempo. Luego de definir cada arreglo r_u, éstos se agrupan para definir la matriz R (Figura 23). En este punto es posible calcular la similitud entre cada par de arreglos.

- Determinar la similitud de los usuarios. Considerando el escenario mencionado, Juan, Marcos, José y Pedro se pueden definir como r_{u7}, r_{u5}, r_{u4}, r_{u2} en la Figura 24(a), y utilizando una función de similitud, se encuentra una cierta similitud entre Juan y Marcos; Juan y José; y, Juan y Pedro; estos usuarios tienen varios lugares en común (P_2, P_4, P_5). Existen diferentes funciones matemáticas que se pueden utilizar para calcular la similitud entre dos elementos. Utilizando la similitud del coseno, los usuarios se representan como arreglos, y la similitud entre éstos se mide a través del coseno de la distancia entre los arreglos:

$$sim(u,v) = cos(\overrightarrow{u}, \overrightarrow{v}) = \frac{\overrightarrow{u}.\overrightarrow{v}}{\parallel \overrightarrow{u} \parallel * \parallel \overrightarrow{v} \parallel} \quad (19)$$

- Evitar la omisión de puntos de interés. Después de conocer la similitud entre los usuarios, para cada uno de los arreglos, se seleccionan los k arreglos más similares a éste. Así, para evitar la omisión de puntos de interés, se consideran los arreglos r_u y r_v (V representa el conjunto de arreglos más similares al arreglo u). Si la similitud entre los arreglos ($sim(u,v)$) es mayor a un cierto umbral, se comparan los índices marcados con 1 en ambos arreglos; si el arreglo r_v tiene una mayor cantidad de índices marcados con 1, se verifica que el usuario i en el día de la semana j, que se encuentra asociado al arreglo r_u tenga registros (POIs candidatos) en aquellas celdas en las que difiere con el arreglo r_v. Si el usuario i en el día de la semana j tiene al menos n cantidad de visitas a los POI candidatos, y las visitas a dichos lugares se realizaron durante el periodo de tiempo que comprende el patrón de movilidad actual, estos lugares se consideran POI.

  Con el propósito de conocer la funcionalidad del modelo de predicción después de agregar nuevos POI, se define un nuevo modelo de predicción. El nuevo modelo de predicción considera los POI que se identificaron en una primera instancia (al utilizar los algoritmos para identificar POI en interiores y exteriores), y los POI agregados en este último proceso.

  Continuando con el ejemplo anterior (Figura 24(a)), y tomando como referencia los lugares visitados por u_2, u_3, u_4, existe cierta probabilidad de que el lugar P_6 se debe considerar como un POI para el usuario i en el día de la semana j que se encuentra asociado al arreglo u_7.

  El proceso anterior se encuentra restringido por la fecha en que se realizaron las visitas a los POI candidatos. Esto es, si se agrega un lugar (POI candidato) que no se visitó dentro del periodo de tiempo que comprende el patrón de movilidad actual, el modelo de predicción no reflejará de manera adecuada el comportamiento de la movilidad del usuario y, por consiguiente, las predicciones que se realicen con dicho modelo de predicción serán imprecisas.

  Este proceso de incorporación de POIs se aplica para cada arreglo r_u y cada valor

de k. Así, para cada usuario y día de la semana se verifica la incorporación de lugares (POIs candidatos) como POIs, considerando los lugares (celdas) visitadas por los k arreglos más similares a éste.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_i	P_N
U_1									...	
U_2		1		1	1	1		1	...	
U_3									...	
U_4		1		1	1	1	1		...	
U_5		1	1	1	1	1			...	
U_6									...	
U_7		1		1	1				...	
U_8									...	
U_j	...	...	...	...	...	...	...	...	...	...
U_M									...	

Figura 23: Matriz de puntuaciones o matriz *R*.

3.3.7.2. Similitud entre lugares

Al considerar la información que se obtiene mediante la matriz R, resulta factible obtener la similitud entre lugares (columnas) como se muestra en la Figura 24(b). Este aspecto resulta de utilidad para tener un entendimiento más profundo acerca de los lugares que visitan los usuarios, y así conocer qué tipo de usuarios visitan determinados lugares, o bien, para realizar recomendaciones basadas en estos lugares.

3.4. De la predicción de la movilidad de un individuo a la predicción de la movilidad de la población.

Finalmente, hasta este punto se ha propuesto un modelo de predicción espacio-temporal que se basa en los modelos ocultos de Markov. El modelo de predicción propuesto incluye dos algoritmos para identificar POIs en interiores y exteriores, así también permite capturar el comportamiento reciente del usuario con el fin de definir de manera adecuada el modelo de predicción, y de tener un mayor entendimiento de la movilidad de los usuarios, se utiliza la similitud entre éstos para solucionar la falta de datos de localización y así evitar la omisión de POI para cada uno de los usuarios.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_i	P_N
U_1									...	
U_2		1		1	1	1		1	...	
U_3									...	
U_4		1		1	1	1	1		...	
U_5		1	1	1	1	1			...	
U_6									...	
U_7		1		1	1				...	
U_8									...	
U_j	...	...	...	...	...	...	...	...	...	...
U_M									...	

(a)

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_i	P_N
U_1									...	
U_2		1		1	1	1		1	...	
U_3									...	
U_4		1		1	1	1	1		...	
U_5		1	1	1	1	1			...	
U_6									...	
U_7		1		1	1				...	
U_8									...	
U_j	...	...	...	...	...	...	...	...	...	...
U_M									...	

(b)

Figura 24: Uso del contexto en la predicción de la movilidad; a) considerando la ubicación actual; b) considerando la ubicación actual y el día de la semana; c) considerando la ubicación actual, el día de la semana, y la hora.

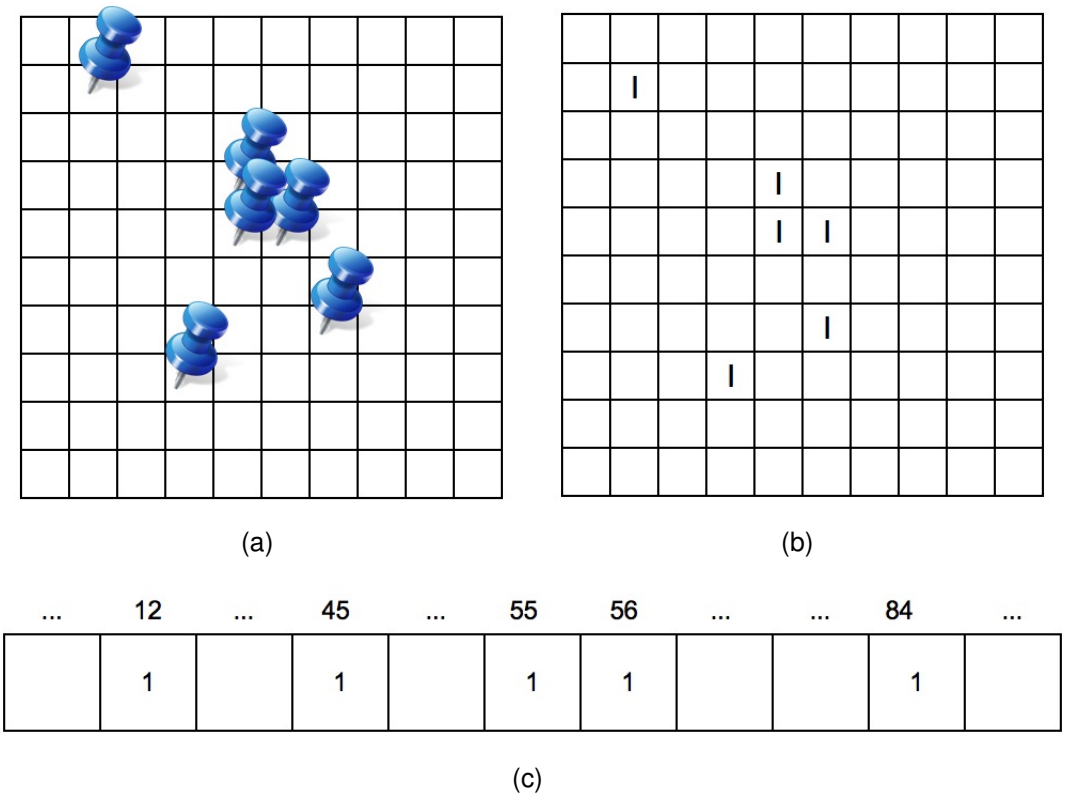


Figura 25: Definición de los arreglos r_u; a) ubicación de los POIs; b) identificación de las celdas en donde se encuentran los POIs; c) etiquetado de los índices correspondientes a las celdas donde se encuentran los POIs

A partir del modelo de predicción propuesto es resulta factible predecir en qué lugares estará el usuario en las próximas horas, y el tiempo de arribo a éstos. Además de realizar la predicción para un usuario en particular, resulta factible y de interés el combinar las predicciones individuales a fin de hacer posible la creación de aplicaciones que sean de interés y/o utilidad para la población en general.

De esta manera, como se comentó en la introducción, existen diversos escenarios que se pueden implementar al combinar las predicciones individuales. Así, cada usuario puede compartir la predicción de su movilidad cada determinado lapso de tiempo (e.g., diariamente, cada 12 horas, etcétera) como se muestra en la Figura 26.

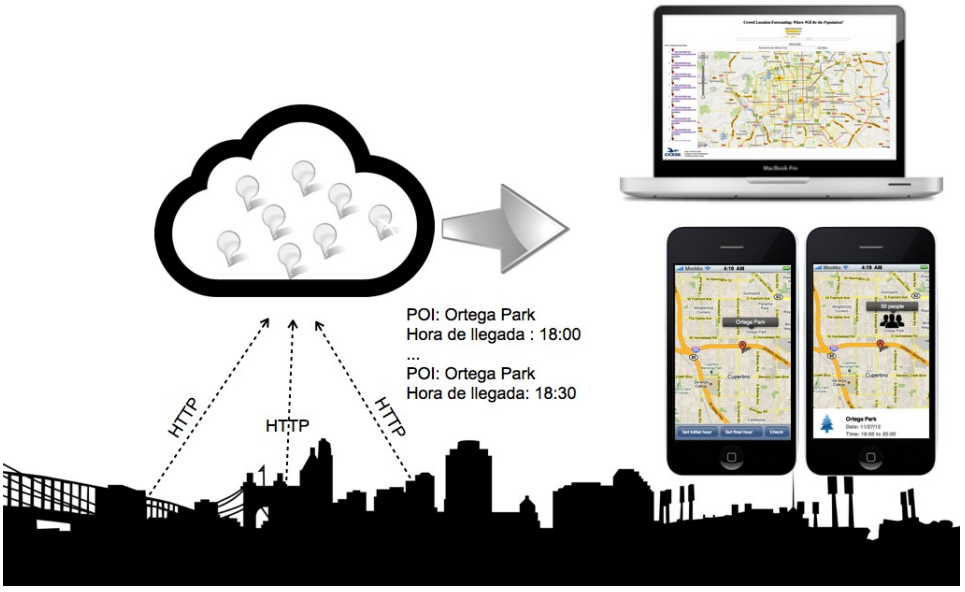
Una vez que se tienen las predicciones de una determinada cantidad de usuarios, éstas pueden servir a varias aplicaciones. Por ejemplo, considerando la Figura 26(b), se conoce la cantidad de personas que habrá en ciertos lugares en un día y periodo de tiempo determinado. Esta aplicación es de utilidad para cuestiones cotidianas como evitar lugares congestionados, para cuestiones de planeación urbana, asignación de recursos, entre otros. En el capítulo 6 se presentan algunos ejemplos significativos de la utilidad de la predicción de la movilidad y la combinación de dichas predicciones para beneficio de la población.

3.5. Resumen

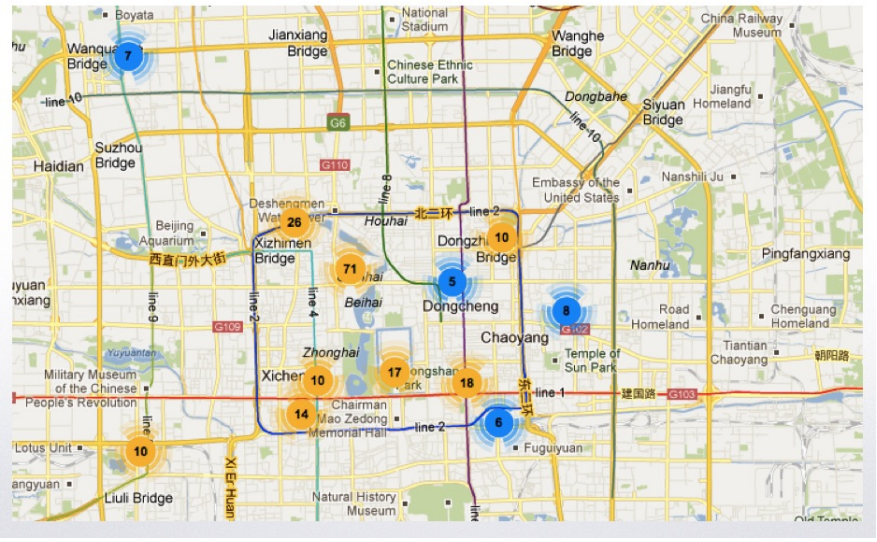
En este capítulo se presentan las características de la movilidad que son relevantes y que se consideran para definir el modelo de predicción espacio-temporal. Además, se describe cómo la movilidad del usuario se puede representar como una cadena de Markov cuando el usuario se moviliza entre los puntos de interés.

Posteriormente, al considerar la relación que se tiene con la hora del día, la movilidad se define como un modelo oculto de Markov, y así es posible realizar la predicción de la movilidad en el aspecto espacial y temporal. Sin embargo, definir la movilidad del usuario como un modelo oculto de Markov no es suficiente para obtener un modelo de predicción robusto.

Por lo tanto, también se describe la función que cumple el algoritmo de Viterbi para



(a)



(b)

Figura 26: De la predicción de la movilidad individual, a la predicción de la población; a) esquema general para combinar las predicciones individuales; b) cantidad de personas que estarán en cierto lugar en un día y periodo de tiempo determinado

identificar la secuencia de estados ocultos de manera eficiente. Aunado a lo anterior, se presentan y discuten los aspectos que son importantes y que se incorporan en el modelo de predicción: identificación de los puntos de interés, entrenamiento del modelo de predicción, la dinamicidad de la movilidad a lo largo del tiempo, y la escasez de datos de localización.

Al considerar la frecuencia de visitas, el tiempo mínimo de estadía, y el periodo de tiempo que comprende el patrón de movilidad más reciente del usuario, permite identificar de manera precisa el área geográfica donde se encuentra cada uno de los puntos de interés del usuario. De igual manera, al definir el modelo de predicción como un modelo oculto de Markov, y considerar los diferentes aspectos, se está en posición de conocer la cantidad de datos necesaria para realizar el entrenamiento del modelo de predicción, así como determinar si la movilidad de un usuario en un día dadode la semana cuenta con la propiedad Markoviana.

Y, al incorporar el mecanismo de actualización de puntos de interés y la ventana deslizante, el modelo de predicción propuesto contempla los cambios en la movilidad del usuario a lo largo del tiempo. Finalmente, al considerar los lugares que visitan los usuarios similares a un usuario en particular, se evita la omisión de puntos de interés y, por consiguiente se tiene un mayor conocimiento de la movilidad de cada uno de los usuarios.

A fin de conocer la funcionalidad del modelo de predicción propuesto, en el capítulo 6 se describen los experimentos que permiten evaluar el funcionamiento de éste. Para ello, se utilizan conjuntos de datos de acceso público, estos conjuntos incluyen registros de la movilidad de usuarios en dos entornos (campus universitario, ciudad). En total se presentan cuatro experimentos, cada uno de ellos incluye uno o varios de los aspectos mencionados. Cada experimento se describe en función del conjunto de datos utilizado, número de usuarios, definición del modelo oculto de Markov, datos de entrenamiento, entre otros.

Capítulo 4. Evaluación

A fin de conocer el desempeño del modelo de predicción propuesto, se realizaron varios experimentos utilizando diferentes conjuntos de datos. Para definir un modelo de predicción robusto, en cada uno de los experimentos se incluyen de manera gradual los aspectos que se describieron en el capítulo anterior. En este capítulo se presenta la descripción de cada uno de los experimentos que se llevaron a cabo. Para ello, se describe el conjunto de datos que se utilizó, la forma en que se definió el modelo de predicción, el entrenamiento de los datos, la cantidad de usuarios, el número de predicciones realizadas, y la efectividad de la predicción.

4.1. Conjuntos de datos

Para realizar la evaluación del modelo de predicción propuesto, se utilizaron dos conjuntos de datos, los cuales se encuentran disponibles de manera pública. A continuación se describen las características de los dos conjuntos de datos.

4.1.1. Conjunto de datos de Dartmouth

Este conjunto de datos contiene registros de conexiones a puntos de acceso (Wi-Fi), los cuales se recolectaron en el campus de Dartmouth College durante 3 años (Kotz *et al.* (2007a)). El campus tiene una extensión geográfica de aproximadamente 0.8 kilómetros cuadrados, con alrededor de 3200 estudiantes, y una red de comunicaciones que incluye 476 puntos de acceso. Sin embargo, el conjunto de datos contiene registros de conexión de 6202 usuarios (se incluyen los registros de conexión de académicos, estudiantes, administrativos, y visitantes) a 450 puntos de acceso. Cada registro incluye: fecha, hora, y nombre del punto de acceso (SSID) al que se encuentra asociado el dispositivo. Para indicar que el dispositivo móvil ha perdido la conexión con un punto de acceso se agregó la etiqueta “OFF”, de esta manera se determina la hora de conexión y desconexión a cada punto de acceso. Este conjunto de datos es utilizado en varios trabajos relacionados a fin de evaluar diversos modelos de predicción. En la Figura 27 se presenta un fragmento de los registros de conexión asociados a un usuario determinado.

```
Timestamp SSID
1026842579 AdmBldg16AP1
1026842620 OFF
1026844420 AdmBldg16AP1
1026845774 OFF
1026847575 AdmBldg16AP1
1026848609 OFF
1026850410 AdmBldg16AP1
1026850410 OFF
1026852211 AdmBldg16AP1
1026852773 OFF
1026854574 AdmBldg16AP1
1026854621 OFF
1026928686 LibBldg2AP7
1026928687 OFF
1026933991 AdmBldg13AP1
1026934902 OFF
1026994453 LibBldg2AP7
1026994457 OFF
1026996258 LibBldg2AP7
1027000802 OFF
1027002602 LibBldg2AP7
1027002603 OFF
1027004404 LibBldg2AP7
1027005861 OFF
1027007663 LibBldg2AP7
1027008870 OFF
```

Figura 27: Fragmento del conjunto de datos de Dartmouth.

4.1.1.1. Pre-procesamiento del conjunto de datos de Dartmouth

Considerando que la conexión a través de WiFi puede ser intermitente, una conexión de larga duración se puede segmentar en varios registros de corta duración. Para evitar este problema, los registros se agrupan si: dada una secuencia de registros al mismo punto de acceso, la diferencia entre el tiempo final del registro x_i y el tiempo inicial del registro x_{i+1} es menor a un cierto umbral. Esto es, $EndTime_x - StartTime_{x+1} \leq \delta$. El valor del umbral δ se definió a 300 segundos (Scellato *et al.* (2011)).

4.1.2. Conjunto de datos de Microsoft Research

Este dataset contiene trayectorias GPS recolectadas en el proyecto Geolife de Microsoft Research Asia (Zheng *et al.* (2008, 2009, 2010)). Las trayectorias corresponden a 178 usuarios, recolectadas en un periodo de 4 años, desde Abril 2007 a Octubre 2011. Una trayectoria GPS se representa como una secuencia de registros GPS, cada uno de los cuales incluye: fecha, hora, latitud, longitud, y altitud. Estas trayectorias las recolectaron diversos dispositivos GPS, y teléfonos celulares con GPS a diferentes frecuencias de muestreo; aproximadamente el 91 % de las trayectorias se recolectaron de manera continua (cada 1 - 5 segundos, o cada 5 - 10 metros). Los usuarios participantes se encuentran distribuidos en 30 ciudades de China, y algunos de ellos residen en ciudades de Estados Unidos y Europa; sin embargo, la mayoría de los datos se recolectan en Beijing, China. Al igual que el conjunto de datos anterior, éste se ha utilizado para evaluar distintos modelos de predicción, sistemas de recomendación, y para identificar diversos patrones de movilidad (Li *et al.* (2008); Zheng *et al.* (2011); Yuan *et al.* (2013, 2010)). En la Figura 28 se presenta un fragmento de los datos de movilidad de un usuario determinado.

4.2. Experimento 1. Predicción de la movilidad dentro de un campus universitario utilizando registros de conexiones a puntos de acceso

4.2.1. Objetivo del experimento.

Al inicio de la investigación sólo se tenía acceso al conjunto de datos de Dartmouth. Por lo tanto, en el primer experimento el objetivo se centró en predecir la movilidad del

```
Latitud,Longitud, 0, Altitud (pies), Días desde 12/30/1899, Fecha, Hora
39.984094,116.319236,0,492,39744.2451967593,2008-10-23,05:53:05
39.984198,116.319322,0,492,39744.2452083333,2008-10-23,05:53:06
39.984224,116.319402,0,492,39744.2452662037,2008-10-23,05:53:11
39.984211,116.319389,0,492,39744.2453240741,2008-10-23,05:53:16
39.984217,116.319422,0,491,39744.2453819444,2008-10-23,05:53:21
39.98471,116.319865,0,320,39744.2454050926,2008-10-23,05:53:23
39.984674,116.31981,0,325,39744.245462963,2008-10-23,05:53:28
39.984623,116.319773,0,326,39744.2455208333,2008-10-23,05:53:33
39.984606,116.319732,0,327,39744.2455787037,2008-10-23,05:53:38
39.984555,116.319728,0,324,39744.2456365741,2008-10-23,05:53:43
39.984579,116.319769,0,309,39744.2456944444,2008-10-23,05:53:48
39.984579,116.319769,0,309,39744.2457291667,2008-10-23,05:53:51
39.984577,116.319766,0,309,39744.2457523148,2008-10-23,05:53:53
39.984611,116.319822,0,304,39744.2458101852,2008-10-23,05:53:58
39.984959,116.319969,0,304,39744.2458680556,2008-10-23,05:54:03
39.985036,116.320056,0,304,39744.2458796296,2008-10-23,05:54:04
39.984741,116.320037,0,304,39744.2458912037,2008-10-23,05:54:05
39.98462,116.32012,0,302,39744.2459143519,2008-10-23,05:54:07
```

Figura 28: Fragmento del conjunto de datos de Geolife.

usuario entre los puntos de acceso del campus universitario. Al considerar T como el tiempo actual, el modelo predice la secuencia de puntos de acceso donde estará el usuario en el lapso $T + \Delta T$, siendo ΔT el periodo de predicción. Por lo tanto, en este experimento se identifican los patrones de conexión a los puntos de acceso dentro del campus. Posteriormente se define el modelo de predicción en base a la transición de cada usuario entre los puntos de acceso, la hora del día de la conexión a estos puntos, entre otros aspectos. Como este fue el primer experimento, no incluyó todos los aspectos mencionados en el capítulo anterior (Figura 13). Por ejemplo, la predecibilidad, la dinamicidad de la movilidad y la falta de datos se incluyen en los experimentos posteriores. De esta manera, la robustez del modelo se incrementa conforme se consideran más aspectos. A continuación se describe este experimento.

4.2.1.1. Usuarios considerados en el experimento

El conjunto de datos de Dartmouth contiene los registros de conexiones de alrededor de 6000 usuarios; sin embargo, en este experimento sólo se consideraron los registros de 200 usuarios. El criterio que se consideró para la selección de estos usuarios fue

la cantidad de datos de conexión asociados a los usuarios. Se seleccionaron aquellos usuarios con registros densos de conexión, a fin de definir de mejor manera el modelo de predicción.

La mayoría de los usuarios no cuentan con registros de conexión durante un periodo prolongado de tiempo, o en su defecto, los usuarios cuentan con registros dispersos durante el periodo de colecta (i.e. un par de registros de conexión por día). Por lo tanto, para estos usuarios no es posible definir de manera adecuada el modelo de predicción. Esto es, no se cuenta con información suficiente para identificar los puntos de interés, y por consiguiente no es posible definir los componentes del HMM.

4.2.1.2. Puntos de interés

Con respecto a la identificación de POIs en este experimento, éstos se definen por el conjunto de puntos de acceso que son importantes para el usuario. Un punto de acceso se considera un POI cuando el usuario lo visita al menos una cantidad n de veces. El valor de n se define a 20 como lo hace Scellato *et al.* (2011) al utilizar el mismo conjunto de datos. Después de varias pruebas, Scellato *et al.* (2011) encontraron que 20 visitas resultan adecuadas para identificar aquellos puntos de acceso que son de interés para el usuario.

Con respecto a duración de cada visita en los puntos de acceso, se definió un valor mínimo de 30 segundos. En una primera instancia el objetivo de este experimento era definir un modelo de predicción para realizar mejores decisiones de ruteo en las redes oportunistas. Por lo tanto, para fines de la aplicación, un tiempo de conexión mínimo de 30 segundos permite la transmisión de aproximadamente 200 MB al utilizar una interfaz de comunición 802.11n. Además, al considerar una duración mínima de conexión permite excluir a aquellos puntos de acceso que no son importantes para el usuario. Por ejemplo, considerando que el usuario transita por un lugar que proporciona conectividad, en dicho recorrido es posible que se realice una conexión de manera transparente para el usuario cuando el dispositivo realiza un escaneo y encuentra un punto de acceso conocido (tiempo atrás el usuario realizó la conexión explícita a dicho punto de acceso); sin embargo, esta conexión se descarta debido a la corta duración de la conexión.

4.2.2. Definición del modelo de predicción

Con respecto al modelo de predicción, los componentes del HMM se definen de la siguiente manera:

4.2.2.1. Estados ocultos

Los estados ocultos se encuentran definidos por el conjunto de POIs. Además, se ha agregado un estado adicional para indicar que el usuario no se encuentra en un punto de interés; el usuario se encuentra en un punto de acceso que no fue definido POI, o bien, se encuentra en movimiento.

4.2.2.2. Observaciones

Un aspecto importante al definir el HMM, se encuentra en especificar las observaciones. Esto es, definir los tiempos que serán considerados por el modelo de predicción para estimar los tiempos en los cuales el usuario estará en los POIs. Para ello, para un día determinado se consideran los tiempos de inter-contacto de las conexiones del usuario a los POIs. De acuerdo a los resultados presentados en trabajos previos (Chaintreau *et al.* (2007); Cai y Eun (2009)), los tiempos de inter-contacto resultan una métrica útil para estimar los contactos futuros. Esto es, considerar el tiempo transcurrido entre la desconexión a un punto de acceso (POI) y la siguiente conexión (ya sea al mismo punto de acceso u otro punto de acceso). Por lo tanto, en este experimento se obtiene el promedio de los tiempos de inter-contacto entre los puntos de acceso que fueron identificados como POIs (Ecuación 20). Posteriormente, se toma como referencia el periodo de tiempo que comprende entre la primera y la última conexión a los POIs, y este periodo se secciona en m segmentos de acuerdo al valor de t_{ic}. Por ejemplo, si los tiempos de la primera y la última conexión son: 7:00 y 19:00 (intervalo de 720 minutos), y el valor de $t_{ic} = 90$ minutos, se definen 9 valores para las observaciones. De esta manera, el conjunto de observaciones es: $O = 7:00, 8:30, 10:00, 11:30, 13:00, 14:30, 16:00, 17:30$, y $19:00$.

$$t_{ic} = promedio\ (\text{tiempos de inter-contacto}) \tag{20}$$

4.2.2.3. Matriz de transición

Esta matriz define la probabilidad de que un usuario se traslade de un POI (q_j) al resto de los POIs (incluyendo el estado correspondiente a no conexión), o bien que el usuario permanezca en el mismo POI (q_i).

4.2.2.4. Matriz de confusión

Esta matriz define la probabilidad de que un usuario se encuentre en un POI determinado en el tiempo definido por la observación $o_j \pm \epsilon$, donde ϵ es un intervalo de tiempo, el cual se definió a 15 minutos puesto que el arribo del usuario al punto de interés no es preciso. Después de realizar diversas pruebas con varios valores de ϵ (i.e., 5, 15, 30 minutos), se encontró que con un margen de15 minutos se identifica con mayor probabilidad ($\approx$ 90 %) el arribo al punto de interés.

4.2.2.5. Vector

Este vector define la probabilidad inicial para cada POI. Esto es, para un día determinado de la semana se define la probabilidad de que la primera conexión del usuario sea en un POI determinado.

4.2.3. Entrenamiento del modelo de predicción

Para realizar el entrenamiento y las pruebas de cada modelo de predicción, se utilizaron los datos de conexión correspondientes a los últimos tres meses de registros. Los datos de conexión del último mes se utilizaron para realizar las pruebas de cada uno de los modelos de predicción; sin embargo, sólo se realizaron pruebas considerando los datos de la primera semana de pruebas. El entrenamiento de cada modelo de predicción se realizó con los datos de conexión correspondientes a los dos meses previos al mes de pruebas. Los modelos de predicción se definieron utilizando un HMM de primer orden; se realizaron pruebas considerando un HMM de segundo orden; sin embargo, aproximadamente sólo en el 10 % de los casos se obtuvo un incremento de hasta 2 %. Por lo tanto, en éste y en el resto de los experimentos, los modelos de predicción se definen como un HMM de primer orden.

4.2.3.1. Número de predicciones

Con respecto a las predicciones, para cada usuario y modelo de predicción se realizaron 5 predicciones considerando diferentes valores para ΔT : 30 minutos, 1, 3, 5, y 7 horas. Un total de 7000 predicciones, 35 predicciones por cada usuario. De esta manera se conoce el rendimiento del modelo de predicción en un término medio, al considerar un ΔT de varias horas.

4.2.3.2. Efectividad de la predicción

Para determinar la efectividad de la predicción, si se desea conocer dónde estará el usuario en el periodo $[T, T + \Delta T]$, la predicción es correcta si el usuario se encuentra en el lugar q_i en el intervalo $[T_{pred} - \epsilon, T_{pred} + \epsilon]$ (Ecuación 21), donde ϵ representa un margen de error. Esto es, la predicción es correcta cuando el usuario está en el POI definido por q_i, en el tiempo indicado por la observación o_i, con cierto margen de error. También es correcta si la predicción indica que el usuario no estará en un POI (en el caso donde q_i corresponde al lugar no definido como POI). El margen de error ϵ se definió en 15 minutos.

$$T_{pred} = T + o_1 \; 1 \leq i \leq \text{Número de observaciones en el periodo de predicción} \qquad (21)$$

Para evaluar los resultados de este experimento, éstos se compararon con aquéllos que se obtuvieron al utilizar un método basado en NextPlace (Scellato *et al.* (2011)). Este método se define como $NP*$.

4.2.3.3. NP* - Método basado en NextPlace

El mecanismo original de NextPlace incluye dos métodos: uno para predecir la ubicación del usuario en un tiempo determinado, y otro para predecir el tiempo de estadía en dicho lugar. El método $NP*$ sólo incluye el mecanismo para predecir la ubicación del usuario en un tiempo determinado.

El modelo de predicción propuesto predice la secuencia de lugares en dónde el usuario estará en un periodo de tiempo dado, mientras que el $NP*$ predice el lugar donde estará el usuario en un tiempo específico. Debido a ello, y a fin de realizar una comparación justa del rendimiento de ambos modelos de predicción, para cada observación presente en el periodo de predicción, se aplica el método $NP*$ para predecir la ubicación del usuario considerando el valor de la observación. Esto es, si se desea conocer la secuencia de lugares en los que estará el usuario en el periodo $[11:00, 15:00]$, y si en este periodo se encuentran 3 observaciones (e.g., 11:30, 13:00,14:30), el método $NP*$ se aplica a cada una de las observaciones. Posteriormente, los resultados se promedian a fin de compararlos con los resultados del modelo propuesto.

El método NP* se utiliza en este y en los experimentos siguientes para evaluar el rendimiento del modelo de predicción propuesto.

4.3. Experimento 2. Predicción de la movilidad del usuario en un área urbana utilizando registros de GPS

4.3.1. Objetivo del experimento.

En el transcurso del primer experimento se tenía acceso a una gran cantidad de datos de localización; sin embargo, con dichos datos sólo fue posible conocer la movilidad del usuario en un área geográfica reducida y durante un periodo de tiempo limitado (durante la estadía de los usuarios en el campus universitario). Por lo tanto, en el segundo experimento se analiza la movilidad del usuario en un área geográfica a nivel urbano y durante un periodo de tiempo mayor a lo largo de cada día.

Con el conocimiento que se adquirió en el primer experimento, en éste se consideran otros de los aspectos mencionados en el capítulo anterior, lo cual permite definir un modelo de predicción más robusto. Por ejemplo, se considera el aspecto de la predecibilidad a fin de demostrar que la movilidad del usuario entre los POIs puede ser representada como una cadena de Markov. Además, se toma como referencia la similitud de la movilidad del usuario para definir de mejor manera el periodo de tiempo que debe ser considerado para realizar el entrenamiento del modelo de predicción. Finalmente, se incluye el aspecto de la actualización de los POIs a fin de agregar y/o eliminar aquellos POIs que dejan

de ser importante para el usuario.

A continuación se describen los detalles de este experimento:

4.3.1.1. Usuarios considerados en el experimento.

En este experimento se utilizan los datos del proyecto Geolife, el cual contiene registros de localización de 178 usuarios; sin embargo, sólo se utilizaron los registros de 63 usuarios. Estos usuarios cuentan con una gran cantidad de registros de localización a lo largo del periodo de sensado (i.e., datos de localización de más de 12 horas en cada día), lo cual permite identificar los patrones de movilidad y así definir de mejor manera cada uno de los modelos de predicción. En cambio, el resto de los usuarios cuentan con registros de localización dispersos durante el periodo de sensado.

4.3.1.2. Puntos de interés.

A fin de identificar los lugares significativos del usuario a diferentes niveles de granularidad, se utilizaron cinco valores para el radio del clúster. De esta manera, los algoritmos encargados de identificar POIs tanto en exteriores como en interiores definen el radio del clúster a 5, 25, 50, 75, y 100 metros. Y por consiguiente, resulta viable realizar la predicción de la movilidad a diferentes niveles de granularidad (e.g., a nivel habitación, hogar, vecindario, colonia). Con respecto al tiempo mínimo de estadía (t), Ashbrook y Starner (Ashbrook y Starner (2003)) recomiendan un valor mínimo de 10 minutos. Los autores argumentan que con este valor, se omiten las falsas estadías, por ejemplo, cuando un usuario espera el cambio de luces en un semáforo, o bien cuando el usuario transita por un túnel subterráneo. Además del tiempo mínimo de 10 minutos, también se consideraron otros valores para t: 30 y 60 minutos. Finalmente, con respecto a la frecuencia de visitas, un lugar debe de contar con al menos n visitas, donde n representa el periodo que cubre el patrón de movilidad actual / 2 (e.g., la cantidad de lunes previos/2).

4.3.2. Modelo de predicción

Con respecto a la definición del modelo de predicción, los componentes del HMM están definidos de la siguiente manera:

4.3.2.1. Estados ocultos

Los estados ocultos se encuentran definidos por el conjunto de POIs. Además, se ha agregado un estado adicional para indicar que el usuario se encuentra en un lugar que no fue definido como punto de interés, o bien que el usuario se encuentra en movimiento. Para este experimento se identificaron POIs en interiores y en exteriores con diferentes radios de clúster. Sin embargo, al momento de considerar los POIs como estados ocultos no se hace alguna distinción. Al igual, cuando se realiza la predicción de la movilidad del usuario, no se especifica si el lugar en el que estará el usuario es un punto de interés en exteriores o en interiores.

4.3.2.2. Observaciones

Entre los factores que determinan la eficiencia del modelo de predicción se encuentra el hecho de definir de manera adecuada las observaciones. Ya que el objetivo del modelo de predicción propuesto es predecir los lugares en los que está el usuario y los tiempos de arribo a éstos, es de suma importancia definir las observaciones. Por lo tanto, para este experimento las observaciones se definen de acuerdo a los tiempos de arribo y partida a cada uno de los POIs. De acuerdo a los resultados presentados por Do y Gatica-Pérez (2012), los tiempos de arribo y partida a algunos lugares no varían mucho. Por ejemplo, considerando las actividades de un estudiante o de una persona de oficina, hay tiempos definidos para la llegada, descanso, y/o salida. De esta manera, es posible definir de manera precisa el tiempo en que los usuarios llegarán a los POIs, el periodo de tiempo en el que los usuarios estarán ahí, y la hora de partida. Al igual, al tener conocimiento de los tiempos de partida se predice que los usuarios estarán en un lugar no definido como POI. Por lo tanto, las observaciones las define el promedio de los tiempos de arribo y partida de cada uno de los POIs.

4.3.2.3. Matriz de transición

Esta matriz define la probabilidad de que un usuario se traslade de un POI (q_i) al resto de los POIs (incluyendo al estado que define que un lugar no es un POI), o bien que permanezca en el mismo q_i.

4.3.2.4. Matriz de confusión

Esta matriz define la probabilidad de que un usuario se encuentre en un POI dado q_i en el tiempo definido por la observación $o_j \pm un\ rango\ de\ tiempo\ \epsilon$, el cual, al igual que en el experimento anterior se definió en 15 minutos.

4.3.2.5. Vector

Este vector define la probabilidad inicial para cada POI. Esto es, la probabilidad de que la primera visita del usuario sea a un POI determinado.

4.3.3. Entrenamiento del modelo de predicción

Como se mencionó en el capítulo anterior, otro de los aspectos importantes en la definición del modelo de predicción, es el entrenamiento de éste. El entrenamiento debe de reflejar en la medida de lo posible la movilidad actual o más reciente del usuario: puntos de interés, transición entre estos puntos, y el tiempo de estadía en éstos. Para ello, un aspecto importante es determinar la cantidad de datos que son necesarios para realizar dicho entrenamiento. Si se considera que los patrones de movilidad no son estáticos a lo largo del tiempo, es necesario identificar el periodo de tiempo que comprende el patrón de movilidad actual del usuario, a fin de incluir los datos que reflejan un comportamiento similar del usuario, y de esta manera evitar incluir información de los patrones de movilidad que a la fecha no son relevantes. Por ejemplo, en el caso de un estudiante, éste tiene un patrón de movilidad cuando se encuentra en periodo de vacaciones, y exhibe otro patrón de movilidad durante los cursos de verano. Así, como resultado del cambio en el patrón de movilidad, existe una variación en los lugares que se consideran como puntos de interés, en los tiempos de arribo o partida de los puntos de interés, o bien una variación en la transición entre los POIs.

4.3.3.1. La movilidad del usuario como un arreglo

A fin de identificar el periodo de tiempo que comprende el patrón de movilidad actual del usuario y así realizar el entrenamiento de cada modelo de predicción, la movilidad del usuario se compara día a día para conocer la similitud de ésta a través del tiempo, y así determinar el periodo que comprende el patrón de movilidad actual.

Debido a lo anterior, cada día de la semana se define como un arreglo de 48 segmentos. Cada segmento representa un periodo de 30 minutos. De esta manera, el primer segmento representa el periodo de las 00:00 a las 00:29, el segundo segmento representa el periodo de las 00:30 a las 00:59, y así sucesivamente. Cada uno de los segmentos contiene un índice (iniciando de cero) que representa el identificador del lugar (POI candidato) en el que el usuario ha estado en el periodo de tiempo correspondiente al segmento (Figura 29). El índice 0 representa que el usuario no se encuentra en algún POI candidato (el usuario se encuentra en movimiento) o bien que no tuvieron datos de localización en dicho periodo. Así, cada día de la semana se puede ver como un arreglo de 48 posiciones. El tamaño del segmento se definió a 30 minutos con el fin de conocer a nivel granular dónde ha estado el usuario. Al igual, se consideraron segmentos de 15 y 60 minutos.

Tomando como referencia que los días de la semana se representan como arreglos, se obtiene la similitud entre estos arreglos a fin de identificar el periodo de tiempo que abarca el patrón de movilidad. Este proceso se realiza para cada uno de los días de la semana. De esta manera, se comparan los arreglos correspondientes al Dia_i y al Dia_{i-1}, si la similitud es mayor a un cierto umbral θ, el periodo de entrenamiento se incrementa; el entrenamiento considera los registros asociados a Dia_i y Dia_{i-1}; en caso contrario, el periodo de entrenamiento omite los registros asociados a Dia_{i-1}, y se procede a conocer la similitud entre el arreglo Dia_i y el arreglo Dia_{i-2}, y así sucesivamente. Si la similitud de m días consecutivos es menor al umbral θ, el periodo de entrenamiento termina; sólo se incluyen los registros que pertenecen al mismo patrón de movilidad, evitando así considerar varios patrones de movilidad en el proceso de entrenamiento. De esta manera, si se toma como referencia un lunes i es posible identificar los lunes previos en los cuales el usuario exhibió un patrón de movilidad similar.

Una vez que se identifica el periodo de tiempo en el cual la movilidad del usuario es similar, los datos asociados a dicho periodo de tiempo se utilizan para identificar los puntos de interés, y posteriormente se definen cada uno de los componentes del HMM. De esta manera, se cuenta con el modelo de predicción.

Time / Week	...	07:30	08:00	08:30	09:00	...	23:30
1	...	1	0	0	2		1
2	...	0	0	0	2		1
3	...	0	0	2	2		1

Figura 29: Comparación de la movilidad del usuario en un determinado día durante varias semanas.

4.3.4. Número de predicciones

A diferencia del experimento anterior, en éste se utiliza el último mes de registros para probar la eficiencia de los modelos de predicción. De esta manera, cada modelo de predicción se puede probar hasta en cuatro días diferentes. Aunque es posible utilizar el modelo de predicción definido en el proceso de entrenamiento para realizar las predicciones en las cuatro semanas de pruebas, el objetivo de este experimento es capturar los cambios en el patrón de movilidad del usuario, y así definir de manera adecuada el modelo de predicción, y por consiguiente, realizar mejores predicciones.

Debido a lo anterior, se utiliza el modelo de predicción definido en el proceso de entrenamiento (e.g., utilizando los registros de 8 lunes, Figura 30) para predecir la ubicación del usuario en la primera semana del periodo de pruebas (i.e., noveno lunes). Después de que se han realizado las predicciones para la primera semana de pruebas, se compara la movilidad del usuario de este día (i.e., noveno lunes) con la movilidad del usuario durante el proceso de entrenamiento (e.g. los 8 Lunes anteriores) utilizando la similitud del coseno. Si la similitud es mayor que un cierto umbral θ $(\theta = 0.50)$, el modelo de predicción se actualiza con los registros de este día (i.e., noveno lunes). En caso contrario, el modelo de predicción se utiliza para predecir la movilidad del usuario en la segunda semana de pruebas (i.e., décimo lunes). Este proceso se repite para las semanas de prueba siguientes (i.e., onceavo y duodécimo lunes).

Una vez aclarado lo anterior, en cada semana de pruebas y para cada modelo de predicción se realizaron 5 predicciones considerando diferentes valores para ΔT (30 mi-

nutos, 1, 3, 5, y 7 horas). Un total de 8820 predicciones; 35 predicciones por usuario y semana de pruebas considerada.

Entrenamiento								Prueba			
8	7	6	5	4	3	2	1	1	2	3	4

Figura 30: Número de semanas que se utilizaron para el entrenamiento y prueba del modelo de predicción.

4.3.5. Efectividad de la predicción

Para determinar la efectividad de la predicción (Ecuación 22), si se desea conocer en cuáles POIs estará el usuario en el periodo $[T, T + \Delta T]$, la predicción es correcta si para cada observación en el periodo de predicción, el usuario se encuentra en el lugar definido por q_i en el intervalo $[T_{pred} - \epsilon, T_{pred} + \epsilon]$, donde ϵ representa un margen de error. Esto es, la predicción es correcta cuando el usuario está en el POI definido por q_i, en el tiempo indicado por la observación o_j con un cierto margen de error. La predicción es también correcta si indica que el usuario no estará en un POI (en el caso q_i corresponda a un lugar no definido como POI). El margen de error ϵ se definió a 15 minutos. Al igual que en el experimento anterior, los resultados de este experimento se comparan con aquéllos que se obtuvieron al utilizar el método $NP*$.

$$T_{pred} = T + o_1 \; 1 \leq i \leq \text{número de observaciones en el periodo de predicción} \qquad (22)$$

4.3.6. Actualización de POIs

Como se mencionó con anterioridad, las actividades del usuario varían con el tiempo, y por ende su movilidad. Por lo tanto, el modelo de predicción incorpora un mecanismo para actualizar los POIs y la información asociada a éstos. Por ejemplo, considérese el siguiente escenario: un usuario dado está iniciando clases de natación los días lunes en un horario de 7:30 AM a 8:30 AM en una alberca pública (parte superior de la Figura

31, bloque D). Los lunes siguientes, el usuario continúa con sus clases de natación a la misma hora; sin embargo, la alberca pública no se puede considerar como un POI debido a que no cuenta con las visitas necesarias (la frecuencia de visitas es baja). Después de algunas semanas, finalmente la alberca cuenta con las visitas necesarias para considerarse un POI (parte inferior de la Figura 31), y así el modelo de predicción se actualiza; se agrega el punto de interés y la información relacionada a éste (hora de arribo, partida, probabilidad de transición a otros POIs, y la probabilidad de confusión). El modelo de predicción también se actualiza cuando un lugar deja de ser significativo. Por ejemplo, cuando el usuario mencionado deja de asistir a las clases de natación. Por ello, día a día se realiza el proceso de actualización de los POIs, esto es, se considera la frecuencia de visitas a estos lugares a fin de agregarlos o eliminarlos del modelo de predicción, se actualiza la información relacionada a éstos, y por consiguiente se tiene un modelo de predicción preciso.

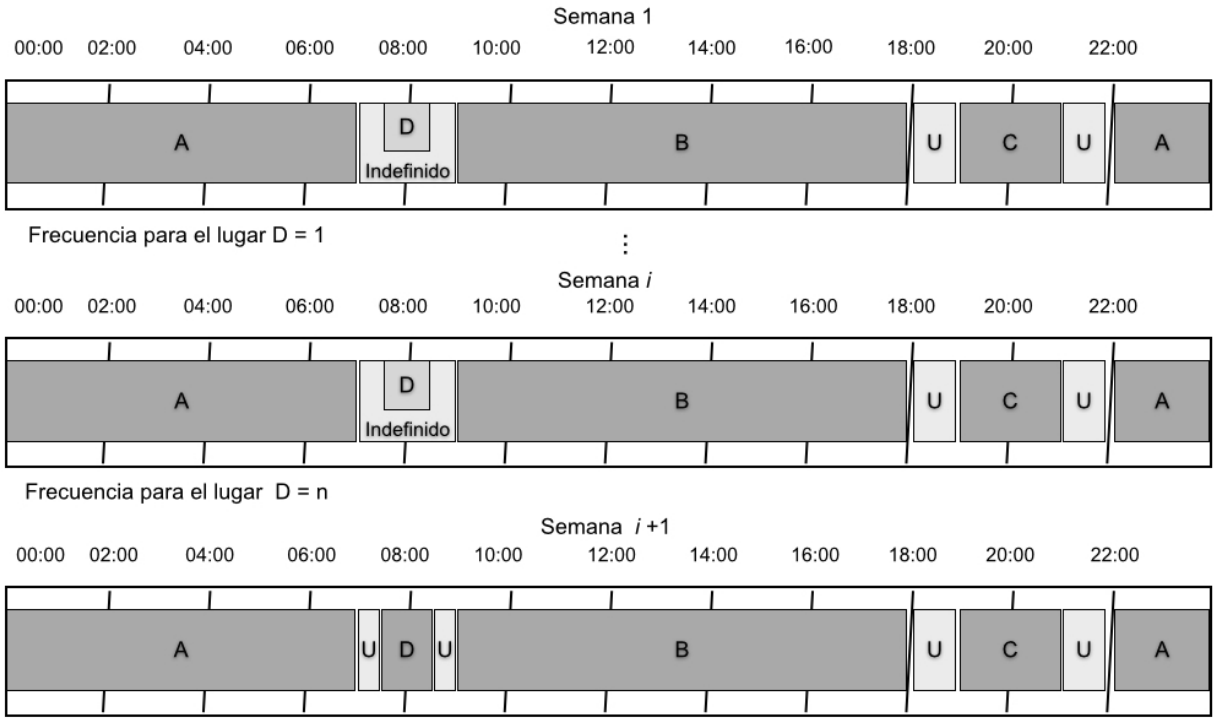


Figura 31: Actualización de los puntos de interés. A,B,C,D representan identificadores de los POIs, U representa a aquel o aquellos lugares que no se consideran POIs.

4.3.7. Predeciblidad de la movilidad del usuario

A diferencia del resto de los experimentos, en este experimento se comprueba la existencia de la propiedad Markoviana al considerar la movilidad del usuario entre los puntos de interés. Para ello, se hace uso de la prueba empírica definida por Zhang *et al.* (2010). Esta prueba toma como referencia la matriz de frecuencia de transición entre los estados ocultos. Suponiendo que la secuencia tiene m estados, f_{ij} denota la frecuencia de transición del estado i al estado j. El cociente de la suma de la j columna de la matriz de frecuencia de transición dividida por el total de todas las columnas y filas de la matriz se llama probabilidad marginal, denotada por $p_{\cdot j}$

$p_{\cdot j} = \sum_{i=1}^{m} f_{ij} / \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij}$, luego

$\chi^2 = 2 \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} \left| log \frac{p_{ij}}{p_{\cdot j}} \right|$ sujeta a la distribución χ^2 con $(m-1)^2$ grados de libertad como su limitación, donde:

$p_{ij} = f_{ij} / \sum_{j=1}^{m} f_{ij}$.

Cuando $\chi^2 > \chi^2_{\alpha}((m-1)^2)$, donde α representa el nivel de significancia, entonces $\{x_i\}$ tiene la propiedad Markoviana; en caso contrario, $\{x_i\}$ no tiene la propiedad Markoviana, y no se puede ver como una cadena de Markov.

En el último caso, al modelar la movilidad del usuario como un modelo oculto de Markov, los resultados serían imprecisos; la existencia de la propiedad Markoviana determina la eficiencia del modelo de predicción propuesto.

4.4. Experimento 3. Predicción de la movilidad del usuario a lo largo del tiempo

4.4.1. Objetivo del experimento

En el tercer experimento, el objetivo se centra en conocer el rendimiento del modelo de predicción al evaluarse durante un periodo prolongado de tiempo. Al analizar los modelos

de predicción de los trabajos relacionados se identificó que ninguno de ellos considera este aspecto. Esto es, los trabajos relacionados no evalúan el rendimiento de sus modelos de predicción en variadas ocasiones (i.e. semanas, meses).

Predecir la movilidad del usuario durante un periodo prolongado resulta relevante debido a la dinamicidad de la movilidad del usuario; los usuarios cuentan con diferentes POIs a lo largo del tiempo, o bien cambian los tiempos de estadía en dichos POIs. En términos de la precisión no es adecuado contar con un único modelo de predicción para estimar la movilidad del usuario a lo largo del tiempo. A fin de considerar los cambios en la movilidad del usuario, el entrenamiento del modelo de predicción se realiza con los datos más recientes; se utiliza una ventana deslizante. En este experimento la definición de las observaciones difiere de los dos experimentos anteriores.

A continuación se presenta la descripción de este experimento.

4.4.1.1. Usuarios considerados en el experimento

Al igual que en el experimento anterior, en éste se utilizaron los datos del proyecto Geolife de Microsoft Research. Aunque el conjunto de datos contiene registros de 178 usuarios, sólo se consideraron los registros de 18 usuarios. El objetivo de este experimento es conocer la efectividad del modelo de predicción durante un periodo prolongado de tiempo; por lo que se requiere contar con usuarios que cuenten con registros de localización durante varios meses. En la Tabla 5 se presenta la información relacionada a los registros de los 18 usuarios seleccionados; cantidad mínima, máxima y el promedio de semanas durante las cuales se tienen registros de localización para cada día de la semana.

4.4.1.2. Puntos de interés

A fin de identificar los lugares significativos del usuario a diferentes niveles de granularidad, se utilizaron tres valores para el radio del clúster. A diferencia del experimento anterior, en éste, los algoritmos para identificar POIs tanto en exteriores como en interiores consideran radios de clúster de 100, 250 y 500 metros. La razón de utilizar radios de clúster más grandes en este experimento, se debe a que conforme se considera un

Tabla 5: Datos de localización de acuerdo al día de la semana.

	Lu	Ma	Mi	Ju	Vi	Sa	Do
Promedio # semanas	15	16.29	16.35	17.47	16.64	16.58	18
Max # semanas	82	94	84	87	89	87	74
Min # semanas	9	12	11	13	13	11	9

periodo de tiempo mayor, se encuentra una mayor dispersión de los registros GPS (originada por el error de las lecturas GPS). Por lo tanto, para un mismo lugar significativo, los centroides del punto de interés varían. Por consiguiente, a fin de evitar la identificación de un lugar como varios puntos de interés, se contemplan radios de clúster más grandes, de esta manera, los clúster incluyen el lugar significativo.

4.4.1.3. Ventana deslizante

Como se mencionó en el capítulo anterior, en este experimento se emplea de la ventana deslizante para considerar únicamente los datos de la movilidad más reciente del usuario durante un periodo de tiempo dado.

Con el objetivo de conocer el rendimiento de los modelos de predicción al utilizar los datos de movilidad asociados a diferentes periodos de tiempo, se utilizaron dos valores para la ventana deslizante: 4 y 8 semanas. De esta manera, se pretende conocer la variación en el número de POIs, y la precisión de la predicción.

4.4.2. Modelo de predicción

Con respecto a la definición del modelo de predicción en este experimento, los componentes del HMM se definen de la siguiente manera:

4.4.2.1. Estados ocultos

Los estados ocultos se definen por el conjunto de POIs. Además, se agregó un estado adicional para definir que el usuario se encuentra en un lugar que no se considera POI (o bien que el usuario se encuentra en movimiento).

4.4.2.2. Observaciones

A diferencia de los experimentos anteriores, en los cuales las observaciones representan un tiempo específico del día, en este experimento las observaciones representan un periodo de tiempo.

Al utilizar la ventana deslizante, se da el caso de que ésta considera los datos asociados a diferentes patrones de movilidad. Por lo tanto, existe una variación y traslape en cuanto a los tiempos de arribo y/o partida a los diferentes puntos de interés. Por consiguiente, se optó por considerar las observaciones como un periodo de tiempo. Para definir las observaciones se utilizaron 3 valores diferentes para el periodo: 2, 3, y 4 horas. De esta manera, al utilizar un periodo de 2 horas, la primera observación (o_1) contempla el lapso de 00:00 - 1:59, la segunda observación abarca el lapso 2:00 - 3:59, y así sucesivamente. Al utilizar diferentes valores para el periodo de tiempo, se pretende identificar si existe un valor que maximice la precisión de la predicción.

4.4.2.3. Matriz de transición

Esta matriz define la probabilidad de que un usuario se traslade de un POI (q_i) al resto de los POIs (incluyendo el estado correspondiente a no conexión), o bien que permanezca en el mismo POI q_i.

4.4.2.4. Matriz de confusión

Esta matriz define la probabilidad de que un usuario se encuentre en un POI dado q_i en el periodo de tiempo que abarca cada observación o_j.

4.4.2.5. Vector

Este vector define la probabilidad de que el usuario inicie su día en un POI q_i.

4.4.3. Entrenamiento del modelo de predicción

En este experimento cada modelo de predicción se entrena considerando los datos correspondientes al periodo de tiempo que abarca la ventana deslizante, ya sea 4 u 8

semanas. Así, para realizar la predicción para un día lunes i, el modelo se entrena con los datos correspondientes a los lunes $i-1$, $i-2$, $i-3$, y finalmente $i-4$, esto cuando el tamaño de la ventana deslizante es de 4 semanas. En el caso de la ventana de 8 semanas el entrenamiento se extiende hasta el lunes $i-8$.

Para realizar el entrenamiento y evaluar el rendimiento de cada modelo de predicción se utiliza la totalidad de los datos disponibles para cada usuario y día de la semana considerado. La cantidad de datos que se utilizaron para cada día de la semana se presenta en la Tabla 5; a lo máximo se cuenta con hasta dos años de datos de localización, lo cual permite evaluar el modelo de predicción en variadas ocasiones.

4.4.4. Predicciones

En cada semana de pruebas y para cada modelo de predicción se realizaron 4 predicciones considerando diferentes valores para ΔT: 1, 2, 4 y 8 horas; 28 predicciones por cada usuario y semana considerada. Al igual que en los experimentos anteriores, los modelos de predicción se definieron utilizando un HMM de primer orden.

4.4.5. Efectividad de la predicción

Al igual que en el enfoque anterior, para determinar la efectividad de la predicción (Ecuación 23), si se desea conocer dónde estará el usuario en el periodo $[T, T+\Delta T]$, la predicción es correcta si el usuario se encuentra en el lugar q_i en el intervalo $[T_{pred}-\epsilon, T_{pred}+\epsilon]$, donde ϵ representa un margen de error. Esto es, la predicción es correcta cuando el usuario está en el POI definido por q_i, en el periodo de tiempo que comprende la observación o_i con cierto margen de error. También es correcta si la predicción indica que el usuario no estará en un POI (en el caso que q_i corresponde a un no POI). El margen de error ϵ se definió a 15 minutos.

En este enfoque los resultados que se obtuvieron no se comparan con el método $NP*$. El objetivo de este experimento no reside en maximizar la precisión, sino conocer el rendimiento del modelo de predicción al considerar la dinamicidad de la movilidad del usuario a lo largo del tiempo.

$T_{pred} = T + o_i \; 1 \leq i \leq$ número de observaciones en el periodo de predicción. (23)

4.5. Experimento 4. Predicción de la movilidad basada en preferencias colectivas

4.5.1. Objetivo del experimento.

El objetivo de este experimento reside en evitar la omisión de puntos de interés para un usuario dado al considerar los lugares que se han visitado por los usuarios similares a éste. Debido a la falta de datos continuos de localización, no se tiene conocimiento completo de la movilidad del usuario, y no es posible identificar en su totalidad los lugares que son significativos para el usuario; el modelo de predicción no reflejará de manera precisa la movilidad de éste. Por lo tanto, en este experimento se infieren los lugares en los que estuvo el usuario al tomar como referencia los lugares que éste y otros usuarios han visitado.

La inferencia de los lugares se logra al utilizar las bases del método filtrado colaborativo utilizado en los sistemas de recomendación.

Al igual que en los experimentos anteriores, en este experimento los modelos de predicción se evaluaron con los datos del proyecto Geolife de Microsoft Research. A continuación se describe este experimento.

4.5.2. Usuarios considerados en el experimento

Con respecto a la selección de los usuarios, ésta se realizó tomando como referencia a aquellos usuarios que comparten un área geográfica en común a fin de obtener la similitud de los usuarios tomando como referencia el aspecto espacial. Considerando que la mayoría de los datos del proyecto Geolife se recolectaron en Beijing, se seleccionaron los usuarios que tienen registros de localización en esta ciudad y en la zona conurbada (Figura 32). Esta área abarca una extensión geográfica de 66 x 52 kilómetros. No se consideró el resto de los usuarios debido a que se encuentran dispersos en distintas ciudades de China, y algunos otros en diversos lugares de Europa y América, por lo que no

tendrían similitud alguna con los usuarios que se seleccionaron. En total se seleccionaron 35 usuarios, aunque el número de usuarios es mayor en el área de Beijing, sólo los 35 usuarios tienen registros continuos de localización, lo que permite realizar el análisis de la movilidad.

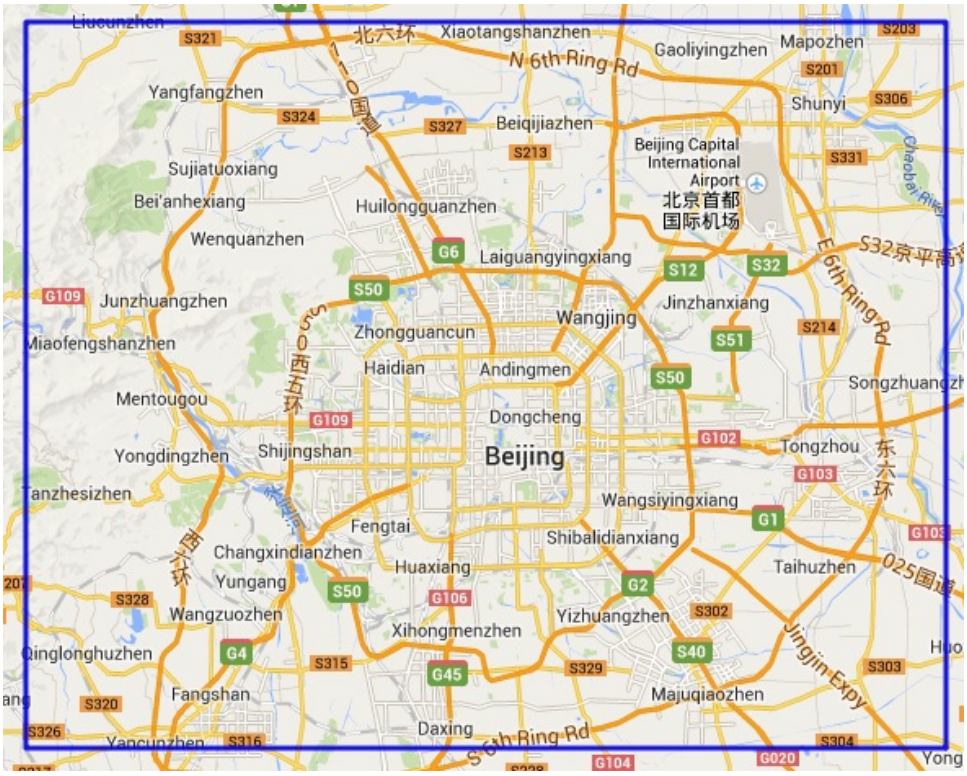


Figura 32: Área geográfica seleccionada para conocer la similitud entre los usuarios.

4.5.3. Puntos de interés

En cuanto a los POIs, se utilizaron aquéllos que se identificaron en el experimento anterior. Así, cada uno de los usuarios cuenta con puntos de interés que abarcan diferentes áreas geográficas con radio de 100, 250 y 500 metros. Estos POIs representan lugares significativos tanto en exteriores como en interiores. Se decidió elegir estos radios de clúster con el objetivo de encontrar una mayor similitud entre los usuarios, y así incorporar los lugares que los usuarios tienen en común como nuevos puntos de interés.

4.5.4. Modelo de predicción

Con respecto a la definición del modelo de predicción, los componentes del HMM están definidos de la siguiente manera:

4.5.4.1. Estados ocultos

Los estados ocultos se encuentran definidos por el conjunto de POIs. Además, se ha agregado un estado adicional para indicar que el usuario se encuentra en un lugar que no se definió como POI. Al igual que en los experimentos anteriores, en este experimento se identificaron POIs en interiores y en exteriores. Sin embargo, al momento de considerar los POIs como estados ocultos no se hace alguna distinción. Al igual, cuando se realiza la predicción no se especifica si el POI se encuentra en exteriores o en interiores.

4.5.4.2. Observaciones

Al igual que en el segundo experimento (Sección 4.3.7), en este experimento las observaciones son definidas de acuerdo al promedio de los tiempos de arribo y partida a cada uno de los puntos de interés.

4.5.4.3. Matriz de transición

Esta matriz define la probabilidad de que un usuario se traslade de un POI (q_i) al resto de los POIs (incluyendo el estado correspondiente a los lugares no definidos como POI), o bien que permanezca en el mismo q_i.

4.5.4.4. Matriz de confusión

Esta matriz define la probabilidad de que un usuario se encuentre en un POI dado q_i en el tiempo definido por la observación $o_j \pm \epsilon$, donde ϵ representa un intervalo de tiempo de 15 minutos.

4.5.4.5. Vector

Este vector define la probabilidad inicial para cada POI. Esto es, la probabilidad de que la primera visita del usuario sea a un POI en particular.

4.5.5. Similitud de los usuarios

Una vez que se definió el área geográfica, se procede a dividirla en celdas de un tamaño dado para definir cada uno de los arreglos r_u. Para ello, se definieron tres tamaños

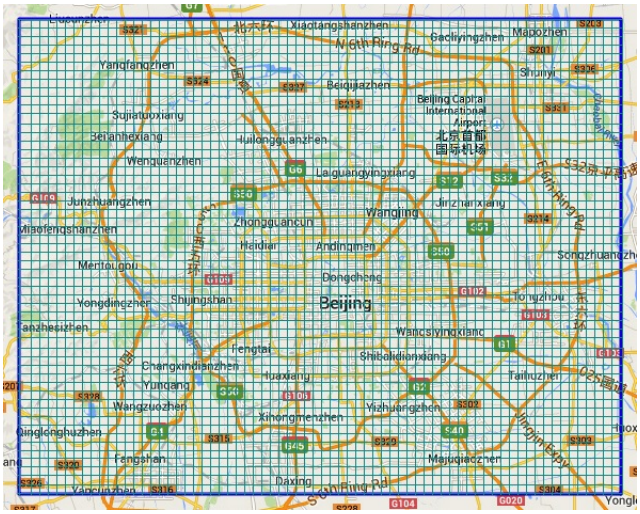
para las celdas: 1000, 500 y 250 metros, esto, debido a que los POIs se identificaron considerando radios de clúster de 500, 250 y 100 metros. Por consiguiente, se definen tres matrices R ($R_{1000}, R_{500}, R_{250}$), una matriz por cada tamaño de celda considerada, así, se conoce la similitud de los usuarios en tres niveles de granularidad.

La división del área geográfica al considerar los tres tamaños de celdas se presenta en la Figura 33. Debido a que cada uno de los modelos de predicción refleja la movilidad del usuario de un día determinado de la semana, cada usuario cuenta con hasta 7 arreglos r_u en cada una de las matrices R; un arreglo por cada día de la semana. En total, cada usuario cuenta con hasta 21 arreglos al considerar las 3 matrices R ($R_{1000}, R_{500}, R_{250}$).

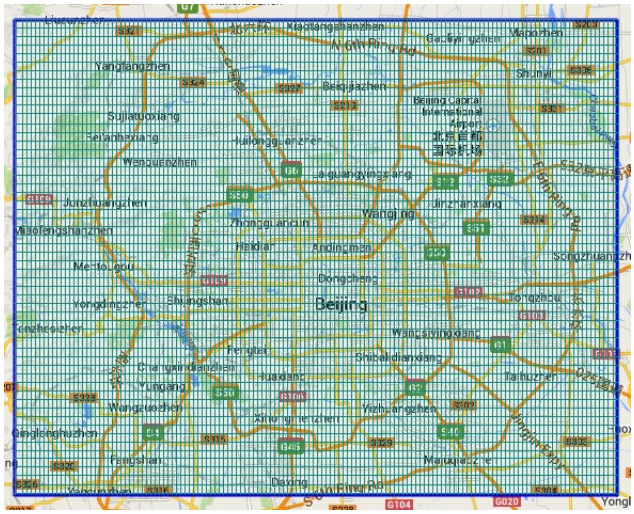
Para definir cada uno de los arreglos de la matriz R_{1000} se utilizan las coordenadas geográficas de los POIs que se identificaron al considerar un radio de clúster de 500 metros. Los arreglos de la matriz R_{500} y R_{250} se definen con las coordenadas de los POIs de 250 metros y 100 metros, respectivamente.

Una vez que se han definido las matrices R, para cada una de ellas se obtiene la similitud de cada uno de los arreglos con respecto a los demás. Cada arreglo representa los lugares que visitó un usuario dado en un día determinado de la semana; cada matriz R contiene hasta 7 arreglos por cada usuario. Así, se conoce la similitud de un usuario dado consigo mismo, y con el resto de los usuarios en cada día de la semana. Para determinar la similitud entre los arreglos se utiliza la similitud del coseno. Posteriormente, para cada arreglo se seleccionan los K arreglo más similares. Se utilizaron tres valores para K: $1, 3,$ y 5, a fin de conocer la influencia de los k similares al momento de realizar el proceso para incorporar nuevos POIs. El valor del umbral θ se definió a 0.75, puesto que al considerar los arreglos con los que se tiene una similitud menor a este umbral no se logran identificar lugares que se consideren POIs.

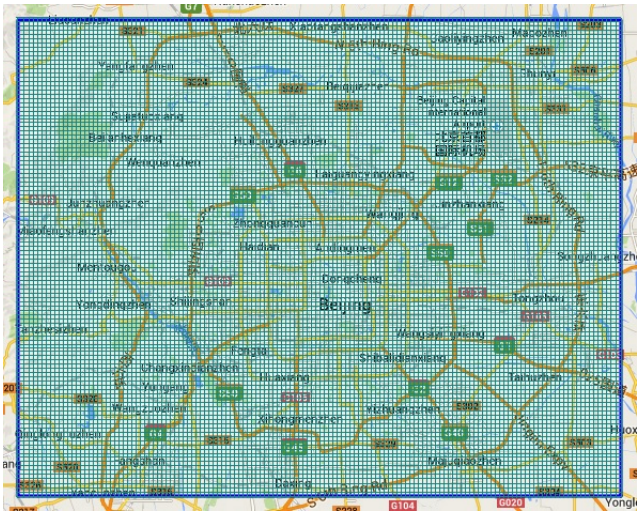
En este experimento sólo se consideraron las características espaciales para determinar la similitud entre los usuarios. Como trabajo futuro se plantea obtener la similitud de los usuarios tomando como referencia el aspecto semántico. Esto es, obtener el significado semántico de los puntos de interés que ha visitado cada uno de los usuarios, y así, conocer que un lugar determinado es un restaurante, bar, hogar, etc. De esta manera,



(a)



(b)



(c)

Figura 33: División del área geográfica; a) celdas de 1000×1000 metros; b) celdas de 500×500 metros; c) celdas de 250×250 metros.

resulta factible comparar no sólo aquellos usuarios que comparten un área geográfica, sino la totalidad de los usuarios que provee el dataset del proyecto Geolife, o algún otro. Además, al considerar la aplicación de la similitud de los usuarios en cuestión de sistemas de recomendación, las recomendaciones se realizarían de manera más puntual; en lugar de recomendar un área geográfica dada, se recomienda un lugar con un significado entendible para los usuarios.

4.5.6. Entrenamiento del modelo de predicción

Con respecto al entrenamiento de cada modelo de predicción, se utilizaron los datos correspondientes al periodo de tiempo en el cual la movilidad del usuario es similar de acuerdo a lo discutido en la Sección 4.3.3 del segundo experimento. Para realizar las pruebas de cada modelo de predicción se utiliza el último mes de registros. Así, el funcionamiento de cada modelo de predicción se puede probar en cuatro días diferentes.

Una vez que se identifica el periodo de tiempo que comprende el patrón de movilidad actual del usuario, se realiza el entrenamiento de cada uno de los modelos de predicción. Posteriormente, se realiza el proceso que determinará la incorporación de nuevos POIs de acuerdo a lo discutido en la Sección 4.5.5. Con el objetivo de conocer la viabilidad y efectividad de la incorporación de nuevos puntos de interés, para cada valor de K considerado se define un nuevo modelo de predicción, si y sólo si, se agregaron nuevos lugares al conjunto base de puntos de interés. De esta manera, para cada uno de los usuarios y día de la semana, se cuenta con hasta 9 modelos de predicción. Esto es, por cada tamaño de POI se cuenta con el modelo de predicción base, y los modelos de predicción resultantes después de considerar la similitud con el arreglo más similar, así como con los 3 y los 5 arreglos más similares. Aunque en la Figura 34 se presentan cuatro modelos de predicción por cada tamaño de POI, sólo se consideraron 3 modelos (base, $K = 1$ y $K = 3$); en la sección de resultados se explica el porqué del número de modelos.

4.5.7. Predicciones

Para realizar la predicción de la movilidad del usuario en la primera semana de pruebas, se utiliza cada uno de los modelos de predicción ($Base$, $k = 1$, y $k = 3$), según sea

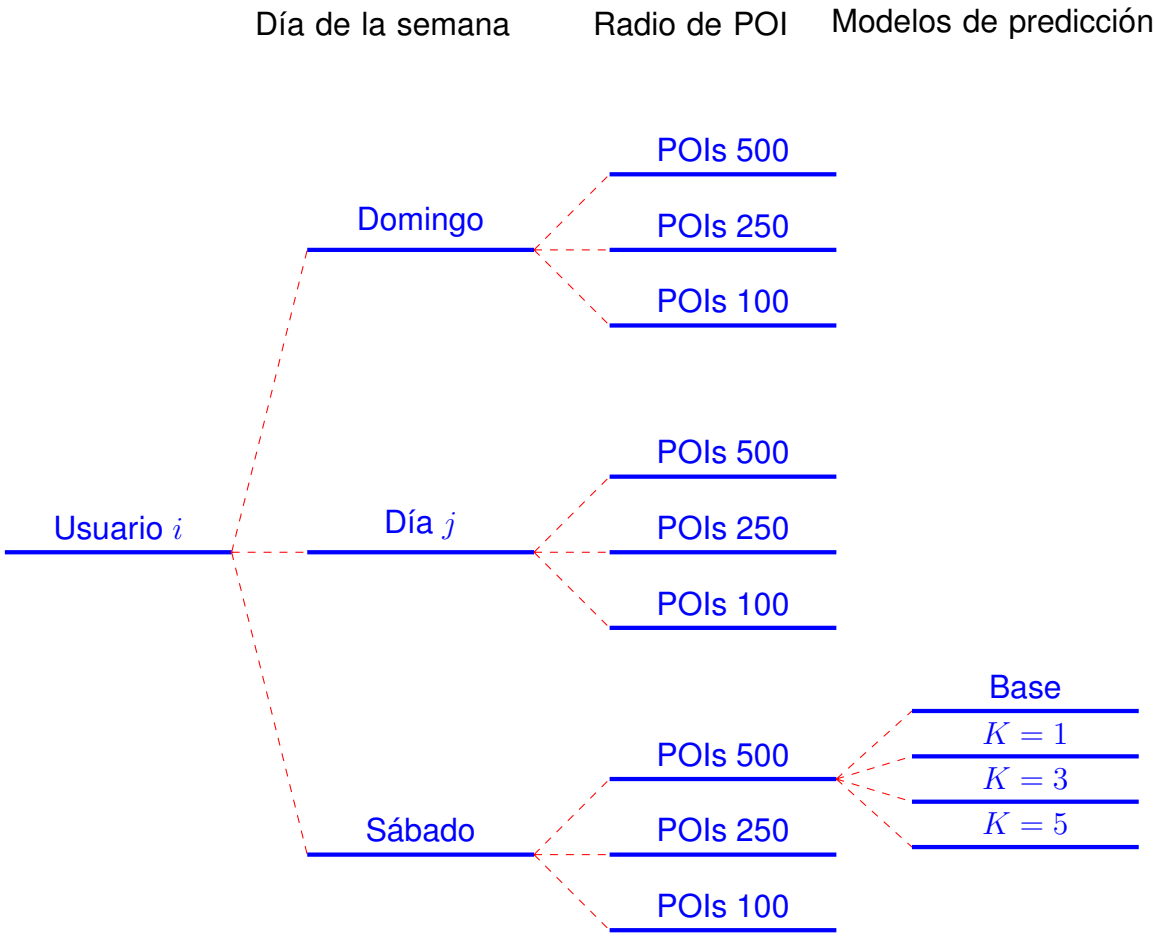


Figura 34: Número de modelos de predicción por usuario.

el caso. Después de que se han realizado las predicciones para la primera semana de pruebas, se compara la movilidad del usuario de este día (*primera semana de pruebas*) con la movilidad del usuario durante el proceso de entrenamiento, utilizando la similitud del coseno. Si la similitud es mayor a un cierto umbral θ, el modelo de predicción se actualiza considerando los registros de este día. En caso contrario, el modelo de predicción se utiliza para predecir la movilidad del usuario en la segunda semana de pruebas. Este proceso se repite para las semanas de prueba siguientes.

En cada una de las semanas de pruebas, cada modelo de predicción se evaluó en 5 ocasiones considerando diferentes valores para ΔT (30 minutos, 1, 3, 5, y 7 horas); 20 predicciones por cada modelo de predicción considerando las cuatro semanas de prueba. Los modelos de predicción se definieron utilizando un HMM de primer orden.

4.5.8. Efectividad de la predicción

Para determinar la efectividad de la predicción (Ecuación 24), si se desean conocer los lugares en donde estará el usuario en el periodo $[T, T + \Delta T]$, la predicción es correcta si el usuario se encuentra en el lugar q_i en el intervalo $[T_{pred} - \epsilon, T_{pred} + \epsilon]$, donde ϵ representa un margen de error. Esto es, la predicción es correcta cuando el usuario está en el POI definido por q_i, en el tiempo definido por la observación o_i con cierto margen de error. También es correcta si la predicción indica que el usuario no estará en un POI (en el caso del q_i correspondiente a un no POI). El margen de error ϵ se definió a 15 minutos.

$$T_{pred} = T + o_i \; 1 \leq i \leq \text{número de observaciones en el periodo de predicción} \tag{24}$$

En este experimento, la comparación de los resultados no se realiza tomando como referencia el método $NP*$; ya que el objetivo de este enfoque es determinar la viabilidad de utilizar la similitud de los usuarios para agregar lugares que no se consideran POIs, y así incrementar la precisión de la predicción, la comparación se realiza tomando como referencia los resultados de las predicciones del modelo de predicción base, y comparándolos con los resultados que se obtienen con los modelos que se definieron al considerar el arreglo y los tres arreglos más similares; cuando es el caso.

4.6. Comparativa de los experimentos

Para finalizar este capítulo, en la Tabla 6 se comparan los experimentos anteriores en función del objetivo, conjunto de datos, número de usuarios, definición del modelo de predicción, y los aspectos que fueron abordados en cada uno de ellos. La esencia del modelo de predicción se mantiene en los cuatro experimentos; la propiedad markoviana al considerar la movilidad del usuario entre los POIs y la relación entre la estadía en los POIs y la hora del día. Las diferencias se encuentran en los aspectos que son abordados en cada uno de los experimentos.

Tabla 6: Comparativa de los experimentos.

	Experimento 1	Experimento 2	Experimento 3	Experimento 4
Objetivo	Conocer el rendimiento del modelo de predicción	Predecir la movilidad en un área urbana	Incluir la dinamicidad de la movilidad	Incluir la escasez de datos de localización
Dataset	Dartmouth	Geolife	Geolife	Geolife
Área geográfica	Campus universitario	Urbana	Urbana	Urbana
Usuarios	200	63	18	35
POIs	Puntos de acceso	Área geográfica que abarca el lugar de interés	Áreas geográfica que abarca el lugar de interés	Áreas geográfica que abarca el lugar de interés
Observaciones	Promedio de los tiempos de inter-contacto	Promedio de los tiempos de arribo y partida a los POIs	Periodos de tiempo de 2,3 y 4 horas	Promedio de los tiempos de arribo y partida a los POIs
Entrenamiento	Dos meses de registros	De acuerdo al periodo que comprende el patrón de movilidad más reciente	De acuerdo al tamaño de la ventana deslizante	De acuerdo al periodo que comprende el patrón de movilidad más reciente
Aspectos	Puntos de interés, Entrenamiento, Observaciones	Puntos de interés, Entrenamiento, Observaciones, Predecibilidad y Similitud de la movilidad	Puntos de interés, Entrenamiento, Observaciones, Dinamicidad	Puntos de interés, Entrenamiento, Observaciones, Escasez de datos

4.7. Resumen

En este capítulo se describe cada uno de los experimentos que se realizaron a fin de conocer la funcionalidad del modelo de predicción. En cada uno de los experimentos se consideran diferentes aspectos de la movilidad, y el enfoque de cada uno de ellos es diferente. La descripción de los experimentos se realiza tomando como referencia el conjunto de datos utilizado, número de usuarios involucrados, la definición de los POIs, la definición de cada uno de los componentes del modelo de predicción basado en HMM, el entrenamiento de los modelos de predicción, y las pruebas de éstos.

Se realizaron varios experimentos debido a que el objetivo es definir un modelo de predicción robusto. Para ello, en cada uno de los experimentos se consideraron diversos aspectos de manera gradual (e.g., predecibilidad, dinamicidad). Además, el enfoque de los experimentos difiere debido a que se utilizan diferentes datos de localización (e.g., registros de conexión a puntos de acceso, y registros de GPS). Con los diversos experimentos se plantea demostrar que el modelo de predicción propuesto obtiene una mejor predicción que el método propuesto por Scellato *et al.* (2011). De igual manera, con los diversos experimentos se desea resaltar la importancia de los aspectos considerados para obtener una mejor precisión, ya sea al realizar la comparación con el modelo presentado por Scellato *et al.* (2011), o bien con el modelo propuesto.

En el Capítulo 5 se presentan los resultados de cada uno de los experimentos. Los resultados se presentan en función de los puntos de interés que se identificaron, y la precisión de la predicción. Además, de acuerdo al experimento, se presentan resultados acerca de la predecibilidad de la movilidad, similitud de los usuarios, similitud de los lugares, entre otros factores.

Capítulo 5. Resultados

En este capítulo se presentan los resultados que se obtuvieron en los diferentes experimentos. Los resultados se presentan en función de la cantidad de puntos de interés que se identificaron, precisión de las predicciones, y predecibilidad de los modelos de predicción. Así también, se presentan aspectos que son relevantes en cada uno de los experimentos.

5.1. Experimento 1. Predicción de la movilidad del usuario en un campus universitario

5.1.1. Puntos de interés

Con respecto a los puntos de interés identificados al utilizar los datos de conexiones a puntos de acceso, cuando sólo se consideraron los registros del mes previo al periodo de pruebas, se identificaron en promedio 5.83 POIs por cada usuario y día de la semana. Al considerar los dos meses previos se identificaron 8.95 POIs, y finalmente al considerar los 3 meses previos se identificaron 9.08 POIs por usuario y día de la semana considerado. Aunque la mayor cantidad de POIs se encontró al considerar los 3 meses de registros, estos POIs no reflejan el comportamiento más reciente del usuario. De igual manera, al considerar solamente los POIs del mes previo, no reflejan de manera adecuada la movilidad del usuario; el usuario visita una cantidad mayor de POIs. En cambio, los POIs que se identificaron con los datos asociados a los dos meses previos corresponden a aquellos lugares que el usuario frecuenta. Por lo tanto, en promedio se consideraron 8.95 POIs para definir cada uno de los modelos de predicción.

5.1.2. Precisión de la predicción.

Con respecto a la precisión que se obtuvo en este experimento, en la Figura 35 se presenta la precisión promedio que se obtuvo tanto por el método $NP*$ como por el modelo de predicción propuesto. Al considerar un periodo de predicción ΔT de 30 minutos, el modelo de predicción propuesto obtiene una mayor precisión, un 85 % en comparación de un 79 % para $NP*$; para un valor de ΔT de 60 minutos se obtiene una precisión de 83 %, y para $NP*$ una precisión de 75 %. Para los periodos de predicción mayores a 60

minutos, la precisión de $NP*$ oscila entre 50 al 70 %, en cambio, para un ΔT de 180 minutos el modelo de predicción propuesto obtiene un 80 % de precisión. Cuando el valor de ΔT se define a 300 minutos, la precisión promedio es de 77 %; finalmente, con un valor para ΔT de 420 minutos, se obtiene una precisión promedio de 73 %. Sin embargo, para los periodos de predicción mayores a 420 minutos, la precisión promedio del modelo propuesto disminuye hasta un 50 %. El hecho de que el modelo propuesto obtiene una mayor precisión, se debe al mecanismo que $NP*$ utiliza.

Cabe mencionar que aunque se tienen 4 semanas para realizar pruebas, los resultados que se presentan sólo corresponden a la precisión que se obtuvo al predecir la movilidad de los usuarios en la primera semana de pruebas. Después de conocer la eficiencia de los modelos de predicción en la primera semana de pruebas, se realizaron pruebas para la siguiente semana (segunda semana de pruebas); sin embargo, la precisión disminuyó. La disminución de la precisión en las semanas de pruebas 2, 3, y 4 se atribuye al hecho de que la movilidad de los usuarios tiene una cierta variación a lo largo del tiempo. De esta manera, el modelo de predicción necesita considerar los cambios en el patrón de movilidad a fin de capturar y modelar de manera adecuada el patrón de movilidad, y así realizar una mejor predicción de ésta. Este aspecto de la movilidad se consideró en los experimentos posteriores, a fin de maximizar la precisión de la predicción.

Resulta de interés el hecho de que la precisión de $NP*$ disminuye a un mayor ritmo que la precisión del modelo propuesto. Este comportamiento puede atribuirse al hecho de que $NP*$ toma como referencia las últimas m apariciones de los últimos 3 lugares que visitó el usuario, para así predecir la próxima ubicación del usuario; estas apariciones no siempre se satisfacen al considerar la movilidad más reciente de los usuarios, sino con datos que no son relevantes a la fecha de la predicción, por lo que existe una diferencia en cuanto al siguiente lugar a visitar y/o a los tiempos de arribo. Debido a ello, se tiene una predicción imprecisa.

Después de conocer la eficiencia del modelo de predicción propuesto, se decidió continuar con experimentos adicionales, los cuales contemplaron diversos aspectos de la movilidad del usuario.

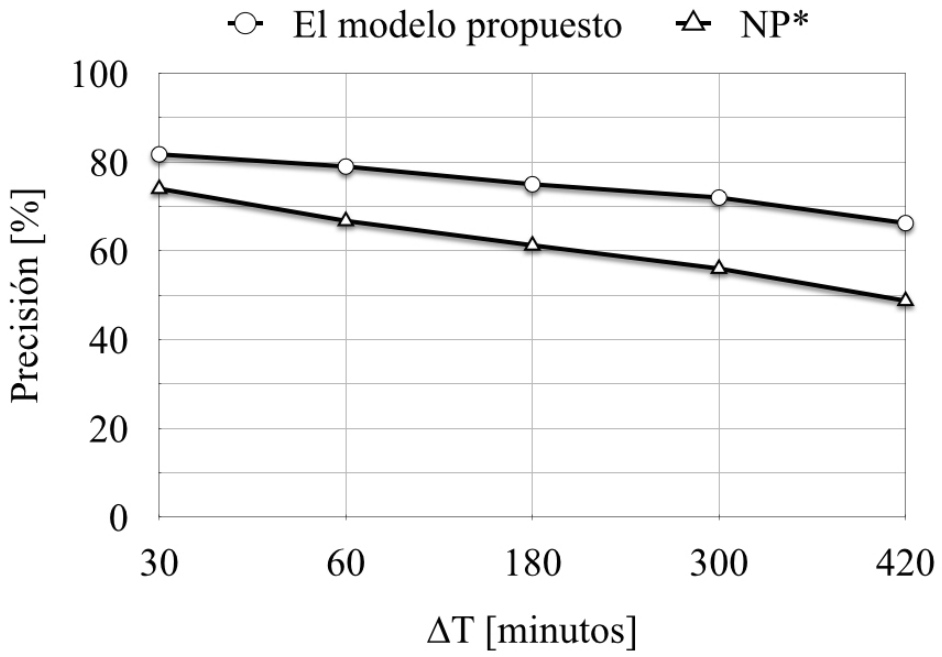


Figura 35: Precisión de la movilidad del usuario dentro de un campus universitario.

5.2. Experimento 2. Predicción de la movilidad del usuario en un área urbana

A diferencia del experimento anterior, en este experimento se presentan los resultados con respecto al periodo de tiempo que comprende el patrón de movilidad actual para cada uno de los usuarios. Así también, se presentan los resultados acerca de la existencia de la propiedad Markoviana en la movilidad del usuario para cada día de la semana.

5.2.1. Puntos de interés

A fin de maximizar el porcentaje de tiempo que los usuarios pasan en los lugares significativos, se utilizaron diversos valores para el tiempo mínimo de estadía y el radio del clúster. Como se observa en la Figura 36, el incremento del porcentaje es mínimo después de considerar un radio de 50 metros (aproximadamente un 1 %). Por lo tanto, tomando como referencia un radio de clúster de 50 metros, en la Figura 37 se presenta el porcentaje del día y cantidad de horas que los usuarios pasan en los puntos de interés al considerar los tres valores del tiempo mínimo de estadía. Como se puede observar, el mayor porcentaje se obtiene cuando el tiempo mínimo de estadía es de 10 minutos, en promedio cada usuario pasa alrededor del 55 % del día en los POIs, lo que equivale a un aproximado de 13.23 horas en los puntos de interés. Por consiguiente, los valores para

el radio del clúster y el tiempo mínimo de estadía se definen a 50 metros y 10 minutos, respectivamente. Cabe mencionar que el tiempo de estadía mínimo corresponde al que sugiere Ashbrook y Starner (2003).

Al maximizar el porcentaje del día que el usuario pasa en los POIs, se incrementa la probabilidad de identificar correctamente los lugares más importantes y significativos para el usuario. Aún cuando el porcentaje promedio que el usuario pasa en los POIs es de 55 %, se obtienen buenos resultados. Lo ideal sería contar con registros continuos de la movilidad del usuario a lo largo del día a fin de tener conocimiento de los lugares que éste visita, y así definir de mejor manera el modelo de predicción.

Con respecto a la cantidad de POIs que se identificaron, en promedio cada usuario tiene 3.69 POIs por día de la semana, el número mínimo de POIs es de 1, con un máximo de 7 POIs para un usuario y día de la semana dado. La cantidad promedio de POIs identificados, corresponde a lo discutido por Chon *et al.* (2012). En su trabajo, Chon *et al.* (2012) argumentan que las personas tienen un alto grado de regularidad al visitar una cantidad limitada de lugares. Posteriormente, al agrupar los POIs de acuerdo a los *días laborales* y *fin de semana*, los usuarios cuentan con más POIs en los días laborales con 3.91. En cambio, los fines de semanas los usuarios cuentan en promedio con 3.16 POIs. Este hecho se puede explicar debido a que en la movilidad de los usuarios en los fines de semanas es en cierta medida más dinámica en comparación de los días laborales; en general no se cuenta con actividades o tareas que se realicen de manera repetitiva. Finalmente, la mayor cantidad de POIs se identificó en los días martes con 4.11, y la menor cantidad de POIs en los días domingo con 3.00 (Tabla 7). Este hecho se atribuye a que en promedio los usuarios cuentan con más registros GPS en los días martes. Con respecto al promedio de registros de los usuarios en los fines de semana, no existe una diferencia significativa entre los registros del sábado y domingo. Por lo tanto, se concluye que la movilidad de los usuarios es más dinámica en los domingos.

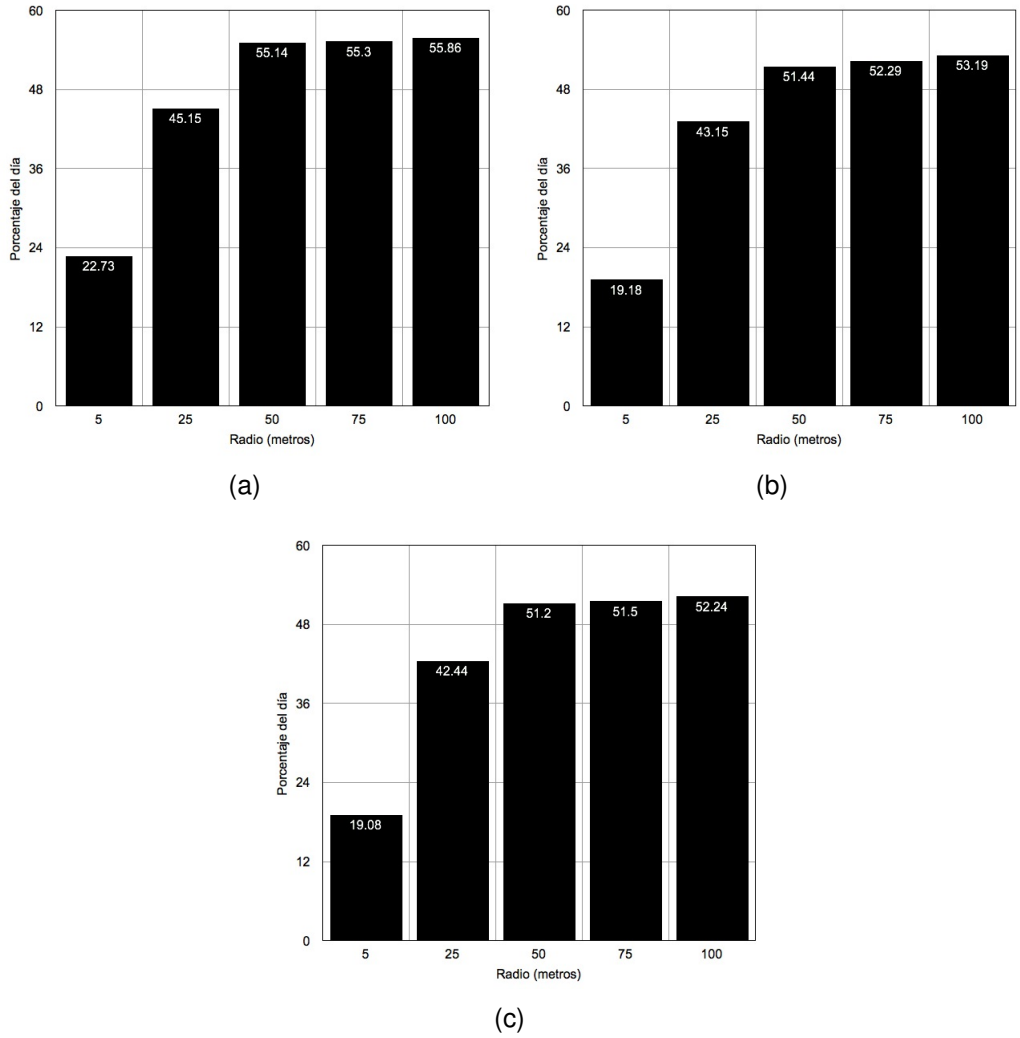


Figura 36: Porcentaje del día que el usuario pasa en POIs de acuerdo a diferentes valores para el tiempo de estadía y el radio del clúster; a) Tiempo mínimo de estadía de 10 minutos b) Tiempo mínimo de estadía de 30 minutos c) Tiempo mínimo de estadía de 60 minutos

Tabla 7: Número de puntos de interés identificados por usuario.

	Min	Max	Promedio
POIs	1	7	3.69
POIs en días laborales	2	7	3.91
POIs en fin de semana	1	6	3.16

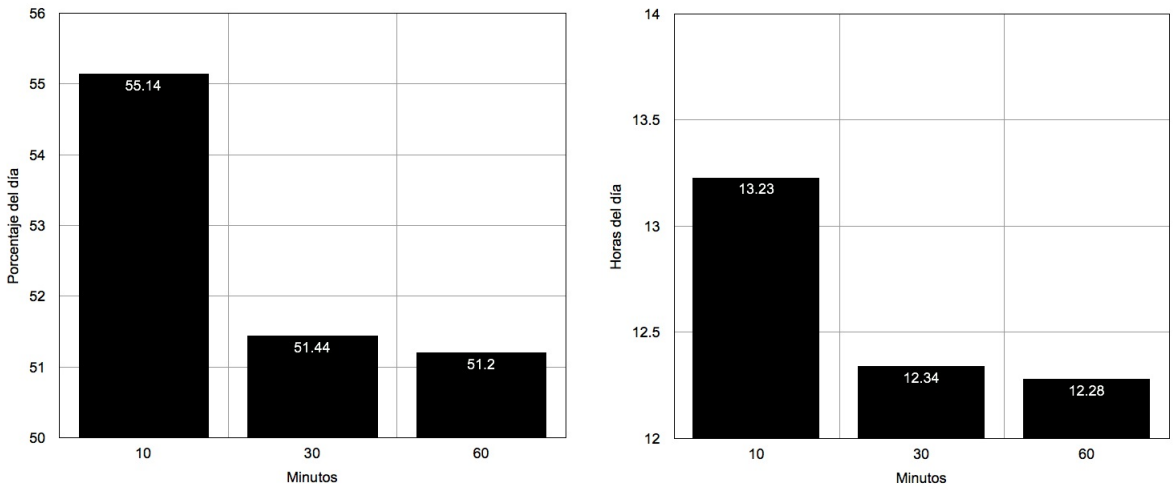


Figura 37: Tiempo que los usuarios pasan en POIs al considerar un radio de 50 metros y varios tiempos de estadía; a) porcentaje del día; b) cantidad de horas

5.2.2. Identificando el periodo de tiempo que comprende el patrón de movilidad del usuario

Considerando un tiempo mínimo de estadía de 10 minutos, y el radio de clúster de 50 metros, se identificó el periodo de tiempo que abarca el patrón de movilidad actual del usuario tomando como referencia una fecha determinada. Los resultados indican que en promedio el periodo de tiempo que comprende el patrón de movilidad es de 7.3 semanas, el periodo mínimo identificado es de 4 semanas, y el periodo máximo es de 11 semanas. Los datos correspondientes a este periodo de tiempo se utilizaron para entrenar cada uno de los modelos de predicción.

5.2.3. Predecibilidad de la movilidad del usuario

Con respecto a la predecibilidad de la movilidad, para cada usuario y día de la semana, se utilizó la prueba definida por Zhang *et al.* (2010). Los resultados demuestran que, en promedio, el 60.31 % de los modelos de predicción de los usuarios son predecibles. Esto es, de los 7 modelos de predicción definidos por cada usuario, la movilidad asociada a 4.22 días de la semana cuenta con la propiedad Markoviana (movilidad del usuario entre los POIs). Por lo tanto, la movilidad de estos días resulta predecible al utilizar los modelos ocultos de Markov. Así también, al agrupar la movilidad de los usuarios considerando los

días *días laborales* y los días de *fin de semana*, los resultados muestran que la movilidad de los usuarios en los fines de semana es más predecible, con un 66.00 %, en lugar de un 60.00 % para los días laborales. A pesar de ello, los resultados fueron mejores que los del método $NP*$, alcanzando hasta un 85 % de precisión al considerar un ΔT de 30 minutos, y hasta un 16 % más que $NP*$.

Al comparar la predecibilidad de la movilidad asociada a cada uno de los días de la semana, los resultados muestran que los días martes tienen el menor porcentaje de predecibilidad, tan sólo 33 %. Esto es, del total de modelos de predicción que se definieron para los martes, sólo el 33 % tiene un resultado positivo para la prueba de Zhang *et al.* (2010). A pesar de que el mayor número de POIs por usuario se identificó al utilizar los registros asociados a los martes, en general, no se tienen registros que permitan identificar la transición del usuario entre los POIs. Por lo tanto, se obtiene un resultado no favorable al utilizar la prueba de predeciblidad. En cambio, la movilidad de los usuarios que corresponde a los días viernes tiene el mayor porcentaje, un 77 %. Esto es, aunque el número de POIs en los viernes es menor que en los martes, se tienen registros acerca de la transición del usuario entre los POIs, lo que permite identificar la existencia de la propiedad Markoviana.

Ahora bien, al considerar el porcentaje de modelos de predicción con un resultado positivo para la prueba de Zhang *et al.* (2010) por cada usuario, se obtuvo que el menor porcentaje es de 14.28 %. Esto es, para un usuario dado, sólo la movilidad de un día de la semana resultó positiva para la prueba de Zhang *et al.* (2010). El mayor porcentaje fue 100 %, esto es, para un usuario la movilidad de los 7 días de la semana resultó predecible al utilizar la prueba de Zhang *et al.* (2010). Al igual que en el párrafo anterior, no todos los usuarios considerados cuentan con registros que permitan conocer la transición de éstos entre los POIs. Es importante mencionar que el porcentaje que se obtiene de la predecibilidad no necesariamente refleja un comportamiento dinámico por parte de los usuarios, este comportamiento se atribuye a la escasez de datos de localización, lo cual impide identificar POIs y en consecuencia definir la transición del usuario entre éstos.

5.2.4. Precisión de la predicción

Como se comentó en la sección de evaluación, para este enfoque se realizaron predicciones en las 4 semanas de pruebas, así cada modelo de predicción se probó en 4 días diferentes. Por lo tanto, los resultados se presentan en dos partes. Primero, en la Figura 38 se presentan los resultados que se obtuvieron en cada una de las semanas de pruebas, tanto por el método $NP*$, como por el modelo de predicción propuesto. Estos resultados representan la precisión promedio que se obtuvo en cada una de las semanas de prueba. Así, el inciso $a)$ presenta la precisión promedio que se obtuvo en la primera semana de pruebas para todos los modelos de predicción; en la segunda semana de pruebas, el 38 % de los modelos de predicción se actualizaron, y la precisión que se obtuvo para esta semana se presenta en el inciso $b)$; mientras tanto, en la tercera semana de pruebas se actualizó el 63 % de los modelos de predicción; el inciso $c)$ presenta la precisión promedio que se obtuvo; finalmente, la precisión promedio correspondiente a la cuarta semana de pruebas se presenta en el inciso $d)$, en la cual se actualizó el 24 % de los modelos de predicción. Además, se obtuvo una precisión de hasta 85, 82,78,75,70 % para los diferentes valores de ΔT (30, 60, 180, 300 y 420 minutos, respectivamente).

En la Figura 39 se presenta la precisión resultante después de promediar los resultados de las 4 semanas de pruebas. Así, al utilizar un periodo de predicción ΔT de 30 minutos, se obtiene una precisión de 81.75 %, lo que representa una mejora con respecto al 74 % que se obtuvo al utilizar el método $NP*$. Al igual, esta mejora se mantiene al utilizar una ΔT de 60 minutos, para este periodo de predicción se obtiene un 79 %, y para $NP*$ sólo se obtiene un 66.75 %. Para el siguiente valor de ΔT (180 minutos), $NP*$ obtiene una precisión de 61.75 %, en cambio el método propuesto obtiene un 75 %; con un ΔT de 300 minutos, se obtiene un 72 % y para $NP*$ una precisión de 56 %; finalmente, para un periodo de predicción de 420 minutos, se obtiene un 66.25 % y 48.75 %, para el modelo de predicción propuesto y $NP*$, respectivamente.

La evaluación demuestra que el modelo propuesto puede obtener una precisión mayor a la que se presenta en el estado del arte al considerar hasta un periodo de predicción de 7 horas. Existe la posibilidad de que la precisión se puede mejorar debido a que existen diversos periodos de tiempo en los cuales no se tienen registros de localización, o bien

las trayectorías GPS no contemplan un periodo prolongado de tiempo. Debido a ello, no fue posible identificar visitas a diversos lugares, definir POIs, o bien conocer la transición del usuario entre los POIs.

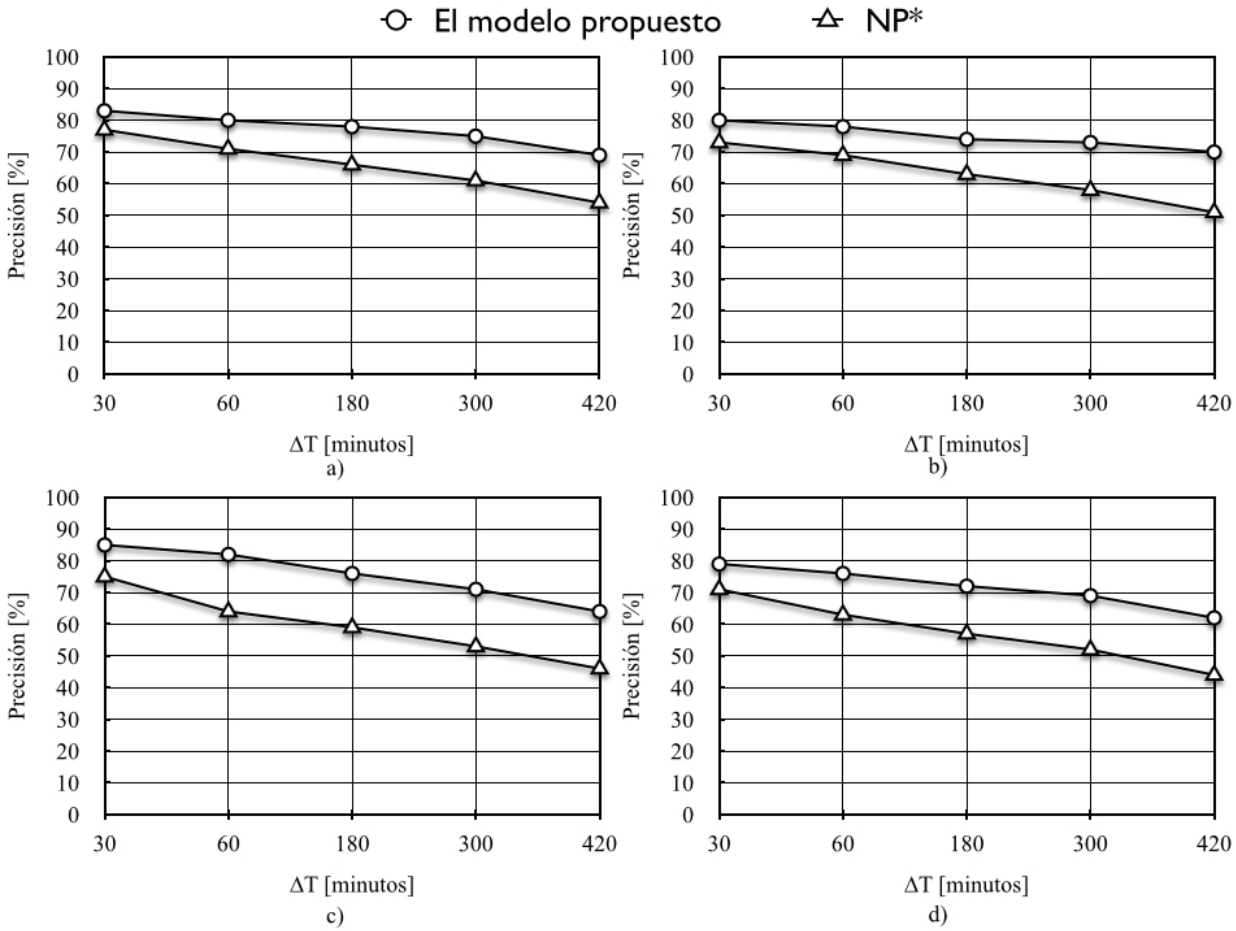


Figura 38: Precisión promedio que se obtuvo en las 4 semanas de prueba.

5.3. Experimento 3. Predicción de la movilidad del usuario a lo largo del tiempo

A continuación se presentan los resultados del tercer experimento, cuyo objetivo fue evaluar el funcionamiento del modelo de predicción a lo largo del tiempo.

5.3.1. Variación en la cantidad de puntos de interés

Al utilizar diferentes valores para el tamaño del radio del clúster, se obtiene una variación en el número de lugares que se identificaron para cada uno de los usuarios (POIs candidatos), y así también, una variación en la cantidad de tiempo que el usuario pasa en estos lugares. Por ejemplo, en las Figuras 40, 41, y 42, se presenta la cantidad de

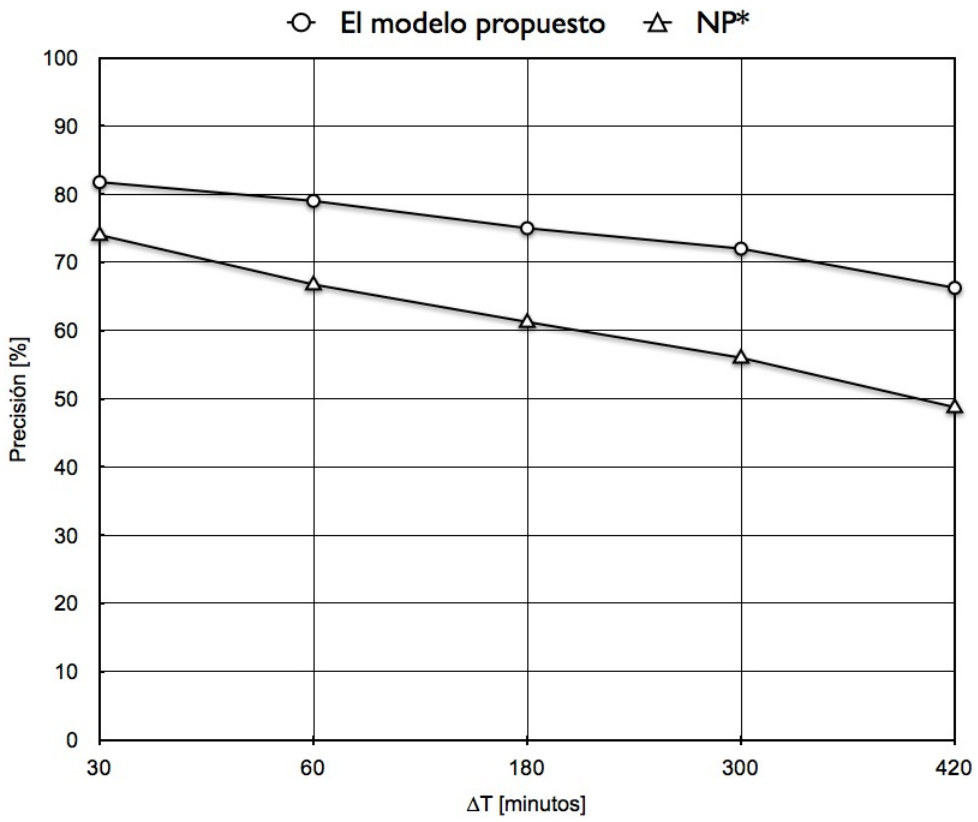


Figura 39: Precisión promedia que se obtuvo considerando las cuatro semanas de prueba.

lugares que se identificaron por los algoritmos al considerar diferentes radios de clúster (100, 250 y 500 metros) en los martes para un usuario con identificador 2. Mediante estas figuras se observan los lugares que el usuario visita, y la cantidad de tiempo que pasa en estos lugares. Cabe mencionar que estos lugares no son puntos de interés, sino aquellos lugares que el usuario visita; puntos de interés candidatos.

Al considerar los registros de localización de este usuario en los martes, se observa una variación significativa en el número de lugares y en la distribución del tiempo; conforme el radio del clúster se incrementa, el número de lugares disminuye de 25 a 14 y posteriormente a 12. Lo anterior se atribuye al hecho de que los lugares se encuentran geográficamente cercanos. De esta manera, los lugares con identificador 1 y 2 que se identificaron con un radio de clúster de 100 metros (Figura 40), se encuentran agrupados en el lugar con identificador 1 cuando se define el radio de clúster a 500 metros (Figura 42). Este comportamiento también se ha identificado para el resto de los días de la semana y para el resto de los usuarios.

A pesar de que para cada usuario se identificaron varios lugares al utilizar diferentes radios de clúster, en las Figuras 40, 41, y 42 se observa que el usuario pasa la mayor parte del día en pocos lugares. Este comportamiento es más notorio al definir el radio de clúster a 250 y 500 metros. Por ejemplo, en la Figura 42 se observa que el usuario pasa gran parte del día en el lugar con identificador 1 y 2. Además, con estas Figuras se corrobora de manera gráfica la existencia de los patrones de movilidad, y la duración de éstos. Así, al considerar la Figura 42 se observa que en la mayoría de los días el usuario visita el mismo lugar a las 9:00 AM; de igual manera, se observa que desde el día $d3$ al día $d10$ en el periodo de las 10:00 a las 18:00 horas, aproximadamente, el usuario se encuentra en el lugar definido con el identificador 1.

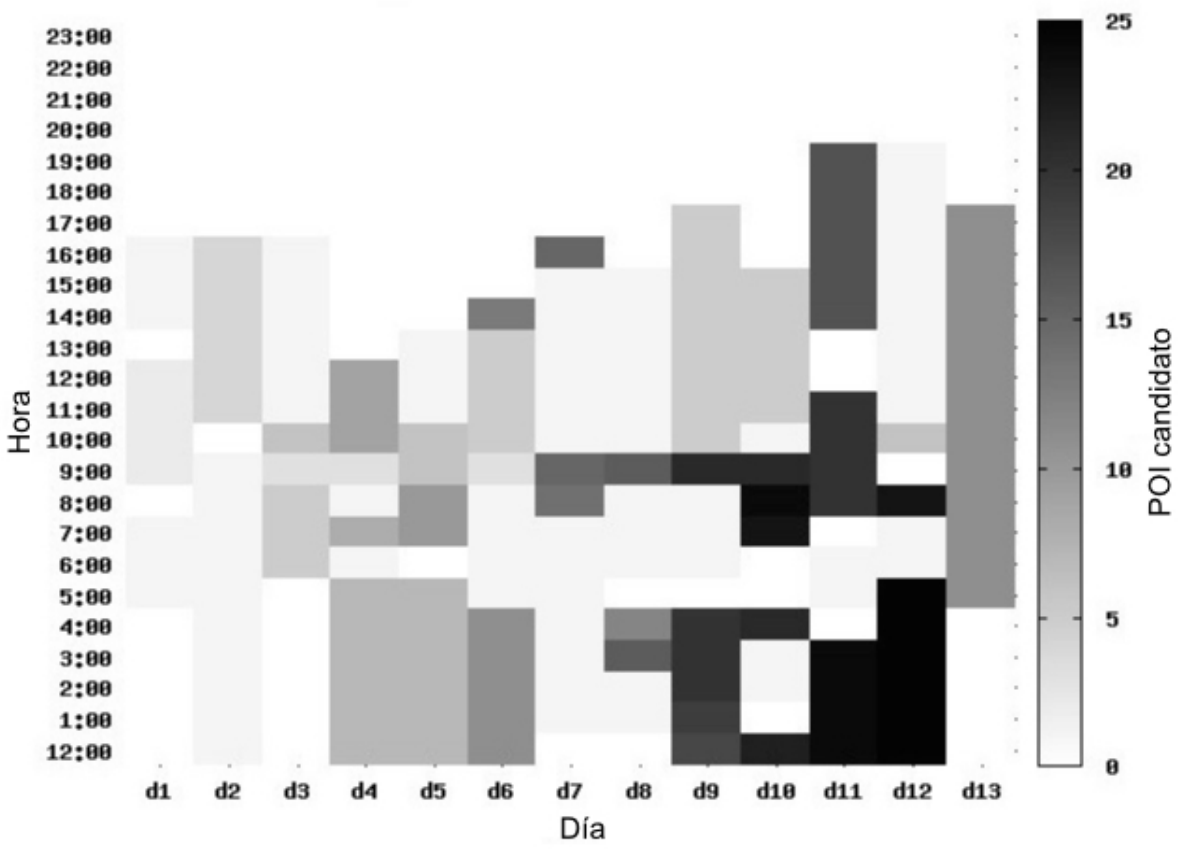


Figura 40: Distribución del tiempo en lugares con radios de clúster de 100 metros.

Después de presentar la distribución del tiempo en los lugares que se identificaron con diferentes radios de clúster, se presentan los puntos de interés. En la Figura 43 se muestra el promedio de POIs por usuario en cada día de la semana al considerar diferentes tamaños de ventana y radios de clúster. En promedio, cada usuario tiene 2.5 POIs

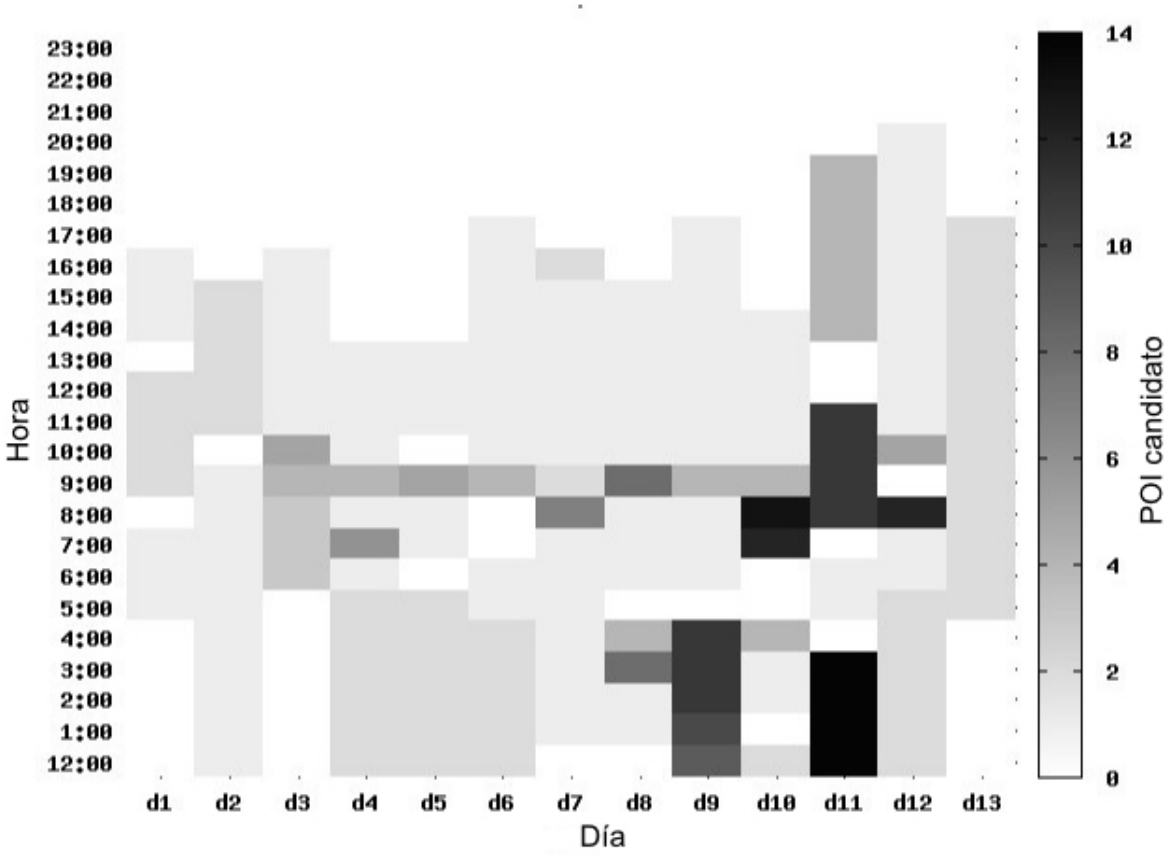


Figura 41: Distribución del tiempo en lugares con radios de clúster de 250 metros.

por cada día de la semana al considerar un radio de clúster de 500 metros y una ventana deslizante de 4 semanas, la cantidad mínima de POIs es de 0, y la cantidad máxima de POIs es de 6 para esta configuración. En cambio, considerando la ventana deslizante de 4 semanas y un radio de clúster de 250 y 100 metros, se identificaron en promedio 1.81 y 1.33 POIs, respectivamente. Al incrementar el tamaño de la ventana deslizante a 8 semanas se obtiene una disminución en la cantidad de POIs identificados, al utilizar un radio de clúster de 500 metros se obtiene un promedio de 1.09 POIs por usuario en cada día de la semana, 0.77 POIs al considerar un radio de 250 metros, y finalmente 0.57 POIs con un radio de 100 metros.

La disminución en la cantidad de POIs conforme se aumenta la ventana deslizante, se atribuye al hecho de que al considerar la movilidad del usuario durante un periodo de tiempo prolongado (e.g., años), sólo un número limitado de lugares son siempre significativos para el usuario, y así, éstos siempre cuentan con visitas. Por ejemplo la casa y el lugar de trabajo del usuario. En cambio, otros lugares son significativos sólo durante un

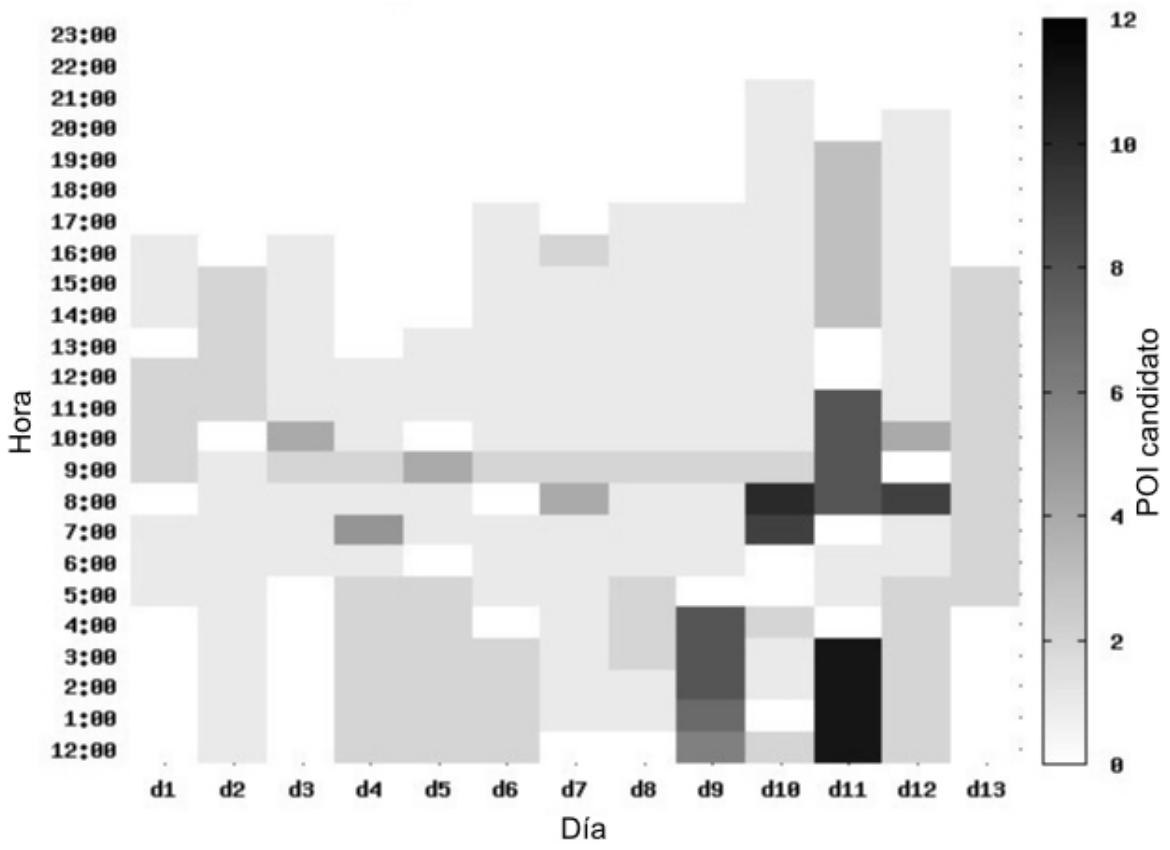


Figura 42: Distribución del tiempo en lugares con radios de clúster de 500 metros.

periodo de tiempo dado (e.g., semanas, meses), tal es el caso de lugares de recreación.

Al considerar un radio de clúster de 500 metros, la ventana deslizante de 4 semanas, y agrupando los POIs de acuerdo a los días laborales y fin de semana, en promedio los usuarios tienen más POIs en los días laborales (2.23); en cambio, los usuarios sólo cuentan con 1.93 POIs el fin de semana. La mayor cantidad de POIs se obtiene en los viernes (2.33), y la menor cantidad en los domingos (1.89). Aún cuando no se cuenta con una gran cantidad de POIs para cada usuario, ni una gran cantidad de registros GPS, los usuarios pasan aproximadamente la mitad del día en los lugares definidos como POIs. En la Tabla 8 se presenta el porcentaje del día que los usuarios pasan en los POIs de acuerdo al día de la semana. Además, se presenta la cantidad de semanas durante las cuales los usuarios cuentan con registros de localización en cada uno de los días de la semana. Por ejemplo, se tiene que un usuario cuenta con registros de localización durante 94 martes, o bien que un usuario cuenta con registros durante 74 domingos.

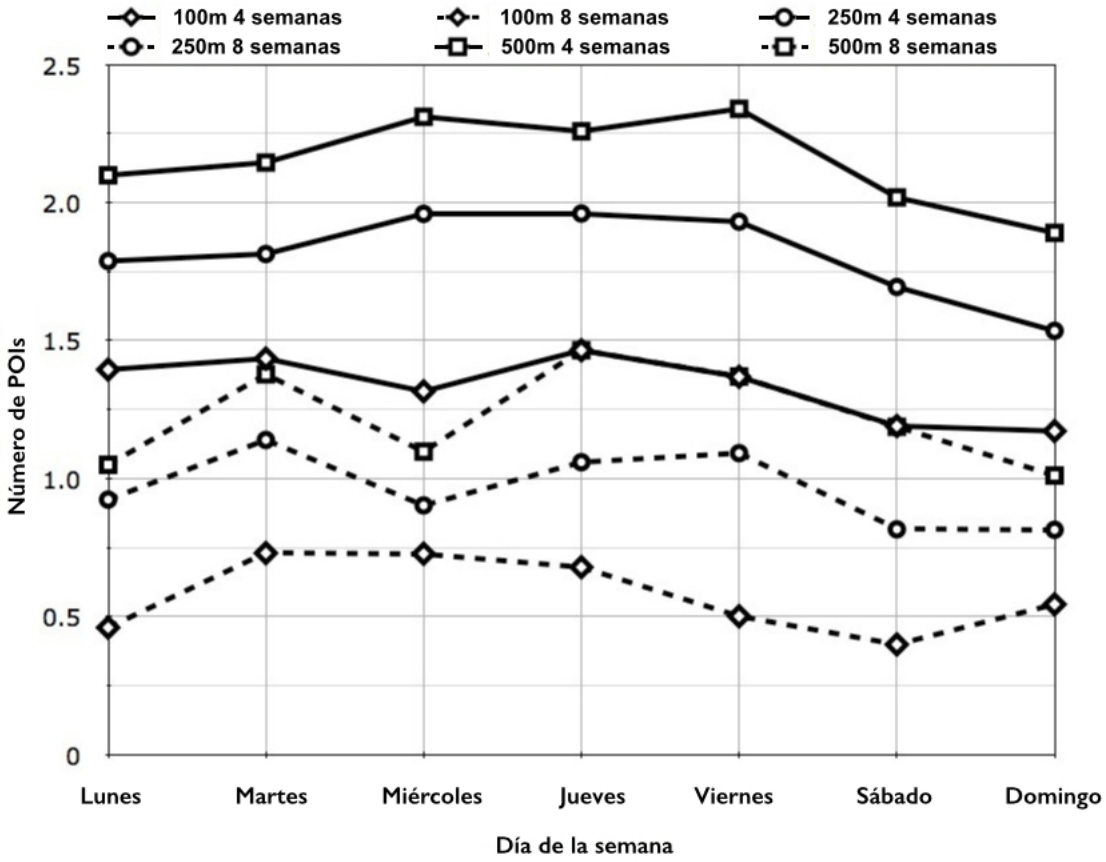


Figura 43: Número de POIs de acuerdo al radio de clúster y ventana deslizante.

Debido a que los usuarios no cuentan con registros de localización durante todo el periodo de sensado, o bien los usuarios tienen datos escasos, para algunas fechas no fue posible definir un modelo de predicción. Por lo tanto, en la Figura 44 se presenta el porcentaje de semanas para las cuales fue posible definir un modelo de predicción. Esto es, semanas para las cuales se identificaron puntos de interés en el periodo de tiempo que abarca la ventana deslizante, y fue posible definir el modelo de predicción. Como se observa en la Figura 44 el mayor porcentaje se obtiene cuando los modelos de predicción se definen con una ventana deslizante de 4 semanas y un radio de clúster de 500 metros. Los porcentajes que se obtienen al considerar una ventana de 4 semanas y radios de clúster de 500 y 250 metros, superan a los porcentajes que se obtienen al definir la ventana deslizante a 8 semanas. La disminución en el porcentaje se debe a la ausencia de POIs al utilizar una ventana deslizante de 8 semanas.

Tabla 8: Porcentaje del día que los usuarios pasan en los puntos de interés y semanas durante las cuales se tienen registros de localización.

	Lu	Ma	Mi	Ju	Vi	Sa	Do
Porcentaje	42.95	52.52	52.78	52.58	49.58	44.90	33.50
Promedio de sem. con registros	15	16.29	16.35	17.47	16.64	16.58	18
Máximo # de sem. con registros	82	94	84	87	89	87	74
Mínimo # de sem. con registros	9	12	11	13	13	11	9

5.3.2. Precisión de la predicción

Debido a las diferentes configuraciones que se consideraron en este experimento, a continuación sólo se presentan los resultados más significativos. Esto es, cuando la ventana deslizante se definió a 4 semanas con un radio de clúster de 500 metros. Así, en la Figura 45 se presenta la precisión promedio que se obtuvo al considerar la configuración anterior y distintos valores para las observaciones. La precisión representa el promedio que se obtuvo después de actualizar el modelo de predicción en todas la semanas disponibles para cada uno de los usuarios y días de la semana. De esta manera, cuando el periodo de predicción (ΔT) es de 1 hora, se obtiene una precisión promedio de 41 %, para un valor de ΔT de 2 horas se obtiene una precisión de 38 %, para un ΔT de 4 horas se obtiene una precisión de 39 %, y finalmente para un ΔT de 8 horas se obtiene una precisión de 43 %. La precisión máxima que se obtuvo fue de 77 % para un ΔT de 1 hora, y 72, 72, y 69 cuando el periodo de predicción se define a 2, 4, y 8 horas, respectivamente.

El objetivo de utilizar diversos valores para las observaciones es el de encontrar el valor que maximice la precisión de las predicciones. Sin embargo, como se observa en la Figura 45, ninguno de los valores considerados fue superior para todos los valores de ΔT. Para el caso de un ΔT con valor de 1 y 2 horas, el mejor valor para las observaciones fue de 2 horas; para un ΔT de 4 horas, el mejor valor para las observaciones fue de igual

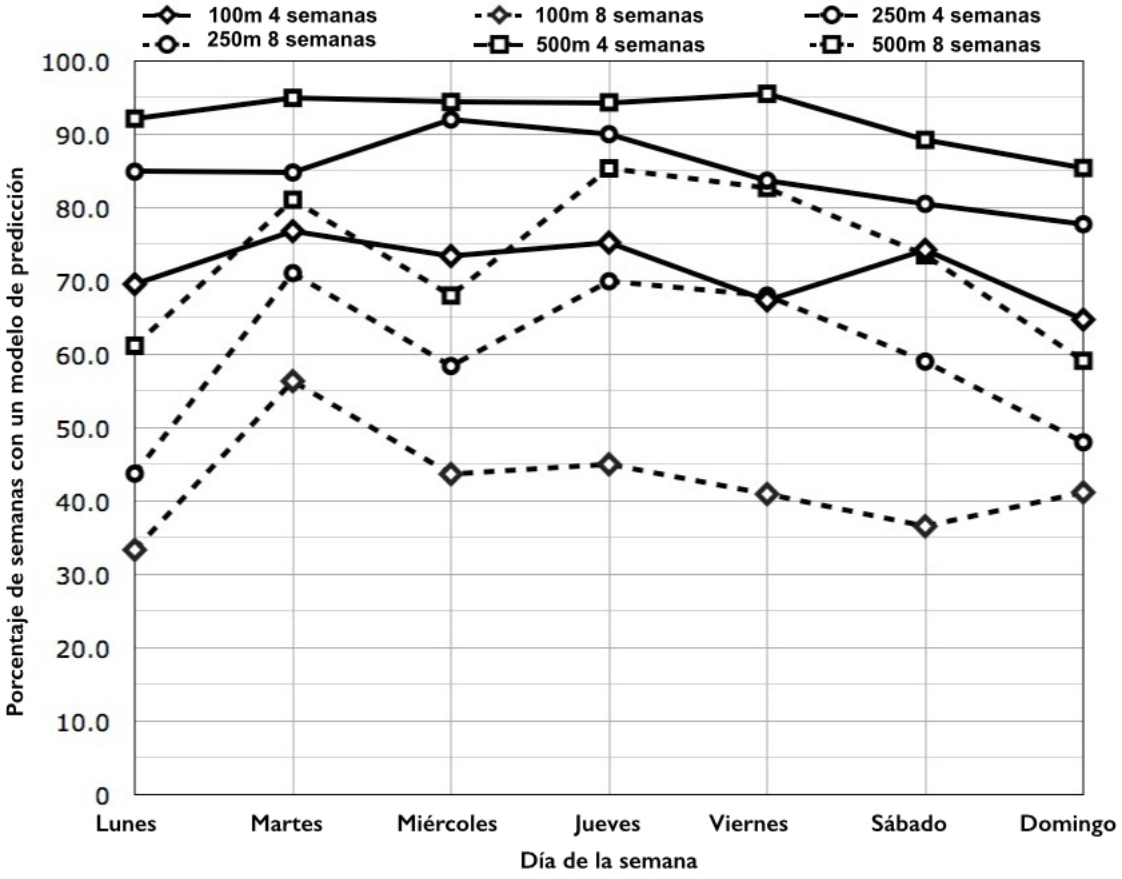


Figura 44: Porcentaje de semanas para las cuales se definió un modelo de predicción de acuerdo al radio del clúster y tamaño de ventana deslizante.

manera de 4 horas; finalmente, para un ΔT de 8 horas, la precisión es muy similar para los 3 valores de las observaciones.

Después de analizar los resultados anteriores, surgen varios comentarios. El hecho de que la mayor precisión se obtuvo al considerar un ΔT de 8 horas, se explica debido a que al tomar dicho periodo de predicción se considera una mayor cantidad de observaciones. Cuando se consideran valores de 1, 2 y 4 horas para ΔT, estos periodos de predicción se pueden ser incluir o cubrir por sólo una observación. En este caso, el POI seleccionado como resultado de la predicción, es aquel con mayor probabilidad de aparecer en dicho periodo. En cambio, si se considera un periodo de predicción ΔT mayor, se toman como referencia más observaciones y así también las probabilidades de la matriz de transición y confusión tendrán un rol importante para identificar la secuencia de POIs en los cuales estará el usuario.

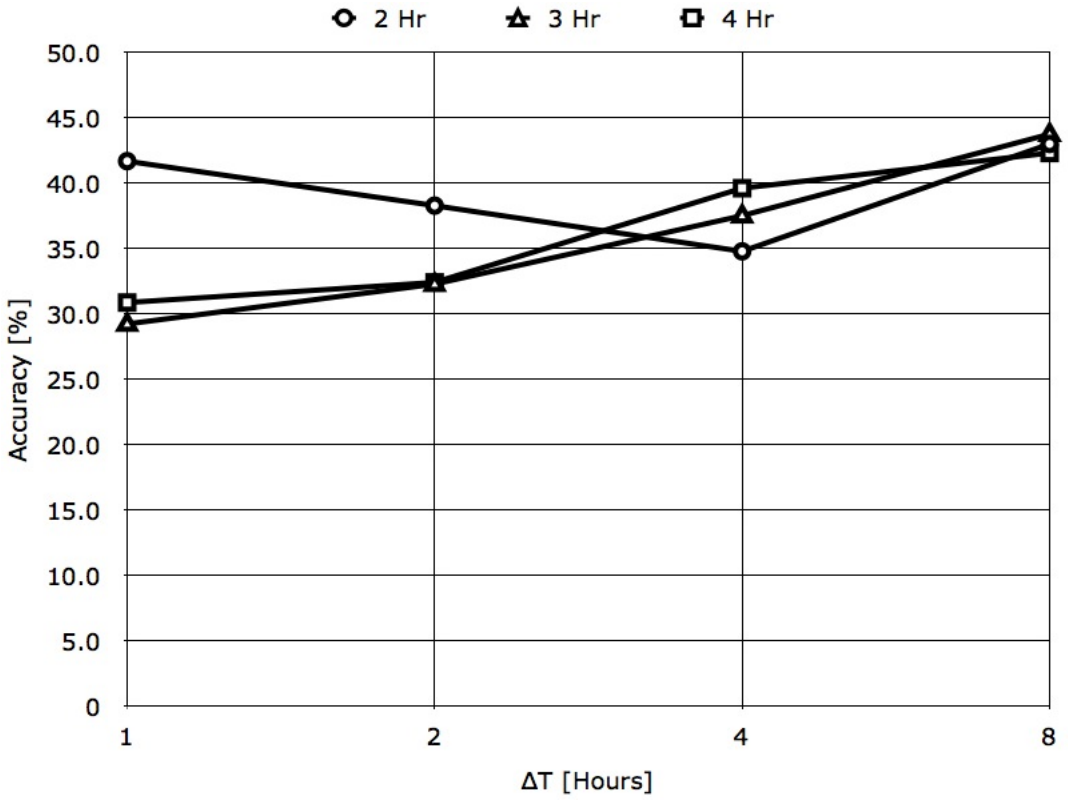


Figura 45: Precisión que se obtuvo al considerar un radio de clúster de 500 metros y una ventana deslizante de 4 semanas.

Considerando lo anterior, sería deseable considerar un periodo de predicción ΔT más grande para conocer el rendimiento de los modelos de predicción. Sin embargo, la mayoría de las trayectorias GPS están distribuidas en un periodo limitado del día por lo que no es posible considerar un periodo de predicción mayor.

Con respecto a la precisión que se obtuvo al utilizar un tamaño de ventana deslizante de 4 semanas, y los radios de clúster de 250 y 100 metros, se obtuvo una disminución en la precisión de alrededor de 20 % y 30 %, respectivamente. Al aumentar el tamaño de la ventana deslizante a 8 semanas, la precisión disminuyó considerablemente, se obtuvo un promedio de 30 % al considerar la precisión para los 3 valores del radio del clúster.

5.4. Experimento 4. Predicción de la movilidad del usuario tomando como referencia las preferencias colectivas

Con respecto a los resultados en el cuarto experimento, a continuación se presenta la cantidad de puntos de interés que se identificaron en cada uno de los días de la semana para cada uno de los usuarios. Así también, se presenta la similitud de los usuarios con

Tabla 9: Cantidad de POIs identificados por usuario y día de la semana.

Tamaño de clúster (m)	POIS
500	3.65
250	3.62
100	3.20

respecto a los lugares que visitan, y finalmente se presenta la similitud de los lugares involucrados.

5.4.1. Puntos de interés

En cuanto a los puntos de interés por cada usuario y día de la semana, en la Tabla 9 se presenta la cantidad promedio de POIs que se identificaron de acuerdo al radio del clúster considerado. Como se puede observar en la tabla, no hay mucha variación en el número de POIs cuando se utilizan diferentes radios de clúster.

Una vez que se han identificado los POIs, éstos se mapearon a las celdas correspondientes de acuerdo a sus coordenadas geográficas, y a los diferentes radios de clúster con el objetivo de definir cada uno de los arreglos r_u.

La Tabla 10 presenta la cantidad de arreglos definidos al considerar los diferentes tamaños de celdas. Como se mencionó en la sección anterior, para este experimento se consideraron los registros de localización de 35 usuarios, de esta manera si por cada usuario y día de la semana se define un arreglo r_u, se tendría un total de 245 arreglos en cada matriz $R(R_{250}, R_{500}, R_{1000})$. Sin embargo, no todos los usuarios cuentan con puntos de interés en cada uno de los días de la semana por lo que se obtuvo un número menor de arreglos. Se obtuvieron 232 arreglos para cada una de las matrices R.

Tabla 10: arreglos definidos al utilizar diferentes valores para el tamaño de celda, y arreglos para los cuales se tienen al menos un arreglo con similitud mayor a cierto umbral.

Tamaño de celda (m)	arreglos	$k = 1$	$k = 3$	$k = 5$
1000	232	136	63	28
500	232	114	60	38
250	232	63	1	0

5.4.2. Similitud de los usuarios

Después de definir cada arreglo r_u, y por consiguiente cada una de las matrices $R(R_{250}, R_{500}, R_{1000})$, se procedió a obtener la similitud entre los arreglos. De esta manera para cada una de la matrices R, se obtuvo la similitud de cada arreglo con respecto al resto de los arreglos. Para los propósitos de este experimento, resultan relevantes aquellos arreglos cuya similitud es mayor a un cierto umbral (Θ); al tener una mayor cantidad de lugares en común, estos arreglos resultan viables para realizar el proceso de incorporación de nuevos puntos de interés. Sin embargo, no para todos los arreglos existe otro arreglo o arreglos con los cuales se tenga una similitud mayor al umbral. De esta manera, en la Tabla 10 se presenta la cantidad de arreglos para los cuales existe al menos un arreglo con el que se obtiene la similitud mayor al umbral Θ. Así también, se presenta la cantidad de arreglo que cuentan con 3 y 5 arreglos similares. El umbral Θ se definió en 0.75, debido a que después de realizar diferentes pruebas con varios valores para Θ, a partir de 0.75 no existe una diferencia significativa en cuanto a la incorporación de nuevos POIs.

Al observar la Tabla 10, resulta de interés que conforme se aumentó el tamaño de la celda, se obtuvo un incremento en la cantidad de arreglos para los cuales se tiene un arreglo similar. Esto se atribuye al hecho de que los usuarios tienen una mayor similitud al considerar los lugares que abarcan un área geográfica mayor. En cambio, al considerar los lugares visitados que abarcan un área geográfica menor, la similitud disminuye. Por ejemplo, al considerar los usuarios del conjunto de datos del proyecto Geolife, la mayoría

de ellos tienen como lugar en común la universidad de Beijing (área geográfica a nivel campus). Pero al considerar un área menor (e.g., edificio, facultad), el número de usuarios con lugares en común disminuye, y así sucesivamente conforme el área considerada es más granular.

Con respecto a la similitud entre los arreglos, los resultados se presentan considerando dos aspectos: *considerando un arreglopor cada día de la semana*, y *considerando un arreglo por usuario*.

5.4.2.1. Considerando un arreglo por usuario

Al obtener la similitud entre los usuarios sin hacer distinción del día de la semana, se obtienen los resultados que se presentan en la Tabla 11. Para fines prácticos se presenta el promedio de la similitud al considerar el usuario más similar, así como los 3, y los 5 usuarios más similares ($k = 1, 3, 5$). De esta manera, al considerar la similitud de los arreglos correspondientes a la matriz R_{1000}, la similitud promedio que se obtuvo fue de 0.82 al considerar el usuario más similar, y de 0.78 y 0.73 al considerar el promedio de la similitud de los 3 y los 5 usuarios más similares con respecto a un usuario dado. En el caso de la matriz R_{500}, se obtuvo 0.84, 0.80, y 0.77 para los 1, 3 y 5 usuarios más similares, respectivamente. Finalmente, para la matriz R_{250} se obtuvo un 0.77, 0.72 y 0.71, respectivamente.

Tabla 11: Similitud entre los usuarios al considerar un único arreglo por usuario.

Tamaño de celda (m)	k=1	k=3	k=5
1000	0.82	0.78	0.73
500	0.84	0.80	0.77
250	0.77	0.72	0.71

5.4.2.2. Considerando un arreglo por cada día de la semana

Cuando se definió un arreglo r_u por cada usuario y día de la semana, la similitud resultó ser mayor que la presentada en la Tabla 11. La similitud resultante se presenta

en la Tabla 12. Como se puede observar, conforme el tamaño de la celda se incrementa, así también se incrementa la similitud al considerar los k arreglos más similares. Cabe mencionar que al considerar un tamaño de celda de 250 metros, ningún arreglo tuvo 5 arreglos similares (k=5).

Aunado a lo anterior, en la Tabla 13 se presenta la similitud promedio mínima y máxima al considerar los arreglo asociados a cada día de la semana. Un aspecto interesante es que, al considerar diferentes tamaños de celdas y valores de k, el promedio máximo se obtiene al tomar como referencia los días laborales. Este fenómeno se puede explicar por el hecho de que la movilidad de los usuarios en los fines de semana (sábado y domingo), tiene una cierta variación, por lo que no se encuentra una cierta similitud con respecto a los demás días de la semana, o bien con la movilidad de otros usuarios. Otro aspecto a destacar, es el hecho de que los arreglos asociados a los días laborales tienen una mayor similitud al compararlos con otros arreglos. Por ejemplo, cuando se tiene un tamaño de celda de 1000 metros, el 41.14 % de los arreglos asociados a los días laborales tiene al menos un arreglo con el cual se obtiene una similitud mayor al umbral. En cambio, sólo el 24.63 % de los arreglos asociados a los días de fin de semana cuenta con al menos un arreglo similar. Este comportamiento se identifica al observar los diferentes tamaños de celda y valores de k (Ver Tabla 14)

Para cada una de las matrices R se presenta la similitud mayor y menor que se obtuvo para cada valor de k, así como el día que se tomó como referencia para obtener dicha similitud.

De esta manera, al considerar la similitud resultante al comparar los arreglos de la matriz R_{1000}, la mayor similitud que se obtuvo al considerar sólo el arreglo más similar, es de 0.88, y se obtiene al tomar como referencia los arreglos correspondientes a los lunes. En cambio, al considerar el arreglo más similar ($k = 1$), la menor similitud es de 0.77 cuando se tomaron como referencia los arreglos asociados a los sábados. Esto es, los arreglos asociados a los lunes y por ende los lugares visitados en este día, tienden a tener una mayor similitud con los arreglos asociados a otros días. En cambio, para los arreglos asociados a los sábados se obtiene la menor similitud. Tanto la mayor como la menor similitud disminuye conforme se incrementa el valor de k al considerar los arreglos

de las tres matrices R.

Lo anterior se atribuye al hecho de que en promedio, los usuarios cuentan con una mayor cantidad de POIs candidatos en los lunes, lo cual da la pauta para identificar usuarios similares. En cambio, los usuarios cuentan con una menor cantidad de POIs candidatos en los sábados, lo cual evita identificar usuarios similares al tomar como referencia los POIs candidatos asociados a este día; los usuarios

Tabla 12: Similitud promedio que se obtuvo al considerar un arreglo por cada usuario y día de la semana, y diversos valores para k.

Tamaño de celda	k=1	k=3	k=5
1000	0.8796	0.8727	0.8499
500	0.8715	0.8530	0.8353
250	0.8470	0.8208	

Tabla 13: Similitud entre los usuarios al considerar un arreglo por cada día de la semana.

Tamaño de celda (m)	k=1	Día	k=3	Día	k=5	Día
1000	0.920	Lunes	0.900	Lunes	0.863	Lunes
	0.861	Sábado	0.853	Martes	0.838	Martes
500	0.901	Martes	0.867	Lunes	0.849	Domingo
	0.830	Viernes	0.815	Viernes	0.802	Viernes
250	0.874	Martes	0.826	Viernes		
	0.782	Lunes	0.786	Lunes		

5.4.2.3. Similitud del usuario consigo mismo

Después de obtener la similitud de cada arreglo en cada una de las matrices R, resulta de interés el conocer la similitud del usuario consigo mismo. La motivación se centra en el

Tabla 14: Porcentaje de arreglos que tienen k arreglos similares (similitud $>$ umbral) de acuerdo al tipo de día: WD: día laboral, y WE: fin de semana.

Tamaño de celda	k=1	k=3	k=5
1000 WD	41.14	14.85	8.57
1000 WE	24.63	5.79	0
500 WD	44.57	23.42	11.42
500 WE	20.28	10.14	4.34
250 WD	17.33	.06	0
250 WE	3.38	0	0

hecho de que conforme se considera una mayor cantidad de usuarios, mayor es la cantidad de arreglos que se tienen, y así el proceso de obtener la similitud de cada arreglo con respecto al resto se incrementa de manera exponencial. Por lo tanto, a fin de evitar el cálculo de la similitud de cada arreglo con el resto de los arreglos, una opción viable es conocer la similitud entre los arreglos correspondientes a cada uno de los usuarios. Por consiguiente, se procede a identificar la cantidad de ocasiones en las cuales los k arreglos más similares corresponden al mismo usuario. Por ejemplo, considerando un tamaño de celda de 1000 metros y k=5, de un total de 28 arreglos que cuentan con 5 arreglos similares, en 4 ocasiones, los 5 arreglos más similares correspondieron al mismo usuario. Esto es, el 14.28 % de las ocasiones. De manera similar, al considerar un tamaño de celda de 500 metros y k=1, de un total de 114 arreglos que cuentan con 1 arreglo similar, en 97 ocasiones el arreglo más similar corresponde al mismo usuario (el 85.08 % de las ocasiones). Cuando el tamaño de la celda es de 250 metros, sólo una pequeña cantidad de arreglos cuentan con un arreglo similar (63), y en 49 ocasiones, estos arreglos corresponden al mismo usuario (77.77 % de las ocasiones). Sin embargo, al considerar k=3, sólo un arreglo cuenta con 3 arreglos similares, y dichos arreglos corresponden al mismo usuario. Finalmente, cuando k=5, ningún arreglo cuenta con 5 arreglos similares. Conforme el tamaño de las celdas disminuye, menor es la cantidad de arreglos simila-

res. Considerando los resultados anteriores, es factible obtener únicamente la similitud entre los arreglos asociados a un usuario dado, debido que al considerar los k arreglos similares, en la mayoría de las ocasiones, estos arreglos corresponden al mismo usuario. De esta manera, el proceso de incorporación de POIs se puede realizar considerando únicamente los arreglos asociados a un usuario dado.

De acuerdo a lo anterior, en la Tabla 15 se presenta la cantidad de ocasiones en las cuales para un arreglo dado, el arreglo más similar ($k = 1$) corresponde al mismo usuario. Así también, para $k = 3$ y $k = 5$, esto es, los 3 y los 5 arreglos más similares corresponden al mismo usuario. Aunque en trabajos relaciones se demuestra que la movilidad de los usuarios difiere para día de la semana, estos resultados indican que los arreglos asociados a los lugares que visita un usuario tienen una cierta similitud.

Tabla 15: Cantidad de ocasiones (primera fila) y porcentaje (segunda fila) en las que los k arreglos más similares corresponden al mismo usuario.

Tamaño de celda	k=1	k=3	k=5
1000	92	31	46
	67.64	49.20	14.28
500	97	25	4
	85.08	41.66	10.52
250	49	1	
	77.77	100	

Debido a estos resultados, resulta viable la opción de únicamente obtener la similitud entre los arreglos asociados a un usuario en particular, evitando así el obtener la similitud con el resto de los usuarios. De igual manera, resulta viable realizar el proceso de incorporación de nuevos puntos de interés considerando únicamente los arreglos asociados a cada uno de los usuarios.

5.4.3. Incorporación de puntos de interés

Después de realizar el proceso de incorporación de POIs, se definieron varios (nuevos) modelos de predicción. En la Tabla 16 se presenta el número de modelos definidos al considerar diferentes tamaños de celda y k=1 y k=3 . Por ejemplo, al considerar un tamaño de celda de 500 metros, de un total de 60 arreglos que cuentan con 3 arreglos similares, sólo en 12 ocasiones se agregaron POIs, y por consiguiente se definieron 12 nuevos modelos de predicción. Cuando el tamaño de la celda es de 250 metros, de un total de 63 arreglos que tienen un arreglo similar, sólo se definieron 9 nuevos modelos de predicción.

Tabla 16: Modelos de predicción definidos después de realizar el proceso de incorporación de POIs.

Tamaño de celda	k=1	k=3	
1000	49	14	
500	33	12	
250	9	0	

Con respecto a los puntos de interés que se incorporaron, en la Tabla 17 se presentan los resultados. Considerando un tamaño de celda de 1000 metros y k=1, en promedio se agregaron 1.62 POIs, en contraste, sólo 1.45 POIs se agregaron cuando k=3. Cuando el proceso de incorporación de POIs consideró un tamaño de celda de 500 metros, en promedio se agregaron 1.36 POIs al definir k=1, y 1.3 cuando k=3. Finalmente, para un tamaño de celda de 250 metros y k=1, en promedio se agregaron 1.3 POIs. Como se puede observar en la Tabla 17, no se presenta la cantidad de POIs que se agregaron al considerar los 5 arreglos más similares. Aunque al considerar los 5 arreglos más similares se identificaron POIs para agregarse, estos POIs correspondían a los que previamente se agregaron al considerar el arreglo y los 3 arreglos más similares. Por lo tanto, con estos resultados se puede comentar que sólo es necesario considerar el arreglo y los 3 arreglos

más similares para realizar el proceso de incorporación de puntos de interés.

Tabla 17: Cantidad de POIs agregados después de realizar con la comparación con los k arreglos más similares.

Tamaño de celda (m)	k=1	k=3
1000	1.62	1.45
500	1.36	1.30
250	1.30	0

5.4.4. Precisión de la predicción

En la Figura 46 se presenta la precisión promedio que se obtuvo durante las 4 semanas de pruebas al considerar los modelos de predicción con POIs de diferentes radios de clúster. Al considerar un periodo de predicción de 30 minutos y POIs con radio de 500 metros se obtiene una precisión promedio de 80 %; para un periodo de 60 minutos, se obtiene una precisión de 76 %. Al definir el periodo de predicción a 180 minutos, la precisión es de 71 %; para un periodo de 5 horas, se tiene una precisión de 70 %, y finalmente para un periodo de predicción de 7 horas, un 63 %.

Al considerar los POIs con radio de clúster de 250 metros, en promedio los modelos de predicción obtienen una precisión de 75, 73, 67, 61, y 59 % al considerar periodos de predicción de 30 minutos, 1, 3, 5 y 7 horas, respectivamente. Finalmente, al considerar los modelos de predicción con POIs de radio de 100 metros, se obtiene una precisión promedio de 72, 67, 66, 54, y 48 %, respectivamente. Cabe mencionar que estos resultados corresponden a la precisión que se obtuvo al utilizar los modelos de predicción base.

Debido a lo anterior, y al considerar que el objetivo de este experimento es conocer la factibilidad de utilizar la similitud de un usuario con otros usuarios, o bien consigo mismo para evitar la omisión de puntos de interés, ahora se presenta el aumento de la precisión

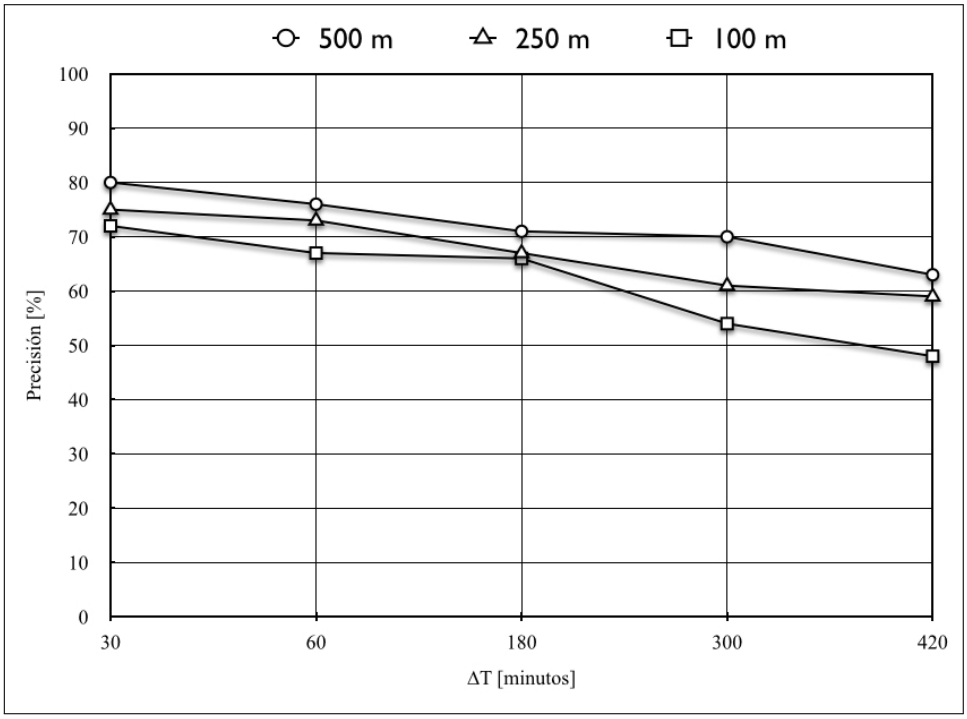


Figura 46: Precisión promedio que se obtiene con los modelos de predicción base durante las 4 semanas de prueba.

que se obtiene con los modelos de predicción que se definieron después de realizar el proceso de incorporación de POIs.

En la Tabla 18 se presenta el incremento promedio de la precisión. Los modelos de predicción que se definieron al comparar las celdas de 1000 metros y k=1, obtuvieron un incremento promedio de 8.14 % sobre los resultados de los modelos de predicción base, y un 7.97 % cuando k=3. De manera similar, cuando el tamaño de la celda fue de 500 metros, el incremento de la precisión fue de 7.84 % y 8.32 % para k=1 y k=3, respectivamente. Finalmente, los modelos de predicción que se definieron al considerar celdas de 250 metros y k=1, obtuvieron un incremento de 8.45 % sobre los resultados de los modelos de predicción base. Es importante mencionar que se obtuvo un incremento de la precisión de hasta 13 % sobre los modelos de predicción base.

A partir de los resultados anteriores, se observa la utilidad de la incorporación de nuevos POIs para así incrementar la precisión de las predicciones. Al considerar una mayor cantidad de POIs por usuario y día de la semana (Tabla 17) fue posible tener un mayor conocimiento acerca de los lugares en los que el usuario realiza sus actividades

cotidianas. Así, cada uno de los modelos de predicción refleja la movilidad del usuario: 1) considerando más POIs y 2) durante un periodo de tiempo mayor a lo largo del día.

Al considerar el incremento de la precisión de los diversos modelos de predicción (Tabla 18), no existe una diferencia significativa al considerar POIs con diferentes radios de clúster. Esto se debe principalmente a que la mayoría de los usuarios son estudiantes o académicos en la universidad de Beijing, por lo tanto, estos usuarios tienen en común el área que abarca la universidad. De esta manera, al realizar el proceso de incorporación de POIs, se agregaron lugares (en su mayoría) que se encuentran en el campus universitario. Así, el área geográfica de los POIs que se identificaron al utilizar celdas de 250 metros, posteriormente los POIs incorporados la incluyen al utilizar las celdas de 500 metros, y finalmente al considerar el proceso de incorporación con celdas de 1000 metros. De esta manera, se tiene un incremento similar en la precisión de la predicción.

Tabla 18: Aumento de la precisión al considerar el arreglo y los tres arreglos más similares.

Radio de clúster (m)	k=1	k=3
500	8.14	7.97
250	7.84	8.32
100	8.45	0

5.4.5. Similitud de los lugares

Después de conocer la similitud de los usuarios al considerar los lugares que éstos visitan de manera general, y/o en cada día de la semana, ahora resulta de interés conocer la similitud de los lugares. Como se había mencionado, el filtrado colaborativo utiliza dos enfoques para realizar predicciones: *basada en usuarios* y *basada en elementos*. En la sección anterior se utilizó el enfoque de *basado en usuarios* a fin de conocer la similitud de los usuarios considerando los lugares que éstos visitan, para posteriormente, basados en la similitud de un usuario dado A con otros k usuarios, determinar si es necesario agregar

otros puntos de interés al usuario A.

Considerando el enfoque *basado en elementos* (en este caso *basado en lugares*), resulta viable realizar recomendaciones de lugares que son similares a aquellos que han visitado cada uno de los usuarios. Además, resulta de interés conocer la similitud de estos lugares. En vista de que no se cuenta con información semántica para determinar la similitud de los lugares, si es posible conocer la similitud de éstos en función de los usuarios que los visitan.

Debido a lo anterior, se obtuvo la similitud de los lugares; para cada una de las matrices R, se obtuvo la similitud de cada columna con respecto al resto de las columnas. Puesto que la representación se realizó considerando celdas geográficas, al igual la similitud se encuentra en función de estas celdas. Primero, en la Tabla 19 se presenta la cantidad de celdas que tuvieron al menos una visita. Esto es, celdas en las cuales alguno de los usuarios tiene al menos un POI. Como se puede observar, cuando el tamaño de la celda se definió a 1000 metros, la cantidad de celdas visitadas fue de 351. En cambio, cuando el tamaño de la celda se fijó en 500 y 250 metros, los usuarios visitaron 606 y 884 celdas, respectivamente.

Posteriormente, en cada una de las matrices R, y para cada celda se identificaron las k celdas más similares. De igual manera, en la Tabla 19 se presenta la similitud promedio que se obtiene al considerar la celda más similar, así como las 3 y las 5 celdas más similares. Como se observa en la Tabla 19, no hay una gran diferencia de la similitud al incrementar el valor de k.

Tomando como referencia la predicción espacio - temporal y la similitud de los lugares, resulta factible realizar sugerencias o recomendaciones de aquellos lugares que son similares a los que el usuario va a visitar. Por ejemplo, al analizar los lugares que visitan los usuarios del proyecto Geolife, al tomar como referencia una cafetería, los lugares que se recomiendan son: una biblioteca de la Universidad de Beijing, una biblioteca pública, un salón de reuniones, y una sala de lectura (Figura 47). A fin de realizar una mejor recomendación, lo ideal sería contar con información de lugares específicos dentro de cada celda, así como contar con celdas de menor tamaño a fin de conocer la similitud de áreas

Tabla 19: Similitud de las celdas.

Tamaño de Celda	Cantidad de Celdas	k=1	k=3	k=5
1000	351	0.9512	0.9335	0.9165
500	606	0.9487	0.9291	0.9113
250	884			

geográficas de menor granularidad, y para que estas celdas no contemplen diversos lugares. Por ejemplo, una celda puede abarcar todos los establecimientos de un centro comercial.

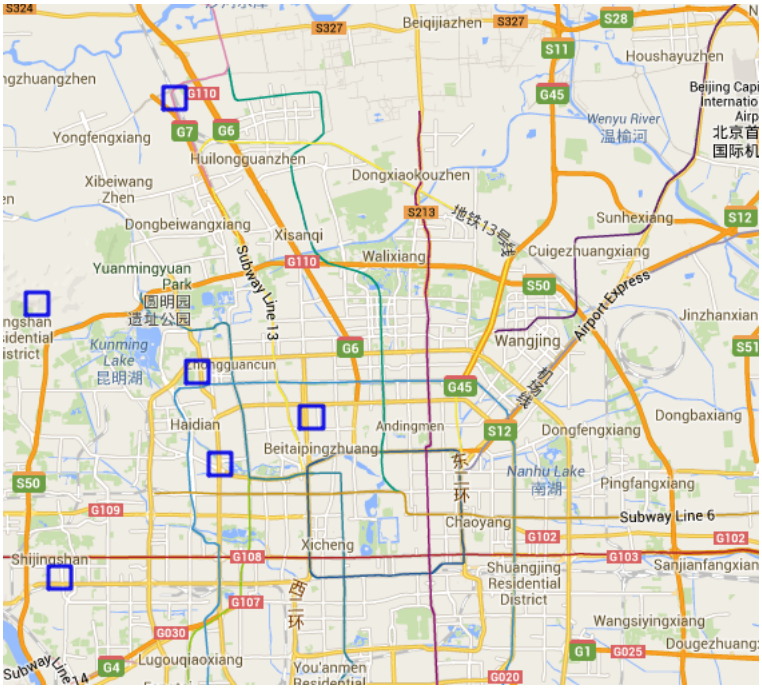


Figura 47: Tomando como referencia la celda en el extremo derecho, se presentan las celdas más similares a ésta.

5.4.5.1. Más allá de la predicción de la movilidad del usuario

Además de conocer la similitud entre usuarios y entre regiones geográficas (*celdas*), fue posible obtener información adicional que resulta de utilidad tanto para cuestiones

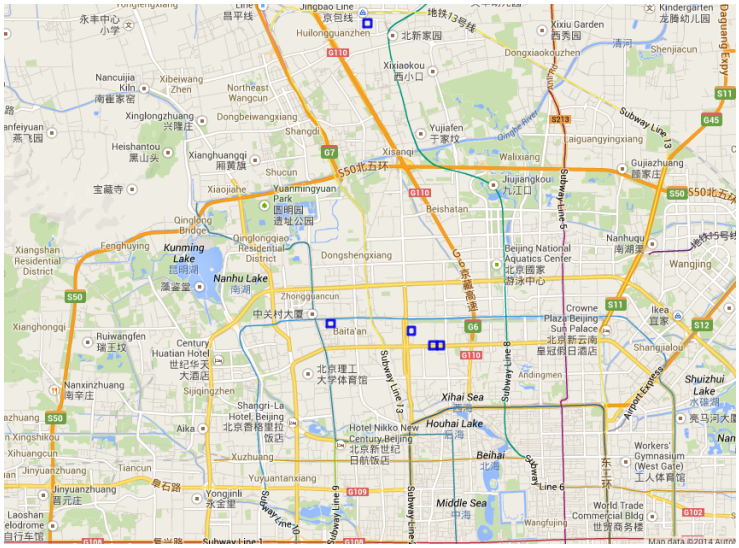
urbanas como para sistemas de recomendación. Por ejemplo, en la Figura 48 se presentan las celdas que tuvieron la mayor cantidad de visitas al considerar diferentes tamaños para las celdas. Estas visitas se realizaron a lo largo del día durante el periodo de tiempo que se consideró para el entrenamiento de los modelos de predicción. Para fines de demostración, sólo se presentan estas figuras, aunque se pueden obtener los lugares más visitados en diferentes periodos de tiempo, y así conocer el denominado pulso de la ciudad. Una vez que se identificaron las celdas que tuvieron la mayor cantidad de visitas, es posible obtener información semántica asociadas a estas. Por ejemplo, en la Tabla 20 se muestra la información de algunos lugares que se encuentran dentro de las 5 celdas más visitadas, al considerar diferentes tamaños para las celdas. De esta manera, es posible conocer el porqué dichas celdas son las más visitadas. En el caso de los lugares presentados en la Tabla 20, éstos corresponden principalmente a instituciones de educación, cultura, y recreación, con lo cual es posible conocer los gustos de la población.

De igual manera, al tener conocimiento de aquellas celdas con mayor número de visitas, puede ser de interés para cuestiones de planeación urbana, sistemas de transporte, y seguridad social. Así, es posible asignar un mayor número de elementos de seguridad al área que comprenden estas celdas, realizar la planeación de las rutas de transporte público para que consideren estas áreas geográficas, entre otros aspectos.

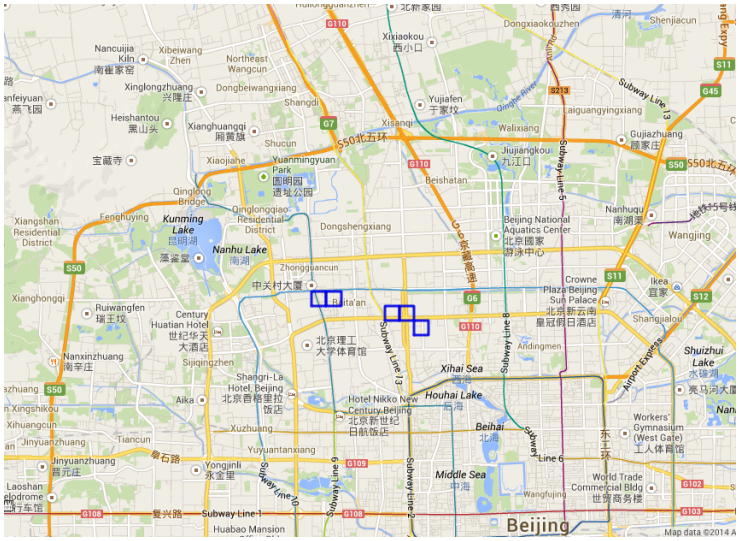
5.5. Resumen

En este capítulo se presentan los resultados de cada uno de los experimentos. Como se puede observar, la precisión del modelo de predicción propuesto supera a los resultados que se obtienen con el método $NP*$. De igual manera, con los resultados que se obtienen en los diferentes experimentos se demuestra la importancia de considerar aspectos tales como: predecibilidad de la movilidad del usuario, actualización de los puntos de interés, similitud de los usuarios, y el uso de la ventana deslizante para realizar el entrenamiento de los modelos de predicción.

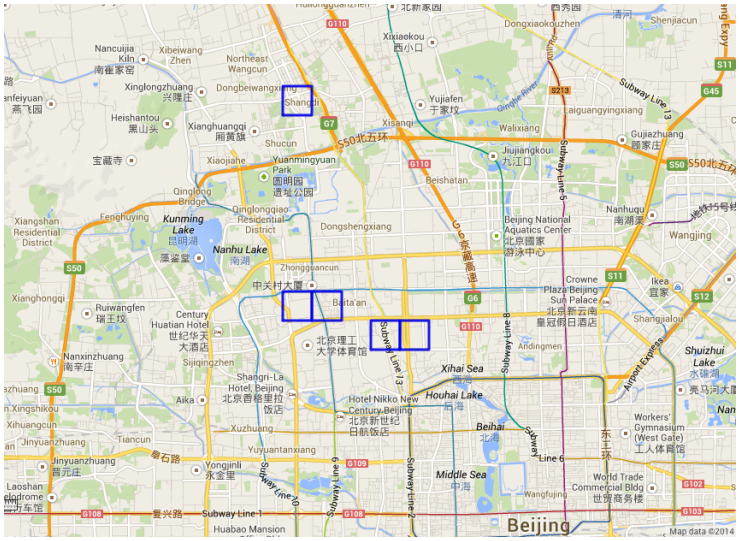
Al utilizar la prueba definida por Zhang *et al.* (2010), se encontró que en promedio la movilidad asociada a 4 días de la semana es predecible. Esto es, la movilidad del usuario entre sus puntos de interés se puede definir como una cadena de Markov, y así,



(a)



(b)



(c)

Figura 48: Celdas con mayor número de visitas; a) con celdas de 250 m; b) con celdas de 500 m; c) con celdas de 1000 m.

resulta viable utilizar los modelos ocultos de Markov para definir el modelo de predicción espacio-temporal. Posteriormente, con el uso de la ventana deslizante y el mecanismo de actualización de puntos de interés, se identificaron los cambios en la movilidad del usuario a lo largo del tiempo, y por consiguiente, se logró redefinir/actualizar el modelo de predicción de manera adecuada.

Con respecto a utilizar las preferencias colectivas para evitar la omisión de puntos de interés, se demostró que éste resulta un buen mecanismo para considerar lugares que una primera instancia no se habían considerado como puntos de interés. Y, al incorporar estos lugares, se tuvo un mayor entendimiento de la movilidad de los usuarios, y por lo tanto fue posible definir de mejor manera el modelo de predicción, y así se logró incrementar la precisión.

En el siguiente capítulo se presentan varios escenarios de aplicación en los cuales la predicción espacio-temporal es crucial a fin cumplir con el objetivo de cada uno de ellos. El dominio de aplicación de cada uno de los escenarios es diferente a fin de mostrar la funcionalidad del modelo de predicción en diversos entornos.

Tabla 20: Celdas pobladas.

Celdas 1000 metros	
Celda	Descripción
1	China University of Political Science and Law Science Hall
	Yuandadu Chengyuan Ruins Park
2	Central University of Finance and Economics
	Beijing Hadian Art Normal University
3	Beijing Zhongguancun High School
	Beijing Yijia Senior High School
4	Restaurants
	Hotels
5	China National Museum
	Xiao Football Stadium
Celdas 500 metros	
1	State Intellectual Property Office
	Yinhe Law Education Center
2	Zhichunli Primary School
	Beijing Zhongguanwn High School
3	National Office for Educational Science Planning
	Beijing Normal University Tengyinchnag
4	Peking University Students Gymnasium
	Capital University of Physical Education and Sports Library
5	Zhichum Park
	Beijing Zhichunli Middle School
Celdas 250 metros	
1	Beijing Film Academy Room
	Beijing Film Academy
2	Crowne Plaza Beijing Zhongguancun
	Zhichunli Community Service Station
3	Beijing Guodu Hospital
	Yuanwanglou Lily Food Garden
4	Jisheng Villa,North Gate
	Jisheng Villa
5	Hotels
	Restaurants

Capítulo 6. Aplicaciones de la predicción espacio temporal

Con el objetivo de mostrar la utilidad del modelo que permite predecir la próxima o próximas ubicaciones del usuario de manera precisa, a continuación se presentan algunos escenarios de aplicación. Estos escenarios de aplicación pertenecen a diversos dominios, de esta manera se muestra que el modelo propuesto no se encuentra restringido a un área específica. Además, los escenarios de aplicación representan situaciones cotidianas, las cuales se pueden estudiar al utilizar el modelo de predicción propuesto.

6.1. Evitar lugares congestionados

Considérese que para un día viernes, Juan y su esposa van a cenar a un famoso restaurante. Cuando la pareja llega al lugar, lamentablemente no hay mesas disponibles, y el tiempo de espera para obtener una mesa es de al menos una hora. Aunque la pareja decide esperar un poco, después de 25 minutos se encuentran molestos y deciden ir a otro lugar. Sin embargo, esta situación se pudo haber evitado al considerar la predicción de la movilidad de la población. De esta manera, cada usuario podría compartir la secuencia de lugares que visitará en un periodo de tiempo dado.

Así, es posible determinar la cantidad de personas que estarán en un lugar dado en un periodo de tiempo determinado. Como resultado, Juan y su esposa podrían haber consultado el restaurante deseado para conocer la cantidad aproximada de personas que habrá a la hora que desean ir a cenar, y considerando la capacidad del lugar, decidir si habrá lugares disponibles (Figuras 49 y 50). En caso contrario, la pareja podría buscar algún otro lugar para cenar.

Este escenario puede ser abordado de manera tradicional, por ejemplo al realizar una reservación de manera anticipada. Sin embargo, la solución que se presenta en este escenario va más allá. Al considerar los patrones de movilidad de un usuario, la inferencia de la movilidad y la creación de aplicaciones proactivas resultan aspectos viables.

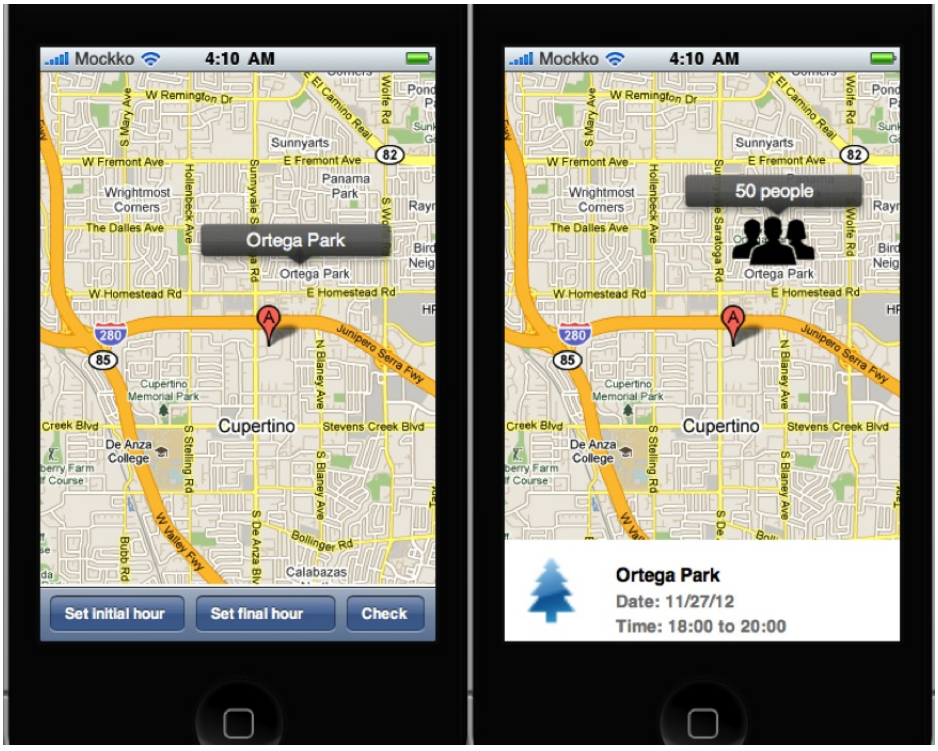


Figura 49: Predicción de la movilidad de la población: conociendo la cantidad de personas que habrá en un lugar determinado, en un día y periodo de tiempo dado.

6.2. Lugares concurridos de acuerdo a la hora y día de la semana

Similar al escenario anterior, otro escenario factible podría brindar una mayor cantidad de información al considerar no solo las predicciones asociadas a un determinado día de la semana, sino también al considerar datos históricos. Tomando como referencia la Figura 49, un usuario puede consultar la cantidad de personas que habrá en cada uno de los lugares de interés, o bien consultar un lugar en particular.

En cualquiera de los dos casos, además el usuario puede obtener información adicional, como: el día de la semana en el cuál el lugar es más visitado, o bien el periodo de tiempo con mayor cantidad de visitas.

Cabe mencionar que la inferencia de la cantidad de usuarios no considera aspectos más allá de las visitas de éstos a los POIs. Resulta factible mejorar la inferencia al considerar aspectos como la condición climatológica, congestionamiento vial, o imprevistos personales que pueden alterar que los usuarios no se encuentren en los tiempos estimados en los POI. Sin embargo, la integración de estos aspectos se encuentra fuera del

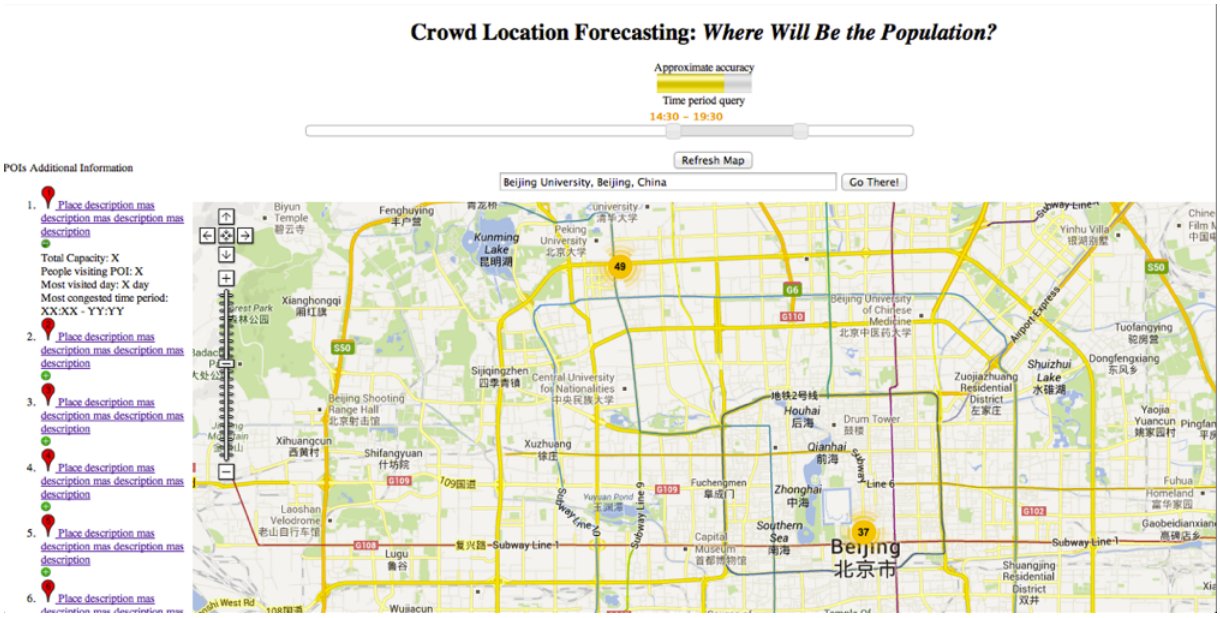


Figura 50: Predicción de la movilidad de la población: conociendo en qué lugares hay mayor concentración de personas de acuerdo al día y hora.

alcance de este trabajo.

6.3. Reserva de recursos

A la fecha, una de las aplicaciones de la predicción de la próxima ubicación del usuario concierne a la reservación de recursos (Nicholson y Noble (2008)); por ejemplo, para reservar ancho de banda en un punto de acceso. Aunque la información de la próxima ubicación de los usuarios es importante para realizar una reservación adecuada, una mejor reservación puede realizarse al considerar el aspecto temporal; esto es, considerar la hora de arribo al próximo o próximos lugares. Así, de manera anticipada a la llegada del usuario al punto de acceso, los recursos se asignan a éste para su uso posterior.

6.4. Comunicación de datos entre regiones desconectadas

Una variante de las redes DTN (redes tolerantes a fallas y retrasos), son las redes oportunistas (OppNet). En las redes OppNet, a diferencia de las redes tradicionales, no existe una ruta predefinida para enviar datos de un nodo remitente a un nodo destino. Así también, en las redes OppNet, los nodos no tienen conocimiento global de la topología de la red. Por lo tanto, en la redes oportunistas se toma ventaja de los encuentros ocasionales de los nodos y de los dispositivos móviles que éstos portan para realizar la

comunicación de datos.

Debido a que no se tiene conocimiento global, las redes oportunistas presentan algunas implicaciones, entre las cuales destaca la baja probabilidad de entrega de los datos, y el aumento del tiempo de entrega de los mismos. Debido a ello, uno de los retos y aspectos interesantes de las redes oportunistas es maximizar la probabilidad de entrega de los datos, y minimizar los tiempos de entrega. Es aquí donde la predicción de la movilidad de los usuarios cobra relevancia. Para ello, considérese el siguiente escenario.

José desea enviar un mensaje a su novia Raquel, pero José vive en una comunidad rural que no cuenta con servicios de telecomunicación, y su novia vive en una ciudad que se encuentra a 1:30 horas de viaje (Figura 51). Afortunadamente, algunos estudiantes como Miguel, Carlos, y Elisa viajan diariamente a la ciudad donde vive Raquel. Por lo tanto, José puede tomar ventaja de la movilidad de estos estudiantes para enviar su mensaje. Sin embargo, ninguno de ellos tiene conocimiento del domicilio de Raquel, por lo que no pueden comunicar el mensaje. Para evitar este problema, José tiene una mejor idea. Él tiene conocimiento de que cuando los estudiantes arriban a la ciudad tienen conexión a Internet en diversos lugares, por lo tanto, José agrega la dirección virtual de Raquel (e.g., correo electrónico). De esta manera, no es necesario entregar el mensaje personalmente, basta con utilizar alguno de los servicios en Internet para comunicar el mensaje.

Al tomar como referencia lo anterior, y considerando los registros históricos de las visitas de estos usuarios a lugares con conectividad a Internet, se está en posición de estimar que Miguel tendrá acceso a Internet en la preparatoria dentro de 5 horas; Carlos tendrá conexión en un café Internet en 7 horas, y Elisa se conectará a Internet en un parque dentro de 2 horas. De esta manera, José puede seleccionar al mejor mensajero de acuerdo a la hora de su próxima conexión. Así, José transmite el mensaje a Elisa, quien a su vez comunicará el mensaje tan pronto tenga acceso a Internet.

6.5. La predicción espacio-temporal y los ambientes inteligentes

Tener conocimiento de los lugares que visitará un usuario, y los tiempos de arribo a éstos, permite el desarrollo de aplicaciones en apoyo a los ambientes inteligentes. Este

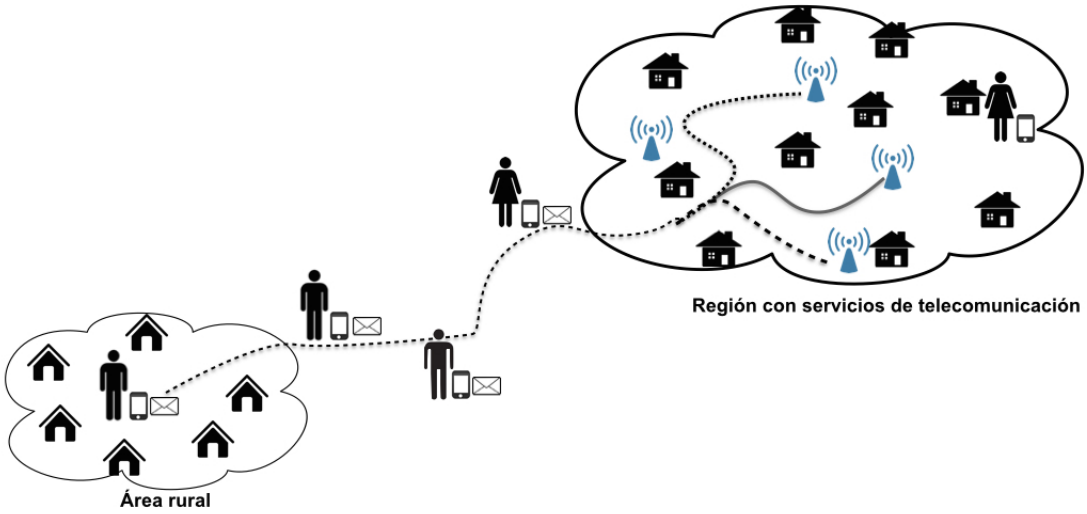


Figura 51: Comunicación de datos entre regiones desconectadas utilizando los contactos oportunistas y la predicción de la movilidad de los usuarios.

conocimiento se puede utilizar para realizar ciertas acciones como encender la calefacción/aire acondicionado previo a la llegada del usuario al hogar u oficina. Ellis *et al.* (2012), Scott *et al.* (2011) y Krumm y Brush (2011) han presentado enfoques similares.

6.6. Discusión

Cabe mencionar que la funcionalidad u objetivo de los escenarios de aplicación presentados no dependen de una predicción 100 % precisa. Estos escenarios pueden tolerar un error en la precisión. Por ejemplo, en el caso del escenario de los lugares congestionados, el costo del error en la precisión representa una demora para acceder a un mesa en el restaurante. En el caso del escenario en el cual se identifican los lugares concurridos, el costo del error se refleja en la cantidad de personas que según la predicción estarán en un lugar y hora determinada. En el escenario de reserva de recursos, el error de la precisión propicia que los recursos no pueden ser utilizados por terceros, ya que éstos se encuentran en reserva.

Con respecto a la comunicación de datos entre regiones desconectadas, el costo del error trae consigo un retraso en la entrega de los datos, lo cual es un aspecto común en las redes oportunistas. Finalmente, en el escenario de ambientes inteligentes una predicción imprecisa tiene como consecuencia el gasto de recursos energéticos (i.e., energía eléctrica). Aunque una predicción imprecisa conlleva un error en la funcionalidad de los

escenarios, este error no es crucial para delimitar la ejecución de los escenarios.

6.7. Resumen

En este capítulo se presentan algunas aplicaciones cuya implementación resulta factible al considerar la predicción espacio-temporal de la movilidad del usuario. Los dominios de aplicación de la predicción no se encuentran restringidos a los que aquí se presentan, de manera general las aplicaciones basadas en localización pueden beneficiarse de este aspecto. De esta manera, aplicaciones como Google Now, Osito, y otras, pueden ofrecer información acorde a las próximas visitas del usuario basándose en la predicción de la movilidad, y no solamente en la información que el usuario provee de manera explícita. La funcionalidad y aplicación del enfoque propuesto no se limita al modelo de predicción.

La movilidad del usuario entre los puntos de interés es de beneficio para aplicaciones urbanas y/o planeación de rutas, así como para conocer el flujo de personas entre ciertos lugares. De igual manera, el hecho de identificar los puntos de interés de cada uno de los usuarios, permite tener conocimiento de los lugares que son significativos para éstos, y posteriormente inferir gustos, preferencias y/o hábitos a partir de los lugares que éstos visitan.

Capítulo 7. Conclusiones y trabajo futuro

7.1. Conclusiones

En este capítulo se presentan las conclusiones del trabajo de investigación, y adicionalmente se presentan las líneas de investigación relacionadas al trabajo realizado, las cuales se pueden explorar a futuro.

La capacidad de predecir la ubicación del usuario de manera precisa es de interés en diferentes campos de aplicación como planeación urbana, salud, sistemas de cómputo ubicuo, redes de computadoras, sistemas de recomendación, éstas sólo por mencionar algunas. Además, con la reciente proliferación de aplicaciones y sistemas proactivos basados en localización, el beneficio de tener conocimiento de la movilidad futura del usuario se incrementa.

Por lo tanto, en este trabajo de investigación se presenta un modelo de predicción que permite predecir los lugares en los cuales estará el usuario en un periodo de tiempo dado, y así también permite conocer el tiempo en que el usuario estará en dichos lugares. Esto es, se definió un modelo de predicción espacio-temporal de la movilidad del usuario. El modelo de predicción propuesto se basa en los aspectos espaciales y temporales de la movilidad del usuario. Después de realizar un análisis de la movilidad de los usuarios, se identificaron aquellos factores que son importantes e imprescindibles para caracterizar la movilidad, y así definir el modelo de predicción.

Debido a lo anterior, se encontró que la movilidad de los usuarios es diferente para cada uno de los días de la semana, por lo que definió un modelo de predicción para cada uno de estos días. Posteriormente, al considerar la movilidad del usuario en cada día de la semana, se encontró que dentro de las actividades cotidianas existen lugares que son significativos o importantes para el usuario. Estos lugares se caracterizan porque el usuario pasa un determinado tiempo en ellos, y además los visita con cierta regularidad.

También, se encontró que la movilidad del usuario entre estos lugares puede se puede modelar como una cadena de Markov. Esto es, la ubicación del usuario en un lugar significativo determina la próxima ubicación de éste. Así, también, al realizar el análisis

de la movilidad, se identificó la relación que existe entre los tiempos de arribo y partida a los lugares significativos y la hora del día; considerando la propiedad Markoviana y la relación con la hora del día, fue posible definir la movilidad del usuario como un modelo oculto de Markov. De esta manera, los aspectos de la movilidad que se consideraron son: el día de la semana, lugares significativos o puntos de interés, hora de arribo y partida a los puntos de interés.

A diferencia de los trabajos relacionados, el modelo de predicción propuesto permite realizar la predicción de la movilidad en el mediano plazo. Esto es, predice la movilidad del usuario de manera precisa considerando un periodo de predicción de hasta 8 horas. Los resultados del modelo de predicción propuesto superan a los resultados que se presentan en los trabajos del estado del arte en cuanto a la precisión se refiere.

Después de la investigación que se ha realizado en este trabajo, y de los resultados que se obtuvieron en los diferentes experimentos, se identificaron varios aspectos que son importantes para asegurar el rendimiento adecuado del modelo de predicción. Cabe mencionar que estos aspectos no se han tratado en su totalidad en los trabajos previos. Por lo tanto, al término del presente trabajo se tienen varios comentarios y conclusiones.

- El modelo de predicción propuesto se basa en aspectos temporales y espaciales de la movilidad, éstos no son exclusivos del modelo de predicción propuesto. En varios trabajos relacionados (McInerney *et al.* (2013), Baumann *et al.* (2013), Do y Gatica-Pérez (2012)) se consideran algunos o la totalidad de los aspectos que se consideran en este trabajo, en conjunto con otros factores. Sin embargo, la diferencia con éstos y otros trabajos reside en cómo se realiza el modelado de la movilidad, el cual se encuentra en función del objetivo y dominio de aplicación del trabajo de investigación. Aunado a ello, el tipo de datos que se tienen disponibles, la resolución, y la cantidad de éstos determinan en gran medida el modelado y la definición de los modelos de predicción. Por consiguiente, se argumenta que más allá de las características que se consideren de la movilidad, existen otros factores que determinan la efectividad del modelo de predicción.

- De manera general, en los trabajos previos, el entrenamiento del modelo de predic-

ción se realiza tomando en cuenta una porción determinada del conjunto de datos que se tiene disponible (e.g., la mitad, un cuarto); sin embargo, esta acción no resulta adecuada cuando se desea predecir la movilidad más reciente del usuario de manera efectiva en los aspectos espacial y temporal. Debido a los múltiples patrones de movilidad del usuario a lo largo del tiempo, al considerar una cantidad arbitraria de datos, éstos pueden abarcar diversos patrones de movilidad. Por lo tanto, el modelo de predicción no reflejaría el comportamiento actual del usuario (e.g., lugares significativos, tiempos de estadía, tiempos de arribo, etc.). Debido a ello, una contribución del presente trabajo es la incorporación de un mecanismo que permite identificar el periodo de tiempo en el cual la movilidad del usuario es similar. Este mecanismo da la pauta para modelar de manera precisa el modelo de predicción con los datos correspondientes al patrón de movilidad actual del usuario.

- Otro factor relevante al definir el modelo de predicción, es la actualización de éste. Debido a los diversos patrones de comportamiento del usuario, la movilidad de éste cambia a lo largo del tiempo, ya sea a través de cambios radicales o bien a través de cambios graduales. A fin de que un modelo de predicción resulte de utilidad a lo largo del tiempo, éste debe de incorporar los cambios en el patrón de movilidad. Debido a ello, una contribución del presente trabajo reside en definir e incorporar un mecanismo que permite identificar e incorporar los cambios en la movilidad del usuario a lo largo del tiempo.

- De manera realista, resulta complicado recolectar continuamente la ubicación del usuario por un periodo de tiempo prolongado, ya sea por cuestiones de recursos de los dispositivos móviles, o bien por cuestiones de privacidad por parte del usuario, la colecta de los datos no se realiza de manera continua. Por lo tanto, el modelo de predicción debe de compensar la escasez de datos de localización con otro tipo de datos con el fin de contar con una mayor cantidad de datos y así identificar el patrón de movilidad del usuario, y por consiguiente modelar la movilidad de manera precisa. Con respecto a este punto, otra contribución de este trabajo reside en definir un mecanismo que toma como referencia la similitud entre un conjunto dado de usuarios para evitar la omisión de puntos de interés para un usuario en particular. Así, el modelo de predicción se define de manera precisa, y por ende se maximiza

la precisión de la predicción.

- A pesar de que la movilidad del usuario es diferente para cada día de la semana, los datos de la movilidad del usuario asociados a un día determinado se pueden utilizar para compensar la falta de datos de la movilidad de otro día de la semana. Como se presentó en el capítulo 5, existe similitud entre los lugares que el usuario visita en cada uno de los días de la semana. Por lo tanto, la contribución que se menciona en el punto anterior, también aplica en esta situación. El mecanismo propuesto obtiene la similitud de la movilidad del usuario en los diferentes días de la semana para que en el caso que no se cuente con datos suficientes en un día determinado, estos datos se pueden extrapolar al considerar los datos de movilidad de aquellos días que son similares.

- Otro de los factores a resaltar es la identificación de los puntos de interés o lugares significativos. Ya que los usuarios pasan la mayor parte del tiempo en lugares que son de interés para ellos, identificar correctamente estos lugares permite conocer a mayor detalle la movilidad del usuario. Además, la identificación de estos lugares da la pauta para el entrenamiento adecuado del modelo de predicción, y por consiguiente una predicción acertada de la movilidad del usuario. Para el enfoque propuesto, los puntos de interés que se consideran son aquellos que se encuentran incluidos en el patrón de movilidad más reciente del usuario. Por lo tanto, otra contribución de este trabajo se encuentra en definir el conjunto de parámetros que permiten identificar, en conjuntos con los algoritmos (para interiores y exteriores), los puntos de interés que son significativos para el usuario en un tiempo determinado. Sin embargo, un punto a considerar es que la identificación de los puntos de interés se encuentra definida por el objetivo del trabajo en cuestión. Algunos trabajos consideran puntos de interés sólo a aquellos lugares que son relevantes durante un periodo de tiempo prolongado. Por lo tanto, no se puede argumentar que los algoritmos para identificar puntos de interés (y los parámetros) utilizados en este trabajo son mejores que los presentados en trabajos relacionados.

De manera específica, también se tienen ciertos comentarios con respecto a los datos de localización:

- En este trabajo se utilizaron datos de localización que corresponden a conexiones a puntos de acceso 802.11, y trayectorias GPS. Utilizar datos de conexiones a puntos de acceso tiene varias ventajas. Actualmente una gran cantidad de dispositivos cuentan con interfaz de comunicación inalámbrica (802.11) y al igual una gran cantidad de lugares cuentan con esta infraestructura (e.g., lugares públicos, hogares), por lo que definir un modelo de predicción basado en datos de Wi-Fi resulta factible. La granularidad de los puntos de acceso representa otra ventaja, al conocer el punto de acceso en el que se encuentra conectado un usuario, se puede determinar en qué sección de un edificio, casa o lugar se encuentra el usuario. Sin embargo, el uso de datos de conexiones a puntos de acceso también tiene ciertas desventajas. A pesar de los puntos de acceso disponibles en el entorno cotidiano, los usuarios no tienen acceso acceso a todos ellos, por lo que sólo es posible recolectar datos de conexiones a puntos de acceso en una cantidad limitada de lugares, y por consiguiente sólo se conoce la movilidad del usuario de manera parcial. Lo anterior conlleva a que el modelo de predicción sólo considere una cantidad limitada de lugares.

- A fin de conocer la movilidad del usuario de manera continua, en este trabajo también se utilizaron datos de GPS. Sin embargo, mediante la colecta de datos con la tecnología GPS la batería de los dispositivos móviles se agota en cuestión de algunas horas. Este aspecto se identificó al utilizar las trayectorias del proyecto Geolife, en donde el mayor periodo de tiempo de colecta de datos fue de 12 horas; la mayoría de las trayectorias contemplan periodos cortos de tiempo. Aunque en algunos casos, el periodo de tiempo entre el fin de una trayectoria y el inicio de la siguiente fue mínimo (e.g., cuestión de minutos), de manera general dichos periodos fueron prolongados (e.g., días). Lo anterior conlleva a la problemática que se tiene con los datos de conexión a puntos de acceso. Aunado a ello, la falta de funcionalidad del GPS en interiores evita la identificación de lugares significativos, y de igual manera al considerar el margen de error del GPS, la labor de identificar los lugares significativos y la información asociada, se complica.

- Con la proliferación de aplicaciones basados en localización, resulta viable recolectar datos de la ubicación del usuario a partir de los registros de presencia, actividad-

des georeferenciadas, entre otros, los cuales pueden compensar la falta de datos. De esta manera con la combinación de datos continuos y discretos de la ubicación del usuario es posible tener una visión más completa de la movilidad de éste, y así realizar el modelado y entrenamiento de los modelos de predicción de una mejor manera.

- La falta de datos de localización, o la escasez de dichos datos es motivo de múltiples discusiones entre los miembros de la academia. Al momento, resulta inadecuada una comparación equitativa entre los diversos modelos de predicción, debido principalmente a los datos que utilizan para realizar las pruebas. Algunos de los conjuntos de datos no se encuentran disponibles de manera pública, por lo que resulta difícil corroborar/comprobar los resultados a partir de estos conjuntos. Los conjuntos que se encuentran disponibles de manera pública, cuentan con ciertos problemas; datos escasos, periodos de sensado limitados, etcétera. Aunque en años recientes, grandes organizaciones han realizado campañas a fin de obtener datos de contexto del usuario, los datos de localización recolectados resultan insuficientes. Una opción alterna reside en recolectar los datos de localización por cuenta propia, lo que conlleva una serie de retos en cuestión de infraestructura, participación ciudadana, privacidad, seguridad de los datos, entre otros.

7.2. Limitaciones del enfoque propuesto

El modelo de predicción propuesto obtiene una precisión mayor a la que que presenta en el estado del arte, sin embargo este modelo no resulta adecuado para todos los usuarios. Se sabe que hay usuarios o grupos de usuarios que no siguen un patrón de movilidad debido a las actividades que éstos desempeñan. Por ejemplo, los políticos, empresarios, entre otros.

En el caso de aquellos usuarios que siguen un patrón de movilidad, no en todos los casos ésta se puede modelar como una cadena de Markov, por lo que el modelo de predicción propuesto resulta inadecuado. Para estos usuarios se requiere contar con otro modelo para así predecir la movilidad de manera adecuada.

Aún en el caso en que la movilidad de los usuarios se puede modelar como una cadena de Markov, hay situaciones en que los usuarios se desvían de sus patrones habituales. Estas situaciones se deben identificar con el fin de que los modelos de predicción no incluyan los datos asociados a éstas, y así evitar definir un modelo de predicción impreciso. O en caso contrario, para este tipo de ocasiones definir otro modelo de predicción que sólo considere los comportamientos que prevalecen durante largos periodos de tiempo como lo hace Eagle y Pentland (2006).

El enfoque propuesto parte de la premisa de contar con predicciones individuales para posteriormente predecir la movilidad de la población y hacer posible el conjunto de aplicaciones presentadas en el capítulo 6. A nivel teórico, esta idea resulta factible; sin embargo, a nivel práctico es necesario considerar varios aspectos. Con el fin de preservar la privacidad y anonimato de los usuarios involucrados se requiere contar con mecanismos de seguridad que permitan la transmisión segura y el buen uso de los datos asociados a las predicciones.

7.3. Publicaciones

Como resultado de este trabajo se publicaron 3 artículos, 2 de ellos en workshop internacionales y uno en revista.

- Jorge Álvarez-Lozano, J. Antonio García-Macías, and Edgar Chávez. 2012. User location forecasting at points of interest. In Proceedings of the 2012 RecSys workshop on Personalizing the local mobile experience (LocalPeMA '12). ACM, New York, NY, USA, 7-12. DOI=10.1145/2365946.2365949
- Jorge Álvarez-Lozano, J. Antonio García-Macías, and Edgar Chávez. 2013. Learning and user adaptation in location forecasting. In Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication (UbiComp '13 Adjunct). ACM, New York, NY, USA, 461-470. DOI=10.1145/2494091.2495978
- Jorge Álvarez-Lozano, J. Antonio García-Macías, and Edgar Chávez. 2013. Crowd Location Forecasting at Points of Interest, International Journal of Ad Hoc and Ubiquitous Computing (Aceptado).

7.4. Trabajo futuro

Con respecto al trabajo futuro, a continuación se describen varios problemas que se pueden estudiar.

7.4.1. Patrones de movilidad

En el desarrollo del presente trabajo y en trabajos relacionados se ha discutido acerca de la dinamicidad de la movilidad a lo largo del tiempo. Aunque se definieron mecanismos para considerar los cambios en el comportamiento de la movilidad, estos cambios no son identificados de manera pronta.

Por lo tanto, como trabajo futuro en una primera instancia se plantea la identificación adecuada de estos patrones de movilidad y el periodo de tiempo que contempla cada uno de ellos. Posteriormente, definir un modelo de predicción base para cada patrón de movilidad. De esta manera, se utiliza el modelo de predicción que sea acorde a la fecha o temporada en curso.

7.4.2. Predicción de la movilidad

7.4.2.1. Modelado de la movilidad a través de datos heterogéneos

En la actualidad, la mayoría de los trabajos que realiza análisis de datos para entender la movilidad de los usuarios y realizar la predicción de ésta, hacen uso de conjuntos de datos especializados. Especializados en el sentido de que los usuarios involucrados en los proyectos correspondientes, son conscientes de que comparten los datos. De esta manera, se tiene una gran cantidad de datos disponibles para analizar.

Sin embargo, de manera realista no es factible obtener una gran cantidad de datos del usuario, ya que debido a las limitaciones de recursos, no es posible recolectar registros de localización mediante GPS, Wi-Fi, o bluetooth durante un periodo de varias horas. Así, también, debido a las restricciones de privacidad no todos los usuarios permiten el acceso a los datos de localización.

Tomando como referencia la importancia de los dispositivos móviles en las actividades

cotidianas de los usuarios, aunado a la proliferación de los servicios basados en localización, es posible obtener datos de localización de diversas fuentes. A través de registros de presencia mediante las redes sociales, registros de conexiones a puntos de acceso, y registros esporádicos de GPS, es posible crear un panorama de la movilidad del usuario. Debido a ello, como trabajo futuro se plantea el uso de datos dispersos y heterogéneos para realizar el análisis de la movilidad del usuario.

7.4.2.2. Predicción semántica

Con respecto a la predicción de la movilidad, este trabajo sólo se enfoca en predecir el próximo lugar a visitar o bien la secuencia de lugares a visitar considerando para ello sólo el aspecto espacial. Sin embargo, un aspecto que resulta de interés para aplicaciones de cómputo ubicuo es que además de conocer la siguiente ubicación geográfica a visitarse, es importante conocer la semántica de dicho lugar.

Realizar la predicción a nivel semántico involucra ciertos retos. El reto más importante reside en obtener la semántica de los lugares que visita el usuario a lo largo del tiempo. A la fecha algunos trabajos han abordado la predicción a nivel semántico, como es el caso del trabajo presentado por Ying *et al.* (2011); sin embargo, estos autores no tratan el problema de obtener la semántica ya que cuentan con un repositorio que mapea cada ubicación geográfica con su respectivo significativo semántico.

Obtener el significado semántico de los lugares no es un reto trivial. Aunque en la actualidad existen diversas herramientas para realizar el mapeo [1], la efectividad de éstas depende de la ubicación exacta del lugar a consultar. Sin embargo, esto no es posible con el enfoque actual, ya que debido al proceso de identificación de los puntos de interés, estos lugares abarcan un área geográfica de tamaño variante en lugar de una ubicación en particular. Debido a ello, dentro del área geográfica se encuentran diversos lugares.

Por lo tanto, no se cuentan con un tupla de coordenadas geográficas, sino con un conjunto de coordenadas que corresponden al área del punto de interés, por lo que al utilizar las herramientas disponibles el resultado sería impreciso.

[1] https://developers.google.com/maps/documentation/geocoding/

Debido a lo anterior, el reto se encuentra en identificar la ubicación exacta que es de interés para el usuario para posteriormente realizar el mapeo. Para ello, resulta viable utilizar la información que brindan los servicios basados en localización o bien el dispositivo móvil del usuario. Utilizando los registros de presencia del usuario (*check-in*) es posible determinar con precisión donde se encuentra el usuario.

Otro enfoque factible para obtener el significado de lugares geográficos es considerar información adicional como lo hace Krumm y Rouhana (2013). Ellos utilizan información acerca del tipo de lugar (e.g., restaurante, supermercado) a fin de estimar el significado semántico del lugar de interés.

Luego de conocer el significado semántico de cada punto de interés, la movilidad se puede modelar nuevamente con un HMM.

7.4.3. Cómputo urbano

La identificación precisa de los puntos de interés tiene un rol fundamental en la predicción de la movilidad del usuario. Una vez identificados estos lugares, no sólo son de utilidad en la predicción espacio-temporal, sino también en otros dominios de aplicación.

Durante el desarrollo de este trabajo se han identificado varios nichos de oportunidad en distintas áreas donde resulta de útil aplicar el conocimiento adquirido, siendo una de ellas el cómputo urbano. Un aspecto importante en la planeación urbana y/o en el sistema de transporte de una ciudad es conocer las áreas en donde se concentra una mayor cantidad de personas.

De esta manera, los urbanistas o planeadores urbanos pueden tomar ciertas decisiones entre las que se encuentran definir rutas de transporte que cubran total o parcialmente estos lugares. Así, también, es posible tomar decisiones en cuanto a mejorar las vías que conectan los diferentes puntos, asignar una mayor cantidad de recursos (e.g., policía), entre otros.

La idea reside en identificar aquellos lugares que son de interés para toda la población, en lugar de identificar los lugares que son significativos para una persona en particular y posteriormente juntar estos lugares como se realizó en este trabajo. Una vez identificados

estos lugares, es necesario tener una visión global de la movilidad de la población en dichos lugares para posteriormente definir un patrón que permita entender la movilidad a nivel urbano.

Por ejemplo, considerando la Figura 52(a) se conocen aquellos lugares en los cuales hay una gran cantidad de personas (Beijing, China) en cada uno de los días de la semana durante un periodo de tiempo dado (e.g., una semana, mes). O bien, la Figura 52(d) presenta los lugares que tienen una mayor afluencia de personas en una hora determinada del día durante un periodo de tiempo dado. Después de tener conocimiento de estos lugares, resulta de interés el conocer la movilidad de los usuarios entre estos lugares.

Por tanto, la Figura 52(b) presenta de manera gráfica la movilidad entre estos lugares de acuerdo al día de la semana, y al igual la Figura 52(d) presenta movilidad de los usuarios considerando la hora del día.

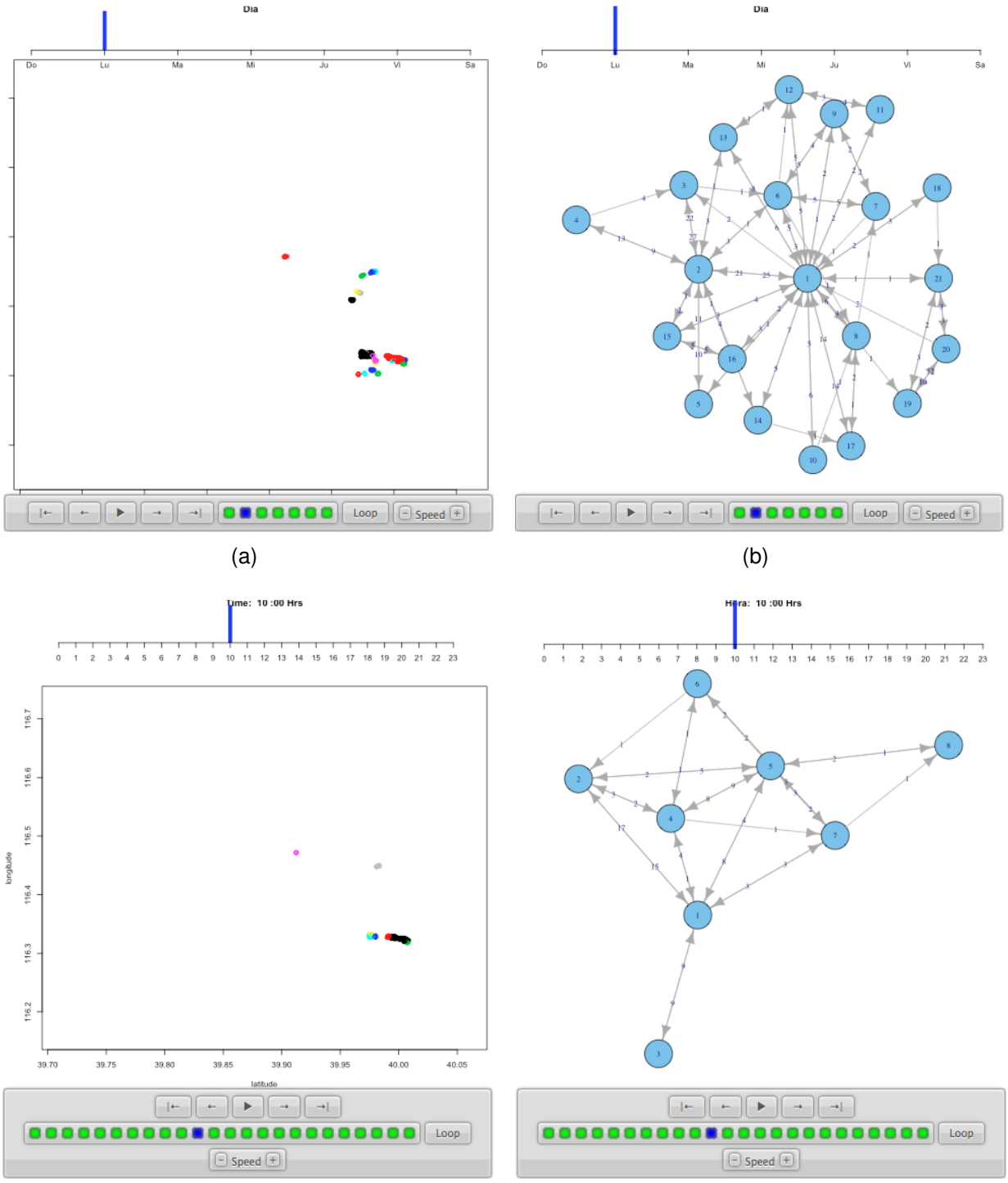


Figura 52: Lugares que son de interés para la población, y la movilidad de los usuarios entre estos lugares; a) lugares con una mayor afluencia de personas de acuerdo al día de la semana; b) movilidad de los usuarios entre los lugares más visitados de acuerdo al día de la semana; c) lugares con una mayor afluencia de personas de acuerdo a la hora del día; d) movilidad de los usuarios entre los lugares más visitados de acuerdo a la hora del día.

Lista de referencias

Aalto, L., Göthlin, N., Korhonen, J., y Ojala, T. (2004). Bluetooth and wap push based location-aware mobile advertising system. En: *Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services*, Boston, MA, USA. ACM, MobiSys '04, pp. 49–58.

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., y Steggles, P. (1999). Towards a better understanding of context and context-awareness. En: *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, London, UK, UK. Springer-Verlag, HUC '99, pp. 304–307.

Adomavicius, G. y Tuzhilin, A. (2005a). Toward the next generation of recommender systems: A survey of the state of the art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, **17**(6): 734–749.

Adomavicius, G. y Tuzhilin, A. (2005b). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, **17**(6): 734–749.

Aleman, D. M., Wibisono, T. G., y Schwartz, B. (2011). A nonhomogeneous agent-based simulation approach to modeling the spread of disease in a pandemic outbreak. *Interfaces*, **41**(3): 301–315.

Ashbrook, D. (2002). Learning significant locations and predicting user movement with gps. En: *Proceedings of the 6th IEEE International Symposium on Wearable Computers*, Washington, DC, USA. IEEE Computer Society, ISWC '02, pp. 101–108.

Ashbrook, D. y Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Computing*, **7**(5): 275–286.

Backstrom, L., Sun, E., y Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. En: *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, North Carolina, USA. ACM, WWW '10, pp. 61–70.

Bardram, J. E., Frost, M., Szántó, K., Faurholt-Jepsen, M., Vinberg, M., y Kessing, L. V. (2013). Designing mobile health technology for bipolar disorder: A field trial of the monarca system. En: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. ACM, CHI '13, pp. 2627–2636.

Baumann, P., Kleiminger, W., y Santini, S. (2013). The influence of temporal and spatial features on the performance of next-place prediction algorithms. En: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA. ACM, UbiComp '13, pp. 449–458.

Begole, G. (2010). It's time to reap the context-aware harvest. Recuperado el 22 de Junio de 2014 de: http://blogs.parc.com/blog/2010/09/its-time-to-reap-the-context-aware-harvest/.

Bellotti, V., Begole, B., Chi, E. H., Ducheneaut, N., Fang, J., Isaacs, E., King, T., Newman, M. W., Partridge, K., Price, B., Rasmussen, P., Roberts, M., Schiano, D. J., y Walendowski, A. (2008). Activity-based serendipitous recommendations with the magitti mobile leisure guide. En: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Florence, Italy. ACM, CHI '08, pp. 1157–1166.

Bhattacharya, A. y Das, S. K. (2002). Lezi-update: An information-theoretic framework for personal mobility tracking in pcs networks. *Wirel. Netw.*, **8**(2/3): 121–135.

Breese, J. S., Heckerman, D., y Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. En: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., UAI'98, pp. 43–52.

Bruner, G. C. y Kumar, A. (2007). Attitude towards location based advertising. *Journal of Interactive Advertisement*, **7**(2).

Burbey, I. (2011). Predicting future locations and arrival times of individuals. Tesis de Doctorado. Virginia Polytechnic Institute and State University. 216 p.

Burbey, I. y Martin, T. (2012a). When will you be at the office? predicting future locations and times. En: M. Gris y G. Yang (eds.), *Mobile Computing, Applications, and Services*, Vol. 76 de *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Springer Berlin Heidelberg, pp. 156–175.

Burbey, I. y Martin, T. L. (2008). Predicting future locations using prediction-by-partial-match. En: *Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments*, New York, NY, USA. ACM, MELT '08, pp. 1–6.

Burbey, I. y Martin, T. L. (2012b). A survey on predicting personal mobility. *Int. J. Pervasive Computing and Communications*, **8**(1): 5–22.

Cai, H. y Eun, D. Y. (2007). Crossing over the bounded domain: from exponential to power-law inter-meeting time in manet. En: *Proceedings of the 13th annual ACM International Conference on Mobile Computing and Networking (MobiCom'07)*. ACM Press, pp. 159–170.

Cai, H. y Eun, D. Y. (2009). Crossing over the bounded domain: From exponential to power-law intermeeting time in mobile ad hoc networks. *IEEE/ACM Trans. Netw.*, **17**(5): 1578–1591.

Calabrese, F., Di Lorenzo, G., y Ratti, C. (2010). Human mobility prediction based on individual and collective geographical preferences. En: *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, Sept. pp. 312–317.

Calabrese, F., Diao, M., Lorenzo, G. D., Jr., J. F., y Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, **26**(0): 301 – 313.

Cao, H., Mamoulis, N., y Cheung, D. (2007). Discovery of periodic patterns in spatio-temporal sequences. *Knowledge and Data Engineering, IEEE Transactions on*, **19**(4): 453–467.

Cardei, I., Liu, C., Wu, J., y Yuan, Q. (2008). Dtn routing with probabilistic trajectory prediction. En: *Proceedings of the Third International Conference on Wireless Algorithms, Systems, and Applications*, Dallas, Texas, USA. Springer-Verlag, WASA '08, pp. 40–51.

Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R., y Scott, J. (2007). Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, **6**(6): 606–620.

Chang, Y.-J., Liu, H.-H., y Wang, T.-Y. (2009). Mobile social networks as quality of life technology for people with severe mental illness. *Wireless Communications, IEEE*, **16**(3): 34–40.

Chen, G. y Kotz, D. (2000). A survey of context-aware mobile computing research. Reporte técnico, Hanover, NH, USA.

Cheng, C., Jain, R., y van den Berg, E. (2003). Wireless internet handbook. CRC Press, Inc., Boca Raton, FL, USA, capítulo Location Prediction Algorithms for Mobile Wireless Systems, pp. 245–263.

Cho, E., Myers, S. A., y Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. En: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA. ACM, KDD '11, pp. 1082–1090.

Chon, Y., Shin, H., Talipov, E., y Cha, H. (2012). Evaluating mobility models for temporal prediction with high-granularity mobility data. En: *Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications*, Lugano, Switzerland. IEEE, Percom, pp. 206–212.

Cook, D. J., Youngblood, M., Edwin O. Heierman, I., Gopalratnam, K., Rao, S., Litvin, A., y Khawaja, F. (2003). Mavhome: An agent-based smart home. En: *2003 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Fort Worth, TX, USA. IEEE Computer Society, p. 521.

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., y Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, **107**(52): 22436–22441.

Das, S., Cook, D., Battacharya, A., Heierman, E.O., I., y Lin, T.-Y. (2002). The role of prediction algorithms in the mavhome smart home architecture. *Wireless Communications, IEEE*, **9**(6): 77–84.

Djordjevic, B., Gudmundsson, J., Pham, A., y Wolle, T. (2011). Detecting regular visit patterns. *Algorithmica*, **60**(4): 829–852.

Do, T. M. T. y Gatica-Pérez, D. (2012). Contextual conditional models for smartphone-based human mobility prediction. En: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, September, Pittsburgh, Pennsylvania, USA. ACM, UbiComp '12, pp. 163–172.

Domenico, M. D., Lima, A., y Musolesi, M. (2013). Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, **9**(6): 798 – 807.

Eagle, N. y Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal Ubiquitous Computing*, **10**(4): 255–268.

Eagle, N. y Pentland, A. (2009). Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiol.*, **63**(7): 1057–1066.

Ellis, C., Scott, J., Hazas, M., y Krumm, J. (2012). Earlyoff: Using house cooling rates to save energy. En: *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, Toronto, Ontario, Canada. ACM, BuildSys '12, pp. 39–41.

Farrahi, K. y Gatica-Pérez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intell. Systems and Technology*, **2**(1): 1–27.

Furletti, B., Gabrielli, L., Renso, C., y Rinzivillo, S. (2013). Analysis of gsm calls data for understanding user mobility behavior. En: *Proceedings of the 2013 IEEE International Conference on Big Data*, Santa Clara, CA, USA. IEEE, pp. 550–555.

Gao, H., Barbier, G., y Goolsby, R. (2011a). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, **26**(3): 10–14.

Gao, H., Wang, X., Barbier, G., y Liu, H. (2011b). Promoting coordination for disaster relief: From crowdsourcing to coordination. En: *Proceedings of the 4th International Conference on Social Computing, Behavioral-cultural Modeling and Prediction*, College Park, MD, USA. Springer-Verlag, SBP'11, pp. 197–204.

Gellert, A. y Vintan, L. (2006). Person movement prediction using hidden markov models. *Studies in Informatics and Control*, **15**(1): 17–30.

Gong, Y., Li, Y., Jin, D., Su, L., y Zeng, L. (2011). A location prediction scheme based on social correlation. En: *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, May, Budapest, Hungary. IEEE, pp. 1–5.

Gonzalez, M. C., Hidalgo, C. A., y Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, **453**(7196): 779–782.

Gupta, M., Intille, S. S., y Larson, K. (2009). Adding GPS-Control to Traditional Thermostats: An Exploration of Potential Energy Savings and Design Challenges. En: *Proceedings of the 7th International Conference on Pervasive Computing (Pervasive '09)*. Springer-Verlag, Pervasive '09, pp. 95–114.

Hossmann, T., Spyropoulos, T., y Legendre, F. (2011). Putting contacts into context: mobility modeling beyond inter-contact times. En: *Proceedings of the Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHocs'11)*. ACM Press, pp. 1–11.

Hsu, W., Spyropoulos, T., Psounis, K., y Helmy, A. (2007a). Modeling time-variant user mobility in wireless mobile networks. En: *Proceedings of the 26th IEEE International Conference on Computer Communications*, Anchorage, Alaska, USA. INFOCOM, pp. 758–766.

Hsu, W., Spyropoulos, T., Psounis, K., y Helmy, A. (2007b). Modeling timevariant user mobility in wireless mobile networks. En: *Proceedings IEEE INFOCOM*.

Kang, J. H., Welbourne, W., Stewart, B., y Borriello, G. (2005). Extracting places from traces of locations. *SIGMOBILE Mobile Computing and Communications Review*, **9**(3): 58–68.

Karypis, G. (2001). Evaluation of item-based top-n recommendation algorithms. En: *Proceedings of the Tenth International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA. ACM, CIKM '01, pp. 247–254.

Kim, M., Kotz, D., y Kim, S. (2006). Extracting a mobility model from real user traces. En: *Proceedings of the 25th IEEE International Conference on Computer Communications*, April, Barcelona, Spain. IEEE, INFOCOM '06, pp. 1 –13.

Korpipaa, P., Mantyjarvi, J., Kela, J., Keranen, H., y Malm, E.-J. (2003). Managing context information in mobile devices. *IEEE Pervasive Computing*, **2**(3): 42–51.

Kotz, D., Henderson, T., Abyzov, I., y Yeo, J. (2007a). CRAWDAD trace dartmouth/campus/movement/infocom04 (v 2004-08-05). Recuperado en 27 de Noviembre de 2014 de: http://crawdad.cs.dartmouth.edu/dartmouth/campus.

Kotz, D., Henderson, T., Abyzov, I., y Yeo, J. (2007b). CRAWDAD trace dartmouth/ campus/movement/01 04 (v. 2005-03-08). Recuperado en 27 de Noviembre de 2014 de: http://crawdad.org/dartmouth/campus/.

Krumm, J. y Brush, A. J. B. (2011). Learning time-based presence probabilities. En: *Proceedings of the 9th International Conference on Pervasive Computing*, San Francisco, California, USA. Springer-Verlag, Pervasive'11, pp. 79–96.

Krumm, J. y Horvitz, E. (2006). Predestination: inferring destinations from partial trajectories. En: *Proceedings of the 8th International Conference on Ubiquitous Computing*, Orange County, California, USA. Springer-Verlag, UbiComp'06, pp. 243–260.

Krumm, J. y Rouhana, D. (2013). Placer: Semantic place labels from diary data. En: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Zurich, Switzerland. ACM, UbiComp '13, pp. 163–172.

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., y Schilit, B. (2005). Place lab: Device positioning using radio beacons in the wild. En: *Proceedings of the*

Third International Conference on Pervasive Computing, Munich, Germany. Springer-Verlag, PERVASIVE'05, pp. 116–133.

Laurila, J. K., Gatica-Pérez, D., Aad, I., Blom, J., Bornet, O., Do, T., Dousse, O., Eberle, J., y Miettinen, M. (2012). The mobile data challenge: Big data for mobile computing research. En: *Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conference on Pervasive Computing*, Newcastle, UK.

Lee, J., Kim, H., y Kim, K. J. (2003a). Resource reservation and allocation based on direction prediction for handoff in mobile multimedia networks. En: *Computational Science — ICCS 2003*, Vol. 2660 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 555–565.

Lee, J., Kim, H., y Kim, K. J. (2003b). Resource reservation and allocation based on direction prediction for handoff in mobile multimedia networks. En: *International Conference on Computational Science*. Springer, Vol. 2660 de *Lecture Notes in Computer Science*, pp. 555–565.

Lee, J.-S. y Hoh, B. (2010). Dynamic pricing incentive for participatory sensing. *Pervasive Mob. Computing*, **6**(6): 693–708.

Lee, M.-J. y Chung, C.-W. (2011). A user similarity calculation based on the location for social network services. En: *Database Systems for Advanced Applications*, Vol. 6587 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 38–52.

Lenczner, M., Grégoire, B., y Proulx, F. (2007). CRAWDAD data set ilesansfil/wifidog (v. 2007-09-07). Recuperado en 27 de Noviembre de 2014 de: http://crawdad.org/ilesansfil/wifidog/.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., y Ma, W.-Y. (2008). Mining user similarity based on location history. En: *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Irvine, California, USA. ACM, GIS '08, pp. 34:1–34:10.

Li, Z., Ding, B., Han, J., Kays, R., y Nye, P. (2010). Mining periodic behaviors for moving objects. En: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA. ACM, KDD '10, pp. 1099–1108.

Lian, D., Zheng, V. W., y Xie, X. (2013). Collaborative filtering meets next check-in location prediction. En: *Proceedings of the 22Nd International Conference on World Wide Web Companion*, Rio de Janeiro, Brazil. International World Wide Web Conferences Steering Committee, WWW '13 Companion, pp. 231–232.

Liao, L., Patterson, D. J., Fox, D., y Kautz, H. (2006). Building personal maps from gps data. *Annals of the New York Academy of Sciences*, **1093**(1): 249–265.

Lin, D.-B., Juang, R.-T., y Lin, H.-P. (2004). Mobile location estimation and tracking for gsm systems. En: *15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2004.*, Sept, Barcelona, Spain. IEEE, PIMRC 2004, pp. 2835–2839 Vol.4.

Liu, G. y Maguire, Jr., G. (1996). A class of mobile motion prediction algorithms for wireless mobile computing and communication. *Mob. Netw. Appl.*, **1**(2): 113–121.

Markov, A. A. (1961). *Theory of Algorithms*. Israel Program for Scientific Translations. Bloomington, IN, USA.

Marmasse, N. y Schmandt, C. (2000). Location-aware information delivery with commotion. En: *Proceedings of the 2Nd International Symposium on Handheld and Ubiquitous Computing*, Bristol, UK. Springer-Verlag, HUC '00, pp. 157–171.

Mathew, W., Raposo, R., y Martins, B. (2012). Predicting future locations with hidden markov models. En: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, Pittsburgh, Pennsylvania, USA. ACM, UbiComp '12, pp. 911–918.

McGee, J., Caverlee, J., y Cheng, Z. (2013). Location prediction in social media based on tie strength. En: *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, San Francisco, California, USA. ACM, CIKM '13, pp. 459–468.

McInerney, J., Zheng, J., Rogers, A., y Jennings, N. R. (2013). Modelling heterogeneous location habits in human populations for location prediction under data sparsity. En: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Zurich, Switzerland. ACM, UbiComp '13, pp. 469–478.

Miluzzo, E., Lane, N. D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S. B., Zheng, X., y Campbell, A. T. (2008). Sensing meets mobile social networks: The design, implementation and evaluation of the cenceme application. En: *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, Raleigh, NC, USA. ACM, SenSys '08, pp. 337–350.

Monreale, A., Pinelli, F., Trasarti, R., y Giannotti, F. (2009). Wherenext: a location predictor on trajectory pattern mining. En: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France. ACM, KDD '09, pp. 637–646.

Montoliu, R. y Gatica-Pérez, D. (2010). Discovering human places of interest from multimodal mobile phone data. En: *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, Limassol, Cyprus. ACM, MUM '10, pp. 12:1–12:10.

Montoliu, R., Blom, J., y Gatica-Pérez, D. (2013). Discovering places of interest in everyday life from smartphone data. *Multimedia Tools and Applications*, **62**(1): 179–207.

Motahari, S., Zang, H., y Reuther, P. (2012). The impact of temporal factors on mobility patterns. En: *Proceedings of the 2012 45th Hawaii International Conference on System Sciences*, Grand Wailea, Maui, Hawaii. IEEE Computer Society, HICSS '12, pp. 5659–5668.

Newman, M. E. (2002). The spread of epidemic disease on networks. *Physical Review Letters*, **66**(1).

Nguyen, H. A. y Giordano, S. (2012). Context information prediction for social-based routing in opportunistic networks. *Ad Hoc Netw.*, **10**(8): 1557–1569.

Nguyen, L. T., Cheng, H.-T., Wu, P., Buthpitiya, S., Zhu, J., y Zhang, Y. (2012). Pnlum : System for prediction of next location for users with mobility. En: *In Nokia Mobile Data Challenge 2012 Workshop*. ACM.

Nicholson, A. J. y Noble, B. D. (2008). Breadcrumbs: forecasting mobile connectivity. En: *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking*, San Francisco, California, USA. ACM, MobiCom '08, pp. 46–57.

Noulas, A., Scellato, S., Lathia, N., y Mascolo, C. (2012). Mining user mobility features for next place prediction in location-based services. En: *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, Brussels, Belgium. IEEE Computer Society, ICDM'12, pp. 1038–1043.

Palma, A. T., Bogorny, V., Kuijpers, B., y Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. En: *Proceedings of the 2008 ACM Symposium on Applied Computing*, Fortaleza, Ceara, Brazil. ACM, SAC '08, pp. 863–868.

Petzold, J., Bagci, F., Trumler, W., y Ungerer, T. (2005a). Next location prediction within a smart office building. En: *Proceedings of 1st International Workshop on Exploiting Context Histories in Smart Environments (ECHISE'05) at the 3rd International Conference on Pervasive Computing*, Munich, Germany. Springer LNCS series.

Petzold, J., Pietzowski, A., Bagci, F., Trumler, W., y Ungerer, T. (2005b). Prediction of indoor movements using bayesian networks. En: *Location- and Context-Awareness*. Springer Berlin Heidelberg, Vol. 3479 de *Lecture Notes in Computer Science*, pp. 211–222.

Piorkowski, M., Sarafijanovic-Djukic, N., y Grossglauser, M. (2009). CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.org/epfl/mobility/.

Pirozmand, P., Wu, G., Jedari, B., y Xia, F. (2014). Human mobility in opportunistic networks: Characteristics, models and prediction methods. *Journal of Network and Computer Applications*, **42**(0): 45–58.

Prasad, P. S. y Agrawal, P. (2010). A generic framework for mobility prediction and resource utilization in wireless networks. En: *Proceedings of the 2nd International Conference on COMmunication Systems and NETworks*, Bangalore, India. IEEE Press, COMSNETS'10, pp. 233–242.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, **77**(2): 257–286.

Ram, A., Jalal, S., Jalal, A. S., y Kumar, M. (2010). A density based algorithm for discovering density varied clusters in large spatial databases. *Int. J. of Computer Applications*, **3**(6): 1–4.

Reddy, S., Estrin, D., Hansen, M., y Srivastava, M. (2010). En: *Proceedings of the 13th International Conference on Ubiquitous Computing (Ubicomp'10)*. ACM, Ubicomp '10, pp. 33–36.

Sadilek, A. y Krumm, J. (2012a). Far out: Predicting long-term human mobility. En: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, Toronto, Ontario, Canada. AAAI Press, AAAI.

Sadilek, A. y Krumm, J. (2012b). Far out: Predicting long-term human mobility. En: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, Toronto, Ontario, Canada. AAAI Press, AAAI.

Sadilek, A., Kautz, H., y Bigham, J. P. (2012a). Finding your friends and following them to where you are. En: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, Seattle, Washington, USA. ACM, WSDM '12, pp. 723–732.

Sadilek, A., Kautz, H. A., y Silenzio, V. (2012b). Modeling spread of disease from social interactions. En: *Proceedings of the Sixth International Conference on Weblogs and Social Media*, Dublin, Ireland. The AAAI Press, ICWSM.

Sarwar, B., Karypis, G., Konstan, J., y Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. En: *Proceedings of the 10th International Conference on World Wide Web*, Hong Kong, Hong Kong. ACM, WWW '01, pp. 285–295.

Satyanarayanan, M. (2001). Pervasive computing: vision and challenges. *Personal Communications, IEEE*, **8**(4): 10–17.

Scellato, S., Musolesi, M., Mascolo, C., Latora, V., y Campbell, A. T. (2011). NextPlace: a spatio-temporal prediction framework for pervasive systems. En: *Proceedings of the 9th International Conference on Pervasive Computing*, San Francisco, CA, USA. Springer-Verlag, Pervasive'11, pp. 152–169.

Schafer, J. B., Konstan, J., y Riedl, J. (1999). Recommender systems in e-commerce. En: *Proceedings of the 1st ACM Conference on Electronic Commerce*, Denver, Colorado, USA. ACM, EC '99, pp. 158–166.

Scott, J., Bernheim Brush, A., Krumm, J., Meyers, B., Hazas, M., Hodges, S., y Villar, N. (2011). Preheat: Controlling home heating using occupancy prediction. En: *Proceedings of the 13th International Conference on Ubiquitous Computing*, Beijing, China. ACM, UbiComp '11, pp. 281–290.

Sodkomkham, D., Legaspi, R., Fukui, K.-i., Moriyama, K., Kurihara, S., y Numao, M. (2013). App: Aperiodic and periodic model for long-term human mobility prediction using ambient simple sensors. En: *Proceedings of the 4th International Workshop on Mining Ubiquitous and Social Environments*, Prague, Czech Republic. Springer LNCS, MUSE 2013, pp. 3–18.

Song, L., Kotz, D., Jain, R., y He, X. (2003). Evaluating location predictors with extensive Wi-Fi mobility data. *SIGMOBILE Mob. Computing and Communications Review*, **7**(4): 64–65.

Song, L., Deshpande, U., Kozat, U. C., Kotz, D., y Jain, R. (2006). Predictability of wlan mobility and its effects on bandwidth provisioning. En: *Proceedings of the 25th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies*, April, Barcelona, Spain. IEEE, INFOCOM '06.

Tarasov, A., Kling, F., y Pozdnoukhov, A. (2013). Prediction of user location using the radiation model and social check-ins. En: *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, Chicago, Illinois. ACM, UrbComp '13, pp. 8:1–8:7.

Terry, M., Mynatt, E. D., Ryall, K., y Leigh, D. (2002). Social net: Using patterns of physical proximity over time to infer shared interests. En: *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, Minneapolis, Minnesota, USA. ACM, CHI EA '02, pp. 816–817.

Viterbi, A. J. (2006). A personal history of the Viterbi algorithm. *IEEE Signal Processing Magazine*, **23**(4): 120–142.

Vu, L., Do, Q., y Nahrstedt, K. (2011a). Exploiting joint wifi/bluetooth trace to predict people movement.

Vu, L., Do, Q., y Nahrstedt, K. (2011b). Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace. En: *Proceedings of the 2011 IEEE International Conference on Pervasive Computing and Communications*, Seattle, WA, USA. IEEE Computer Society, PERCOM '11, pp. 54–62.

Weiser, M. (1991). The computer for the 21st century. *Scientific American*, **265**(3): 66–75.

Xiong, H., Zhang, D., Zhang, D., y Gauthier, V. (2012). Predicting mobile phone user locations by exploiting collective behavioral patterns. En: *2012 9th International Conference on Ubiquitous Intelligence Computing and 9th International Conference on Autonomic Trusted Computing*, Sept, Fukuoka, Japan. IEEE, UIC/ATC, pp. 164–171.

Yavas, G., Katsaros, D., Ulusoy, O., y Manolopoulos, Y. (2004). A data mining approach for location prediction in mobile environments. *Data & Knowl. Engineering*, **54**(2005): 121–146.

Ye, Y., Zheng, Y., Chen, Y., Feng, J., y Xie, X. (2009). Mining individual life pattern based on location history. En: *Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, Taipei, China. IEEE Computer Society, pp. 1–10.

Ying, J. J.-C., Lu, E. H.-C., Lee, W.-C., Weng, T.-C., y Tseng, V. S. (2010). Mining user similarity from semantic trajectories. En: *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, San Jose, California, USA. ACM, LBSN '10, pp. 19–26.

Ying, J. J.-C., Lee, W.-C., Weng, T.-C., y Tseng, V. S. (2011). Semantic trajectory mining for location prediction. En: *Proceedings of the 19th ACM SIGSPATIAL International*

Conference on Advances in Geographic Information Systems, Chicago, Illinois, USA. ACM, GIS '11, pp. 34–43.

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., y Huang, Y. (2010). T-drive: Driving directions based on taxi trajectories. En: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, San Jose, California, USA. ACM, GIS '10, pp. 99–108.

Yuan, J., Zheng, Y., Zhang, L., Xie, X., y Sun, G. (2011). Where to find my next passenger. En: *Proceedings of the 13th International Conference on Ubiquitous Computing*, Beijing, China. ACM, UbiComp '11, pp. 109–118.

Yuan, J., Zheng, Y., Xie, X., y Sun, G. (2013). T-drive: Enhancing driving directions with taxi drivers' intelligence. *Knowledge and Data Engineering, IEEE Transactions on*, **25**(1): 220–232.

Yun, J., Zheng, Y., Zhang, L., Xie, X., y Sun, G. (2011). Where to find my next passenger. En: *Proceedings of the 13th international Conference on Ubiquitous Computing*, Beijing, China. ACM, UbiComp '11, pp. 109–118.

Zhang, W. (2001). *Algorithms for partially observable Markov decision processes*. Tesis de doctorado, Hong Kong University of Science and Technology.

Zhang, Y. F., Zhang, Q. F., y Yu, R. H. (2010). Markov property of markov chains and its test. En: *Proceedings of the International Conference on Machine Learning and Cybernetics*, Qingdao,China. IEEE, ICMLC'10, pp. 1864 –1867.

Zheng, Y. y Xie, X. (2010). Learning location correlation from gps trajectories. En: *Proceedings of the 2010 Eleventh International Conference on Mobile Data Management*, Washington, DC, USA. IEEE Computer Society, pp. 27–32.

Zheng, Y., Li, Q., Chen, Y., Xie, X., y Ma, W.-Y. (2008). Understanding mobility based on gps data. En: *Proceedings of the 10th International Conference on Ubiquitous Computing*, Seoul, Korea. ACM, UbiComp '08, pp. 312–321.

Zheng, Y., Zhang, L., Xie, X., y Ma, W.-Y. (2009). Mining interesting locations and travel sequences from gps trajectories. En: *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain. ACM, WWW '09, pp. 791–800.

Zheng, Y., Xie, X., y Ma, W.-Y. (2010). Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Engineering Bulletin*, **33**(2): 32–39.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., y Ma, W.-Y. (2011). Recommending friends and locations based on individual location history. *ACM Trans. Web*, **5**(1): 5:1–5:44.

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., y Terveen, L. (2007). Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems*, **25**(3): 56–68.

Ziebart, B. D., Maas, A. L., Dey, A. K., y Bagnell, J. A. (2008). Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. En: *Proceedings of the 10th International Conference on Ubiquitous Computing*, Seoul, Korea. ACM, UbiComp '08, pp. 322–331.