



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

# **Método para el cálculo de la similitud entre dos trayectorias basado en los sitios visitados, las actividades ejecutadas y la cronología de los episodios**

**Santiago Román Fernández**

Universidad Nacional de Colombia  
Facultad de Minas, Departamento de Ciencias de la Computación y la Decisión  
Medellín, Colombia

2018



# **Método para el cálculo de la similitud entre dos trayectorias basado en los sitios visitados, las actividades ejecutadas y la cronología de los episodios**

**Santiago Román Fernández**

Tesis presentada como requisito parcial para optar al título de:

**Magister en Ingeniería de Sistemas**

Director (a):

Ph.D., Francisco Javier Moreno Arboleda

Codirector (a):

Ph.D., Jaime Alberto Guzmán

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y la Decisión

Medellín, Colombia

2018



## Resumen

Hallar trayectorias semánticas similares es de gran utilidad en campos como el mercadeo o las redes sociales, ya que permite encontrar usuarios con gustos y preferencias similares. No es una tarea trivial puesto que gran cantidad de factores influyen en el cálculo de la similitud. En esta tesis se propone un método para calcular la similitud de acuerdo con los sitios visitados, basado en un árbol de categorías que permite relacionar los tipos de sitios de manera jerárquica, brindando flexibilidad y adaptabilidad a diferentes dominios de aplicación. Dicho método se puede utilizar para calcular la similitud basada en las actividades ejecutadas por los usuarios en lugar de los sitios visitados; y se propone una extensión de este para tener en cuenta la cronología de los hechos como criterio temporal. A través de datos de usuarios reales se evidencia el funcionamiento del método y se compara con otros métodos actuales.

Se analizan también diversos factores temporales que pueden influir en la medida de similitud de dos trayectorias y como han sido abordados por diferentes autores. Se plantea una serie de retos y situaciones a considerar en el aspecto temporal, y se propone un método basado en la cronología (orden) para abordar el problema.

**Palabras clave:** trayectorias semánticas, similitud de trayectorias, datos espacio temporales.

## Abstract

Finding similar semantic trajectories can be useful in fields such as marketing or social networks, since it allows you to find users with similar likes and preferences. However, it is not a trivial task since many factors influence the calculation of similarity. In this thesis a method is proposed to calculate the similarity according to the visited sites, based on a tree of categories that allows to relate the types of sites in a hierarchical way, providing flexibility and adaptability to different application domains. This method can be used to calculate the similarity based on the activities executed by the users instead of the visited sites; and an extension of this is proposed to consider the chronology of events as temporal criteria. Data collected from real users is used to see how the method operates and it is compared with other current methods.

Different temporal factors are analyzed and is shown how they can influence the similarity measurement of two trajectories. How they have been addressed by different authors are also analyzed. Some challenges and situations to consider in the temporal aspect and a method based on chronology (order) are proposed to address the problem.

**Keywords: semantic trajectories, trajectories similarity, spatiotemporal data.**

# Contenido

	Pág.
<b>Resumen .....</b>	<b>V</b>
<b>Lista de figuras.....</b>	<b>IX</b>
<b>Lista de tablas .....</b>	<b>XI</b>
<b>Introducción.....</b>	<b>1</b>
Objetivos.....	3
<b>1. Estado del arte .....</b>	<b>5</b>
1.1 Motivación .....	11
1.2 Trabajos relacionados .....	16
1.3 Trabajo 1: Xiao et al. ....	19
1.4 Trabajo 2: Tiakas et al.....	21
1.5 Trabajo 3: Chang et al.....	23
1.6 Trabajo 4: Kondaveeti .....	25
1.7 Análisis.....	29
<b>2. Similitud de trayectorias semánticas basada en sitios y actividades .....</b>	<b>33</b>
2.1 Similitud de trayectorias .....	36
2.2 Método 1 .....	42
2.3 Método 2 .....	44
2.4 Diferencias e interpretación de los dos métodos .....	45
2.5 Algoritmos de similitud entre trayectorias .....	47
2.6 Similitud combinada: Sitios y actividades .....	49
2.7 Experimentos y resultados .....	50
<b>3. Similitud de trayectorias basada en la cronología de los episodios .....</b>	<b>57</b>
3.1 Motivación .....	58
3.2 Métodos para identificar subsecuencias comunes.....	61
3.3 Cálculo de la similitud temporal .....	67
3.4 Algoritmo .....	69
3.5 Niveles de análisis.....	71
3.6 Similitud multidimensional .....	73
3.7 Experimentos .....	75
<b>4. Conclusiones y recomendaciones.....</b>	<b>79</b>
4.1 Conclusiones.....	79
4.2 Cumplimiento de objetivos .....	80
4.3 Recomendaciones.....	81

VIII Método para el cálculo de la similitud entre dos trayectorias basado en los sitios visitados, las actividades ejecutadas y la cronología de los episodios

---

4.4 Difusión.....81

**Bibliografía ..... 85**

## Lista de figuras

	Pág.
Figura 1. Dos trayectorias similares en forma.....	5
Figura 2. Dos trayectorias similares en velocidad.....	6
Figura 3. Dos trayectorias similares en modo de transporte .....	7
Figura 4. Una trayectoria representada como un conjunto de <i>stops</i> y <i>moves</i> .....	8
Figura 5. Trayectorias con actividades .....	8
Figura 6. Actividades ejecutadas por tres usuarios A, B y C durante un día entre las 8 am y las 6 pm.....	11
Figura 7. Actividades ejecutadas por tres usuarios A, B y C durante una semana .....	12
Figura 8. Ejemplos de combinaciones de algunos criterios.....	15
Figura 9. Taxonomía de clasificación de criterios .....	18
Figura 10. Relación entre el árbol de categorías y las diferentes capas. Fuente (Xiao et al., 2012).....	20
Figura 11. Ejemplo de una red espacial.....	21
Figura 12. Ejemplo de trayectorias con 4 nodos.....	23
Figura 13. Ejemplo de dos trayectorias similares. Fuente (Kondaveeti, 2012).....	26
Figura 14. Ejemplo de creación de clase sintética. Fuente (Kondaveeti, 2012) .....	26
Figura 15. Trayectorias 1 y 9.....	27
Figura 16. Ejemplo de dos trayectorias semánticas A y B .....	34
Figura 17. Ejemplo de CTCS.....	37
Figura 18. Ejemplo de CTCA.....	38
Figura 19. CTCS con valores de similitud para T <sub>1</sub> y T <sub>2</sub> usando el método 1 .....	43
Figura 20. CTCS con valores de similitud para T <sub>1</sub> y T <sub>2</sub> usando el método 2 .....	45
Figura 21. Diferentes valores de similitud obtenidos con el método 1 y método 2 .....	46
Figura 22. CTCA con valores de similitud para T <sub>1</sub> y T <sub>2</sub> usando el método 1 .....	49
Figura 23. CTCA con valores de similitud para el nodo <i>ns</i> = Entretenimiento para T <sub>1</sub> y T <sub>2</sub> usando el método 1.....	50
Figura 24. CTCS usado para los experimentos .....	51
Figura 25. CTCA usado para los experimentos .....	51
Figura 26. Resultados para los pares de trayectorias que obtuvieron la mayor similitud respecto a los sitios.....	53
Figura 27. Comparación de resultados cuando <i>nmsw</i> = 0 .....	55
Figura 28. Comparación de resultados cuando <i>nmsw</i> = 0.5 .....	56
Figura 29. Comparación de resultados cuando <i>nmsw</i> = 1 .....	56
Figura 30. Ejemplo de dos trayectorias .....	58
Figura 31. Subsecuencias en común de dos personas A y B .....	58

X Método para el cálculo de la similitud entre dos trayectorias basado en los sitios visitados, las actividades ejecutadas y la cronología de los episodios

---

Figura 32. Similitud temporal por cada uno de los posibles caminos según el método 1a para T3 y T4 .....	68
Figura 33. Relación de un episodio con el CTCS y CTCA.....	72
Figura 34. Foco de esta tesis .....	74
Figura 35. Resultados del método 1a.....	76
Figura 36. Resultados del método 1b.....	76
Figura 37. Resultados del método 2.....	76
Figura 38. Resultados del método 3.....	77
Figura 39. Similitud promedio para cada uno de los métodos .....	77

## Lista de tablas

	<b>Pág.</b>
Tabla 1. Parejas más similares de acuerdo con los criterios seleccionados .....	13
Tabla 2. Rangos de tiempo de un día. Fuente (Chang et al., 2007) .....	24
Tabla 3. Grupos de rangos de tiempo. Fuente (Chang et al., 2007) .....	24
Tabla 4. Falencias encontradas en los trabajos analizados .....	29
Tabla 5. Eventos de la trayectoria Ti .....	39
Tabla 6. Episodios de la trayectoria T2.....	42
Tabla 7. Resultados de los métodos 1 y 2 respecto a los sitios .....	52
Tabla 8. Resultados de los métodos 1 y 2 respecto a las actividades .....	52
Tabla 9. Resultados de similitud combinada con el método 1.....	54
Tabla 10. Resultados de similitud combinada con el método 2.....	54
Tabla 11. Grado de dureza de los diferentes métodos .....	64



# Introducción

Una trayectoria representa la posición a través del tiempo de un objeto móvil en el espacio, e.g., la posición de un usuario en una ciudad, o las coordenadas de un vehículo en una red de carreteras. Esta información es recolectada como datos de la forma  $(x, y, t)$ , donde  $x$  y  $y$  representan la longitud y latitud del objeto en el espacio y  $t$  la marca de tiempo. El conjunto de datos de este tipo es denominado trayectoria cruda (*raw trajectory*) puesto que está expresado en la forma más básica de representación, tal como fue capturada y sin ningún procesamiento posterior (Spaccapietra et al., 2008). Actualmente estos datos pueden ser recolectados gracias al amplio uso de dispositivos móviles con sistemas de ubicación integrados como el GPS, teniendo en cuenta las implicaciones éticas y legales que esto conlleva, tales como aprobaciones del usuario, aceptación de términos y seguridad de la información.

Si bien las trayectorias crudas permiten ubicar un objeto en el espacio a través del tiempo, son incapaces por sí solas de brindar información acerca de los sitios visitados por el objeto que representan. Diversos autores (Alvares et al., 2007; Parent et al., 2013) han propuesto métodos para enriquecer las trayectorias crudas con información que pueda ser más relevante y útil en diversos campos. Estas trayectorias se denominan trayectorias semánticas (*semantic trajectory*), y brindan información más detallada acerca de los sitios visitados por el usuario. La definición más aceptada de trayectoria semántica considera a la misma como un conjunto de *stops* y *moves*, donde un *stop* representa una parada del usuario en un sitio (e.g., un centro comercial, un supermercado) y un *move* el desplazamiento entre dos *stops* consecutivos (Spaccapietra et al., 2008).

La similitud de trayectorias puede estar basada en criterios espaciales, temporales, semánticos o en una combinación de varios de estos. Encontrar trayectorias similares basadas en criterios espaciales como la forma o dirección permite, por ejemplo, agrupar vehículos con rutas de transporte similares en la red de carreteras de una ciudad, o grupos

de animales inmigrantes con la misma ruta o destino de inmigración. Mientras que los criterios semánticos como los sitios visitados, permite encontrar usuarios con gustos y preferencias similares, e.g., personas que les guste visitar restaurantes de comida italiana, o que prefieran hacer deporte en la mañana antes de ir a la oficina.

Diversos autores (Liu & Schneider, 2012; Xiao, Zheng, Luo, & Xie, 2012; X. Zhao, 2011) han propuesto diversos métodos para el cálculo de similitud de trayectorias basado en criterios temporales y espaciales, apoyados en herramientas como la distancia euclidiana, distancia entre nodos de un grafo o funciones de tiempo como DTW (*Dynamic time warping*). Sin embargo, pocos trabajos han considerado la similitud entre trayectorias basado en aspectos semánticos. Algunos (Furtado, Kopanaki, Alvares, & Bogorny, 2015; Liu & Schneider, 2012) se han centrado en utilizar el sitio, o tipo de sitio (e.g., comida italiana, deportes, entretenimiento) como criterio base, pero carecen de un método flexible que permita relacionar los tipos o categorías de los sitios de una forma más acorde con la realidad y el dominio de aplicación. Tampoco han considerado otros aspectos como las actividades ejecutadas en cada sitio, la duración, el momento del día o el orden de los episodios.

El capítulo 1 realiza un análisis del estado del arte actual, presentando diversos trabajos y métodos relacionados con la similitud de trayectorias y diversos aspectos temporales que se pueden considerar en la similitud de trayectorias, tales como forma, distancia, velocidad, dirección, orden, entre otros, y analiza a profundidad los trabajos más representativos. En el capítulo 2 se propone un método para calcular la similitud entre dos trayectorias basado en los tipos de sitios visitados por los usuarios y las actividades ejecutadas. Se define un árbol jerárquico de categorías de sitios (o actividades) que puede ser definido por el analista según el dominio de interés y permite relacionar entre sí los sitios visitados (o actividades ejecutadas) por diferentes usuarios. El método es extendido en el capítulo 3 para considerar ambos criterios al tiempo, permitiendo asignar pesos para definir prioridades y calculando la similitud en diferentes niveles de profundidad del árbol, brindando flexibilidad al analista según sus intereses; este capítulo deja en evidencia lo complejo que puede ser considerar dicho criterio debido a la cantidad de combinaciones y situaciones que se pueden presentar, con diferentes niveles de importancia según el dominio de aplicación y los intereses del analista.

## Objetivos

Objetivo general: El objetivo general de esta tesis es proponer un método para el cálculo de la similitud entre dos trayectorias semánticas basado en los sitios visitados, las actividades ejecutadas y la cronología de los episodios (*stops*). Para esto, la misma se desarrolla de la siguiente forma.

Objetivos específicos:

1. Identificar los métodos para el cálculo de similitud en trayectorias semánticas y determinar las ventajas y desventajas de cada uno.
2. Desarrollar un método para el cálculo de la similitud entre dos trayectorias semánticas basado en los sitios visitados y las actividades ejecutadas.
3. Extender el método desarrollado en el objetivo anterior considerando además la cronología de los episodios.
4. Desarrollar un prototipo y evaluar el método desarrollado frente a otro método.



# 1. Estado del arte

El descubrimiento de comportamientos similares a partir de los datos de trayectorias puede ser útil en diferentes dominios, e.g, en turismo se pueden diseñar rutas turísticas. Considere, por ejemplo, algunos turistas que están visitando una ciudad en un día específico. Cada turista visita un conjunto de sitios  $s_1, s_2, \dots, s_n$ . Si se identifica un subconjunto de sitios que son visitados por muchos turistas, se podría diseñar una ruta que incluya todos o algunos de esos sitios. En este ejemplo, se considera la similitud de trayectorias con respecto a los sitios visitados (el movimiento de cada turista a través de la ciudad durante el día representa una trayectoria). Sin embargo, la similitud de trayectorias puede ser considerada desde otro punto de vista, e.g, dos trayectorias pueden ser consideradas similares respecto a su forma (Liu & Schneider, 2012; Yanagisawa, Akahani, & Satoh, 2003), ver en la Figura 1.

También se puede considerar la similitud respecto a la velocidad (Liu & Schneider, 2012), ver Figura 2. Si bien esas dos trayectorias son diferentes en forma, su promedio de velocidad es el mismo.

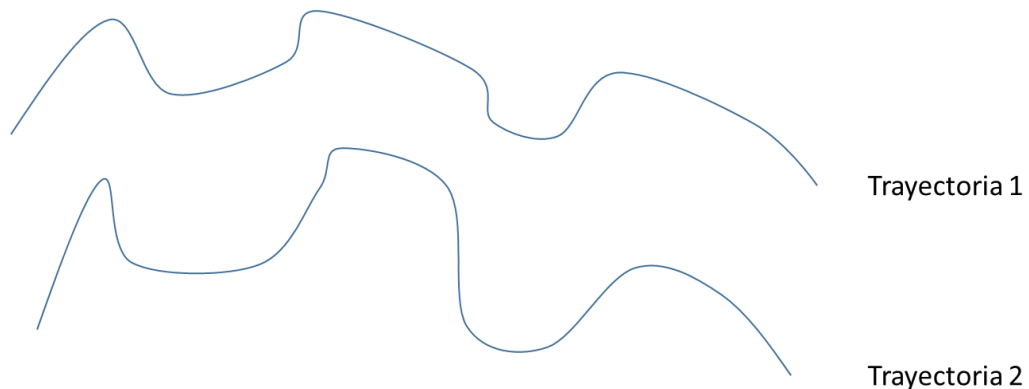


Figura 1. Dos trayectorias similares en forma

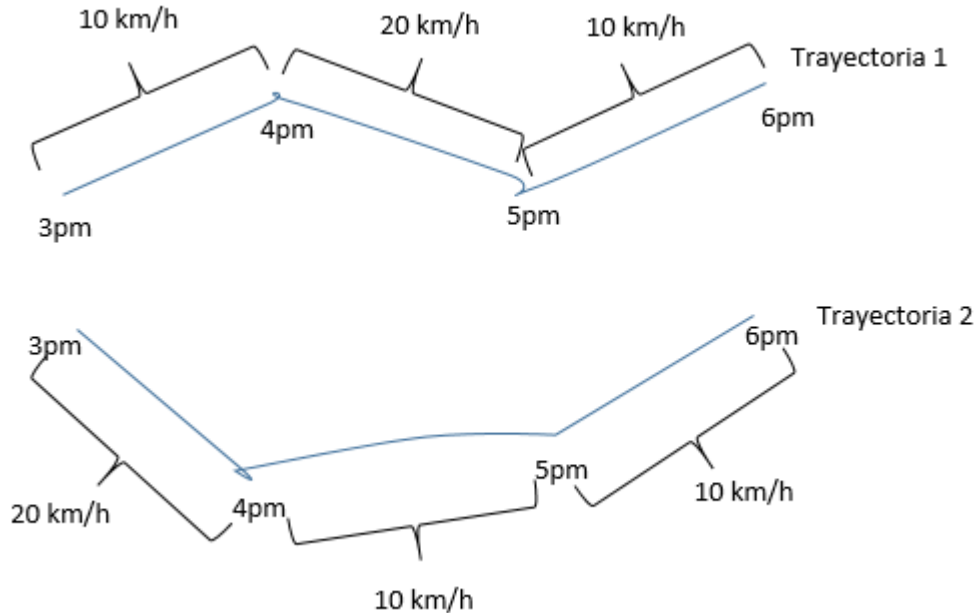


Figura 2. Dos trayectorias similares en velocidad

Otra alternativa para determinar la similitud de trayectorias es considerar el modo de transporte (Kondaveeti, 2012). Considere, por ejemplo, la Figura 3, se puede ver que, aunque las dos trayectorias tienen diferente forma, y su promedio de velocidad también es diferente, usan el mismo modo de transporte. Nótese también que usan el mismo medio de transporte en el mismo orden y cada medio de transporte fue usado por el mismo tiempo (una hora). Dependiendo de los requisitos de similitud, las trayectorias pueden ser consideradas similares, por ejemplo, si usan el mismo medio de transporte, pero no necesariamente por la misma cantidad de tiempo (un umbral puede ser establecido). La frecuencia de uso también puede ser un factor por considerar en la similitud de trayectorias. Suponga, por ejemplo, que, durante un día, un turista usó tres veces un carro, dos veces una bicicleta, y cuatro veces el subterráneo. Otro turista usó cuatro veces el carro, cuatro veces la bicicleta, y dos veces un tranvía. Las trayectorias pueden ser consideradas similares, por ejemplo, si el requisito es que los turistas tengan en común al menos dos medios de transporte y cada uno de ellos debe ser usado al menos dos veces.

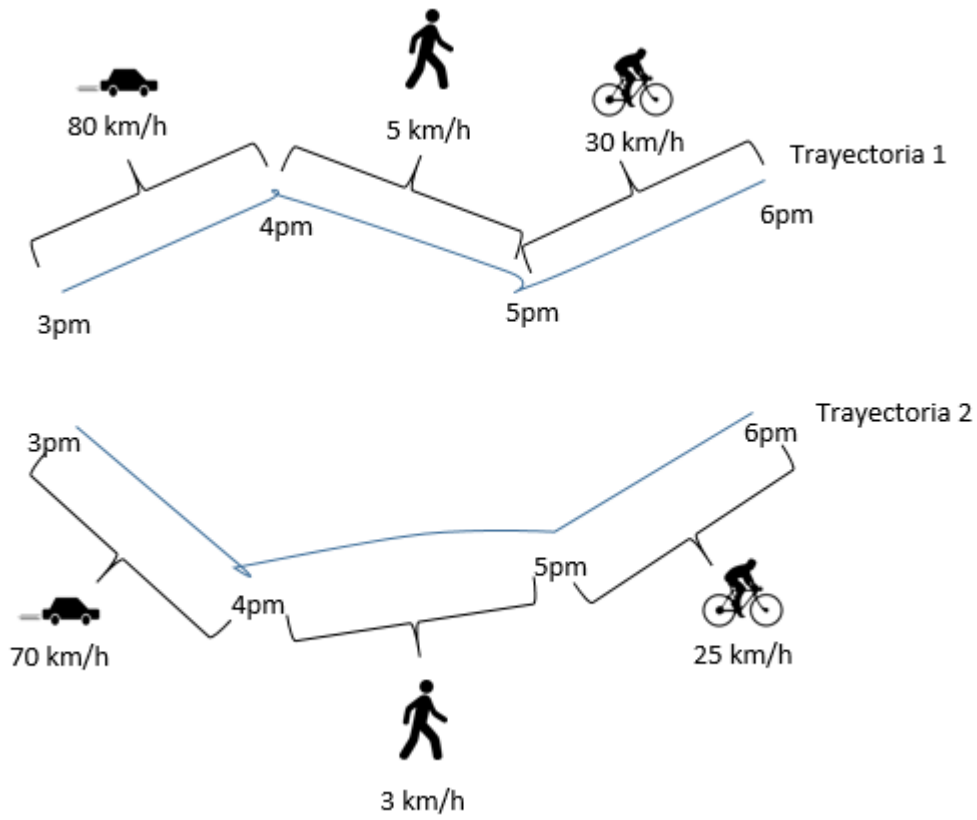


Figura 3. Dos trayectorias similares en modo de transporte

Empíricamente, una trayectoria de un objeto en movimiento es una secuencia de puntos ordenados en el tiempo de la forma  $(x, y, t)$ , donde  $(x, y)$  representa las coordenadas espaciales y  $t$  el tiempo. Esta definición corresponde a una trayectoria cruda (Spaccapietra et al., 2008). En la última década, ha habido gran cantidad de propuestas que consideran trayectorias crudas enriquecidas. Por ejemplo, Spaccapietra et al. (2008) define una trayectoria como una secuencia de *stops* y *moves*. Un *stop* representa un periodo de la trayectoria durante el cual un objeto no se movió (el objeto estuvo posiblemente visitando un sitio, por ejemplo, un restaurante). Por otro lado, un *move* representa un periodo de una trayectoria durante el cual un objeto estuvo efectivamente moviéndose (El objeto estuvo moviéndose de un punto A a un punto B). Por ejemplo, en la Figura 4, se muestra una trayectoria con cuatro *stops* (la persona durmió, almorzó, nadó y fue al teatro) y tres *moves* (note que un *move* es definido por dos *stops*).

También se puede considerar las actividades ejecutadas en los sitios. Por ejemplo, en la Figura 5, se muestra dos trayectorias de dos personas. Esas dos personas visitaron los

mismos sitios, pero ejecutaron actividades diferentes. Así, esas trayectorias son similares respecto a sitios, pero no respecto a las actividades que la persona ejecutó en esos sitios.

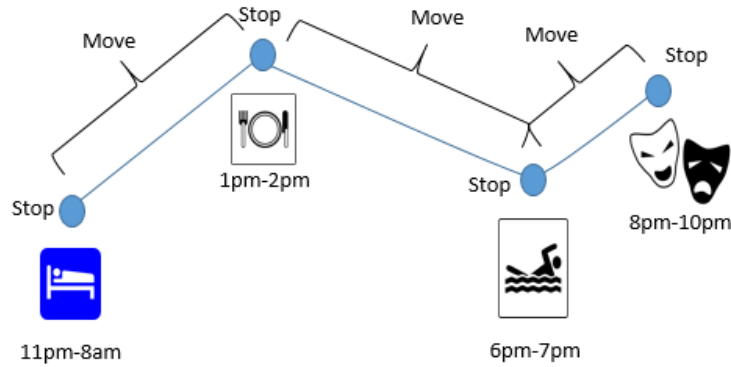


Figura 4. Una trayectoria representada como un conjunto de *stops* y *moves*

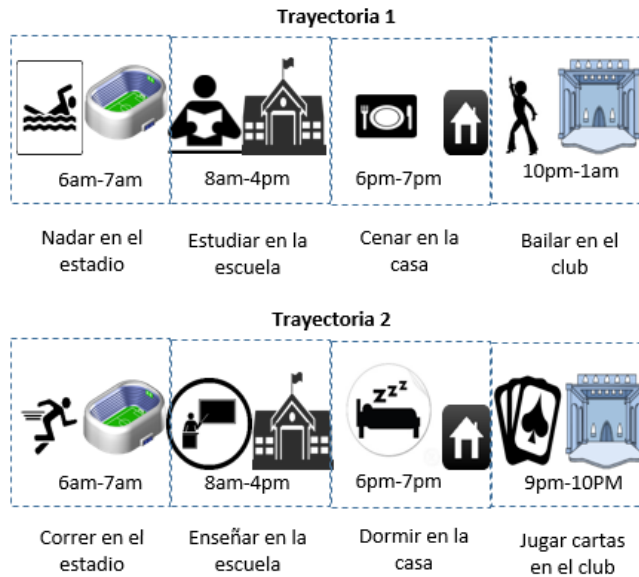


Figura 5. Trayectorias con actividades

De una forma más general, se podría decir que una trayectoria es una secuencia de episodios, donde la naturaleza de cada episodio es definida por el analista. Por ejemplo, se podría definir una trayectoria como una secuencia de episodios, se podría considerar

solo dos tipos de episodios: trabajar y estudiar, y considerar que el resto del tiempo el objeto estuvo en movimiento porque solo nos interesa ese tipo de episodios.

Durante los últimos años, diferentes aproximaciones para medir la similitud de trayectorias crudas han sido propuestas. Entre las principales está el DTW (*Dynamic Time Warping*) (Keogh & Ratanamahatana, 2004; Kruskal, 1983), desarrollado para series de tiempo, LCSS (*Longest Common Subsequence*) (Vlachos, Kollios, & Gunopulos, 2002) y EDR (*Edit Distance on Real Sequences*) (Chen, Özsu, & Oria, 2005).

Recientemente, se han hecho otros esfuerzos por enriquecer las trayectorias crudas, esto es, transformar una trayectoria cruda en una trayectoria semántica (Alvares et al., 2007; Parent et al., 2013). Además del tiempo y el espacio, las trayectorias semánticas tienen información, como el nombre de los sitios visitados por un objeto y las actividades ejecutadas en cada sitio (Bogorny et al., 2014).

Diferentes definiciones para trayectorias semánticas son encontradas en la literatura (Alvares et al., 2007) y (Bogorny et al., 2014) pero por simplicidad, se va a considerar una trayectoria semántica como una secuencia de sitios importantes llamados *stops*, como originalmente propuso Spaccapietra et al. (2008).

La primera aproximación, propuesta en 2012 por Liu & Schneider (2012), divide una trayectoria semántica en sub-trayectorias y calcula la similitud semántica de dos trayectorias basado en la máxima subsecuencia común de los sitios visitados. En esta aproximación solo una coincidencia completa es considerada, es decir, 1 si hay coincidencia y 0 si no hay coincidencia en el nombre del sitio.

Xiao, et al. (2012) propusieron una medida semántica de similitud que consiste en la semántica de los *stops*, la secuencia de los sitios visitados (*stops*), el tiempo de viaje entre los sitios y la frecuencia en que un sitio es visitado. Dos trayectorias son consideradas similares si visitan la misma secuencia de sitios, varias veces, y con tiempos de viaje similares.

En este capítulo, se propone una nueva función de similitud para trayectorias semánticas, donde consideramos una jerarquía de sitios para calcular la similitud de las trayectorias en diferentes niveles de abstracción.

A partir de las trayectorias se pueden inferir preferencias de un objeto móvil, e.g., las tiendas preferidas para comprar, las avenidas preferidas para conducir, los sitios de la selva preferidos para dormir. También se pueden inferir características de un objeto móvil (e.g., su velocidad, su forma de manejar, sus hábitos durante el día) y de su entorno (e.g., identificar las tiendas de ropa más visitadas, los periodos del día de mayor congestión, las zonas de la selva más transitadas por animales en migración).

Una trayectoria se puede enriquecer con información como, e.g., los sitios visitados o por los que pasó cerca el objeto móvil durante su desplazamiento, e.g., las tiendas visitadas por una persona en un centro comercial, los hoteles o museos por los que pasó cerca (e.g., a menos de 1 km) un taxi y los ríos atravesados por un animal en la selva.

Por su parte, con el crecimiento de las tecnologías de geolocalización como el GPS y el auge de dispositivos móviles como smartphones y tablets, hoy se generan enormes volúmenes de datos (del orden de exabytes) que pueden ser usados para analizar el comportamiento del movimiento de las personas, animales y en general de cualquier objeto móvil. Las redes sociales ofrecen también datos sobre los sitios visitados por los usuarios, e.g., muchas publicaciones, reseñas o comentarios publicados en Facebook, Foursquare y Twitter suelen ir acompañadas de datos de geolocalización (Cheng, Caverlee, Lee, & Sui, 2011). Estos datos pueden ser usados para encontrar relaciones de similitud entre los usuarios, para recomendar sitios o actividades, o para detectar comportamientos inusuales (e.g., una persona que acude a un cajero electrónico en un horario en el que usualmente no hace transacciones).

En particular, la similitud de las trayectorias permite agrupar individuos con comportamientos similares lo cual puede ser útil para recomendar nuevos amigos, productos o servicios (M.-J. Lee & Chung, 2011). La identificación de grupos de objetos móviles similares tiene aplicaciones en campos como mercadeo, turismo, entretenimiento, meteorología, tráfico, migración, monitoreo de animales, entre otros (Kondaveeti, 2012; Lee et al., 2007).

Las trayectorias se pueden agrupar por similitud según diferentes criterios (e.g., la forma, la distancia, los sitios visitados, entre otros). En este capítulo se analizan diversos trabajos relacionados con la similitud de trayectorias basada en criterios semánticos y temporales. Se analizan las características principales de cada método, sus ventajas y desventajas y

su contribución en este campo. A partir de estos resultados se propone algunas líneas de investigación.

### 1.1 Motivación

Varios trabajos se han enfocado en determinar la similitud entre trayectorias considerando criterios como la forma, la velocidad, la orientación geográfica, los sitios y los tipos de sitios que estas visitan (e.g., sitios educativos o de entretenimiento), entre otros. Unos pocos trabajos (Lee et al., 2007; Moreno, Borgony, & Román, 2017) han considerado también las actividades ejecutadas (e.g., estudiar, trabajar) en los sitios visitados; sin embargo, no han considerado aspectos relacionados con el tiempo de ejecución de las actividades, e.g., la duración, el orden y la frecuencia.

Ejemplo 1. Considérense tres usuarios A, B y C. La Figura 6 muestra las actividades ejecutadas por estos usuarios entre las 8 am y las 6 pm.

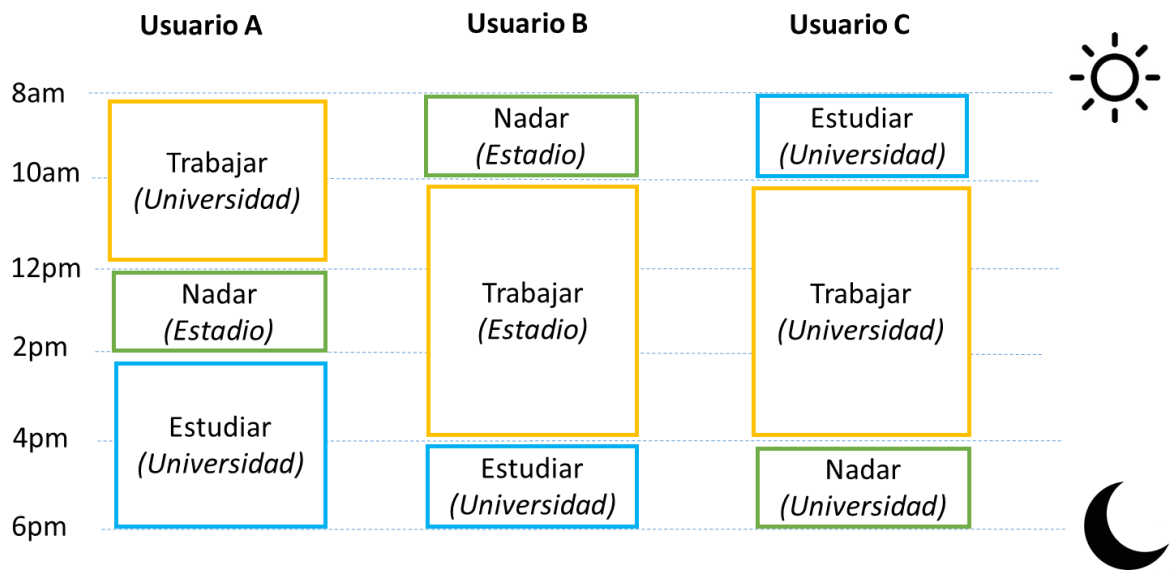


Figura 6. Actividades ejecutadas por tres usuarios A, B y C durante un día entre las 8 am y las 6 pm

Se observa que los tres usuarios trabajaron; no obstante, los usuarios A y C trabajan en una universidad, mientras que el usuario B en un estadio. Por lo tanto, si se considera el sitio donde se ejecuta esta actividad, A y C es la pareja de usuarios más similar. Sin

embargo, los usuarios B y C trabajan durante seis horas, mientras que el usuario A durante cuatro. Así, si se considera la duración de esta actividad, B y C es la pareja más similar.

Por otro lado, los tres usuarios nadaron durante dos horas; sin embargo, los usuarios A y C nadan en la tarde, mientras que el usuario B nada en la mañana. De esta forma, A y C es la pareja de usuarios más similar en cuanto al momento de ejecución de esta actividad. Sin embargo, si se considera el sitio de ejecución de esta actividad, A y B es la pareja más similar.

Considérese ahora el orden de ejecución de las actividades. Se observa que los tres usuarios trabajaron, nadaron y estudiaron. Sin embargo, los usuarios A y C trabajaron antes de nadar, a diferencia del usuario B que lo hizo después. Según este aspecto, A y C es la pareja más similar.

Ejemplo 2. Considérese ahora la Figura 7 que muestra las actividades “Mercar” y “Hacer deporte” ejecutadas por tres usuarios durante una semana.

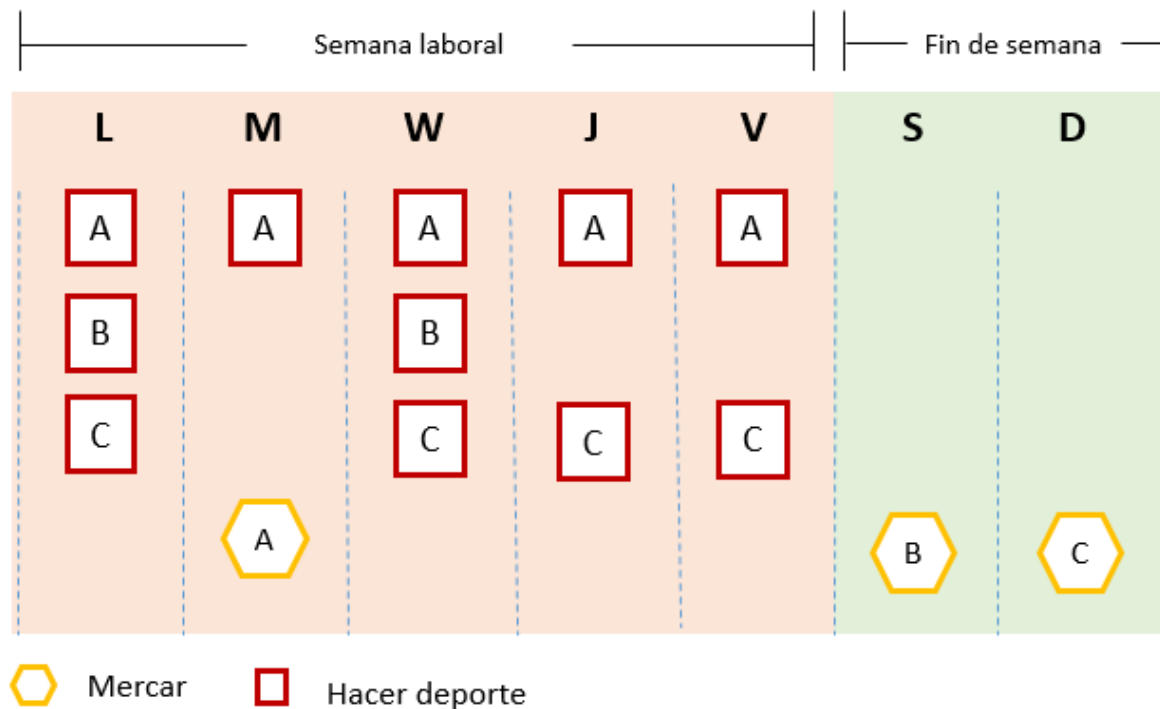


Figura 7. Actividades ejecutadas por tres usuarios A, B y C durante una semana

El usuario A hizo deporte cinco veces en la semana, el usuario B dos veces y el C cuatro veces. Si bien todos hicieron deporte, los usuarios A y C fueron los que ejecutaron la actividad más veces. Así, según la frecuencia de esta actividad, la pareja A y C es la más similar.

Por otro lado, nótese que todos los usuarios mercaron; sin embargo, B y C lo hicieron el fin de semana, mientras que A lo hizo el martes. Según el momento de la semana (con respecto a esta actividad), B y C es la pareja más similar.

La Tabla 1 relaciona las parejas de usuarios más similares según un criterio seleccionado. Por ejemplo, los criterios 1 y 2 consideran la similitud con respecto a los sitios visitados o a las actividades ejecutadas; de esta forma, los tres usuarios son igualmente similares ya que según el criterio 1 todos visitaron los mismos sitios y según el criterio 2 todos ejecutaron las mismas actividades. Así, si un usuario X y un usuario Y visitan una universidad, el criterio 1 es suficiente para decir que son similares, sin importar la actividad que ejecutan en la universidad.

Tabla 1. Parejas más similares de acuerdo con los criterios seleccionados

Nro.	Criterio	Actividad o Sitio	Usuarios más similares	Observación	Ejemplo motivador
1	Sitios visitados		A, B y C	Todos los usuarios visitaron los mismos sitios	1
2	Actividades ejecutadas		A, B y C	Todos los usuarios ejecutaron las mismas actividades	1
3	Sitios de ejecución de las actividades	Trabajar	A y C	A y C trabajan en una Universidad. B en un Estadio	1
		Nadar	A y B	A y B nadan en un Estadio. B en una Universidad	1
		Estudiar	A, B y C	Los tres usuarios estudian en una Universidad	1
4	Actividades ejecutadas en el sitio	Universidad	A y C	A y C trabajan y estudian en una Universidad. El usuario B solo estudia	1

		Estadio	A y B	C no ejecuta actividades en un Estadio	1
5	Duración de las actividades	Trabajar	B y C	B y C ejecutan la actividad durante 6 horas. A durante 4 horas	1
		Nadar	A, B y C	Los tres usuarios ejecutan la actividad durante 2 horas	1
		Estudiar	B y C	B y C ejecutan la actividad durante 2 horas. A durante 4 horas	1
6	Momento de ejecución de las actividades (parte del día)	Trabajar	B y C	B y C trabajan en la mañana y en la tarde. A trabaja solo en la mañana	1
		Nadar	A y C	A y C nadan en la tarde. B en la mañana	1
		Estudiar	A y B	A y B estudian en la tarde. C estudia en la mañana	1
7	Orden de ejecución de las actividades	Trabajar vs Nadar	A y C	A y C nadan después de trabajar. B nada antes de trabajar	1
		Trabajar vs Estudiar	A y B	A y B estudian después de trabajar. C estudia antes de trabajar	1
		Nadar vs Estudiar	A y B	A y B estudian después de nadar. C estudia antes de nadar	1
8	Frecuencia de las actividades	Hacer deporte	A y C	A hizo deporte cinco veces en la semana, C lo hizo cuatro veces y B dos veces	2
		Mercar	A, B y C	Los tres usuarios mercaron 1 vez en la semana	2
9	Momento de ejecución de las actividades (parte de la semana)	Hacer deporte	A, B y C	Los tres usuarios hacen deporte en días laborales (lunes a viernes)	2
		Mercar	B y C	B y C mercan los fines de semana. A merca el martes	2

Para determinar la similitud con mayor detalle se puede recurrir a una combinación de los dos criterios. De esta forma, los criterios 3 y 4 de la Tabla 1 muestran, dependiendo del sitio o actividad escogida, los usuarios más similares (pero en estos casos importa que actividad ejecutan los usuarios en un sitio y que sitio visitan para ejecutar una actividad). Por ejemplo, si X trabaja en una universidad y Y estudia en una universidad, su similitud se verá reducida por la diferencia en las actividades que ejecutan.

Los criterios relacionados con el tiempo pueden ser considerados con respecto a la actividad ejecutada o al sitio visitado. En la Tabla 1 los criterios del 5 al 9 se consideraron con respecto a las actividades, pero pueden ser considerados de forma análoga con respecto a los sitios. Por ejemplo, el criterio 5 considera la duración de las actividades (sin importar el sitio donde se ejecutan). Análogamente, aunque no está en la Tabla 1, también se puede considerar la duración de la estadía en un sitio (sin importar la actividad ejecutada). Nótese además que se puede considerar una combinación de los tres criterios donde importa el aspecto temporal, el sitio y la actividad. Por ejemplo, si X trabaja en una universidad durante 8 horas y Y estudia en una universidad durante 2 horas, su similitud se verá afectada por la diferencia en sus actividades y en la duración.

La Figura 8 esquematiza de manera gradual la combinación de 1, 2 o 3 criterios para establecer la similitud.

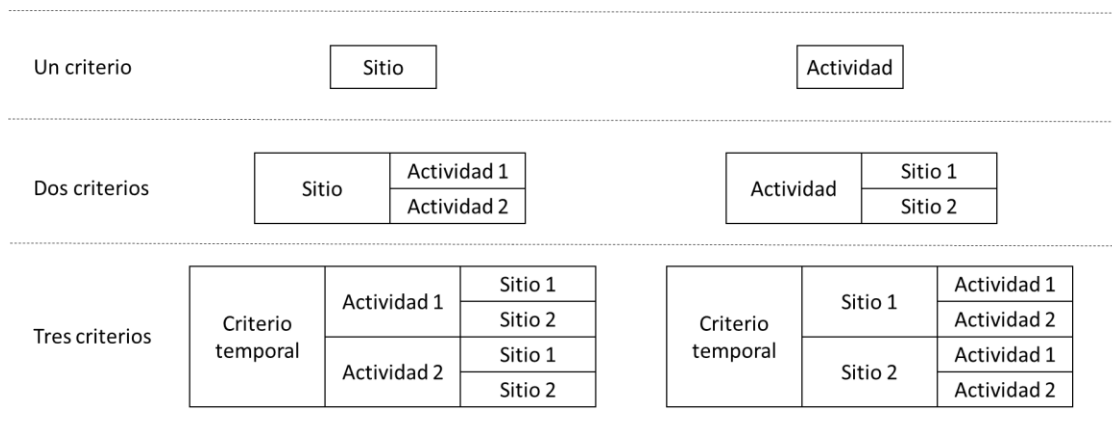


Figura 8. Ejemplos de combinaciones de algunos criterios.

Hasta el momento se ha considerado una combinación de máximo tres criterios; sin embargo, con el fin de establecer la similitud con mayor detalle se podrían considerar aún más. Estos dos ejemplos motivadores, muestran que el establecimiento de la similitud no

es una tarea trivial: depende en gran medida de los intereses del analista y de los criterios con que desea establecerla.

## 1.2 Trabajos relacionados

En (Z. Li, Kays, & Nye, 2010) se propone un algoritmo de dos fases llamado *Periodica*, el cual recibe un conjunto de trayectorias: en la primera fase encuentra sitios de referencia (*reference spots*) y detecta periodos (e.g., de días, de semanas) de visita para cada uno; un *reference spot* es un área frecuentemente visitada por las trayectorias. En la segunda fase trata de descubrir comportamientos habituales (e.g., un usuario llega a la casa de lunes a viernes alrededor de las 8 pm, y los sábados alrededor de las 11 pm) y con base en estos hace agrupar las trayectorias

En (Liu & Schneider, 2012) se define una medida de similitud entre trayectorias determinada por aspectos geográficos y semánticos. La distancia geográfica está determinada por criterios como la dirección, el centro de masa y la velocidad de las trayectorias. La distancia semántica considera la secuencia de sitios visitados por cada trayectoria y calcula la máxima subsecuencia de sitios en común, i.e., el conjunto más grande de sitios comunes visitados por las dos trayectorias en el mismo orden. Sin embargo, no se consideran las actividades ejecutadas por los objetos móviles ni los tipos de sitios visitados.

En (Lee & Chung, 2011) se define un modelo de similitud de trayectorias que considera el tipo de sitio visitado (e.g., de Entretenimiento, de Comida, Cultural). Se define además una jerarquía de categorías (tipos) de sitios, e.g., el tipo de sitio "Cultural" es más general que el tipo de sitio "Museo"; y se seleccionan los sitios más frecuentados por los usuarios (i.e., no considera los sitios visitados esporádicamente por los usuarios). La similitud se ve determinada por el número de sitios coincidentes y sus categorías, e.g., dos usuarios que visitan museos diferentes son menos similares que dos usuarios que visitan el mismo museo, pero son más similares que dos usuarios que visitan uno un museo y el otro un gimnasio. Este método permite establecer relaciones semánticas más detalladas con respecto a los sitios; sin embargo, no se consideran las actividades ejecutadas en ellos ni aspectos temporales.

En (Zhao, 2011) se propone un algoritmo para el clustering de trayectorias basados en tres fases. En la primera fase se crea una matriz lógica (*boolean*) donde cada trayectoria es una fila y cada columna es un *stop* (i.e., un sitio visitado por un objeto móvil), y apoyado en el índice de Jaccard (mide la similitud entre dos conjuntos) agrupa las trayectorias que tienen sitios en común. La segunda fase usa el *Dynamic Time Warping* (DTW, mide la similitud entre dos series de tiempo) para agrupar las trayectorias que comparten secuencias cronológicas (i.e., el orden en que se hacen los *stops*) similares. Finalmente, en la tercera fase se agrega una condición de tiempo que permite agrupar las trayectorias que tienen secuencias cronológicas en común con tiempos similares.

En (Ying, Lu, Lee, Weng, & Tseng, 2010) se propone una medida de similitud denominada *MSTP-Similarity (Semantic Trajectory Pattern Similarity)* que calcula la similitud entre dos trayectorias de acuerdo con etiquetas semánticas de los sitios (e.g., Escuela, Universidad, Hospital) y a la secuencia de sitios visitados. El método se basa en una modificación del algoritmo de la máxima subsecuencia común más larga (LCS, *Longest Common Subsequence* (Bergroth, Hakonen, & Raita, 2000)). El algoritmo recibe dos trayectorias (con los sitios visitados etiquetados), e.g., la máxima subsecuencia común más larga entre las trayectorias  $P = \{\text{Parque, Cine, Teatro, Gimnasio, Universidad}\}$  y  $Q = \{\text{Parque, Teatro, Escuela, Universidad}\}$  es  $LCS = \{\text{Parque, Teatro, Universidad}\}$ . El método propuesto considera etiquetas semánticas para los sitios, pero no una relación entre los mismos, e.g., no es posible relacionar los sitios Escuela y Universidad como sitios de tipo Educativo.

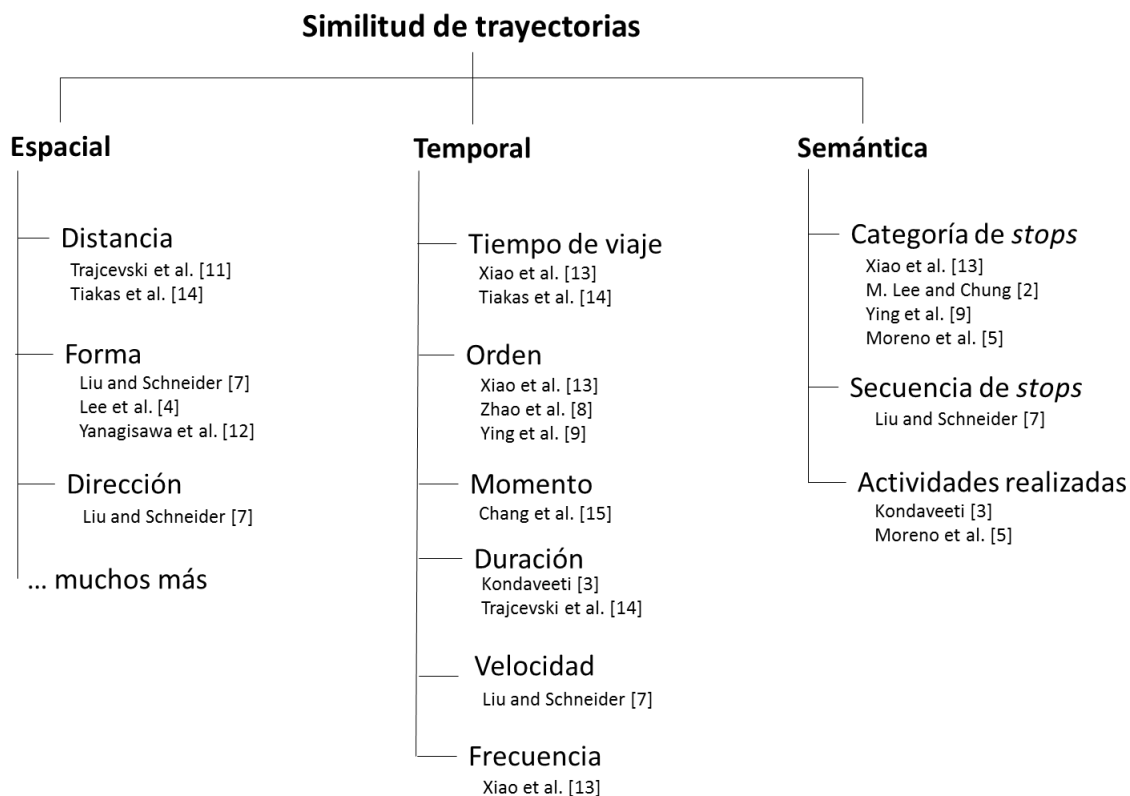


Figura 9. Taxonomía de clasificación de criterios

La Figura 9 muestra una taxonomía de diferentes criterios que se pueden aplicar para determinar la similitud de trayectorias y algunos trabajos que consideran dichos criterios. Se hace la clasificación a partir de tres criterios: espacial, temporal y semántica. Los trabajos cuyos métodos abarcan más de un tipo de similitud aparecen en más de una rama. Debido a que la similitud espacial ha sido ampliamente estudiada (Trajcevski, Ding, Scheuermann, Tamassia, & Vaccaro, 2007; Yanagisawa et al., 2003) en este trabajo se analizan las propuestas que consideran criterios temporales y semánticos.

En la siguiente sección se analizarán en detalle cuatro de los trabajos más representativos, a partir de los aspectos no contemplados por estas propuestas, se proponen algunas ideas que pueden ser la base para futuros trabajos de investigación en este campo: a) El trabajo de Xiao et al (2012) es el más maduro en cuanto al manejo de la semántica de los sitios, ya que considera un árbol de categoría de sitios que permite relacionar los sitios visitados por los usuarios según su tipo (e.g., hostel y hotel, bar y discoteca, gimnasio y piscina), b) el trabajo de Tiakas et al. (2009) considera tanto aspectos espaciales como temporales,

los cuales se tratan de forma independiente y se combinan al final según la importancia definida por el analista, brindando adaptabilidad a diferentes aplicaciones e intereses, c) por otra parte, Chang et al. (2007) son los únicos que tratan el aspecto momento (i.e., cuando ocurren los eventos), fraccionando los días, semanas y años en diferentes secciones, con el objetivo de relacionar los eventos que ocurren en momentos similares, y por último, d) Kondaveeti (2012) es el que más aspectos diferentes considera (sitios, actividades, orden, duración, entre otros), pero que también evidencia una serie de limitaciones a partir de las cuales se pueden plantear diferentes líneas de investigación.

### 1.3 Trabajo 1: Xiao et al.

Xiao et al. (2012) proponen un método para calcular la similitud entre usuarios de acuerdo con la historia de sus localizaciones. Cada punto de interés (POI) tiene asociado una categoría que le confiere una categoría semántica al sitio, e.g., “Restaurante”, “Parque”, “Hotel”. Se define un conjunto de capas, donde las capas de menor granularidad agrupan a los sitios con semántica similar (e.g., “Restaurante”) y las capas de mayor granularidad diferencian a los sitios de una forma más específica (e.g., “Restaurante de Comida China”, “Restaurante de Comida Árabe”); esto permite que el cálculo de la similitud entre usuarios sea flexible de acuerdo con el nivel deseado por el analista.

Se define un árbol de jerarquía de categorías que representa la relación entre las diferentes categorías asignadas a los sitios y donde cada nivel corresponde a una capa, ver la Figura 10. Nótese que al descender por los niveles del árbol (se aumenta la granularidad) los sitios son asociados con categorías más específicas, e.g., los sitios 1 y 2 podrían estar agrupados con la categoría “Restaurante” en la capa 2, pero ser agrupados con categorías diferentes (de comida china y de comida árabe) en la capa 3.

La similitud temporal viene determinada por el orden en que se visitan los sitios y el tiempo de viaje. Un *travel match* es una secuencia común de localizaciones semánticas visitadas por dos usuarios en periodos de viaje similares. Se hallan las subsecuencias en común entre las secuencias de dos usuarios y se determinan los *travel match* entre las localizaciones de ambas secuencias, con el fin de detectar secuencias similares de sitios visitados en tiempo similares. El método para encontrar las subsecuencias en común no exige una continuidad en los sitios visitados (i.e., permite “huecos” en las secuencias), e.g., un usuario que visitó los sitios Universidad -> Restaurante será similar a otro usuario que

hizo una parada (*stop*) en un centro comercial antes de visitar un restaurante, i.e., que tenga la secuencia Universidad -> Centro comercial -> Restaurante, pero solo si la estancia en el centro comercial es de corta duración, ya que si las diferencias de tiempo son muy grandes ya no serán consideradas subsecuencias en común. La similitud está determinada por las secuencias compartidas entre usuarios (i.e., los usuarios que compartan más secuencias en común más largas serán más similares), la granularidad (i.e., los usuarios que compartan secuencias en común a un nivel más fino de granularidad serán más similares) y la popularidad de la localización (i.e., los usuarios que compartan sitios poco frecuentados por los demás usuarios serán más similares). La combinación de estos tres factores determina la similitud entre dos usuarios.

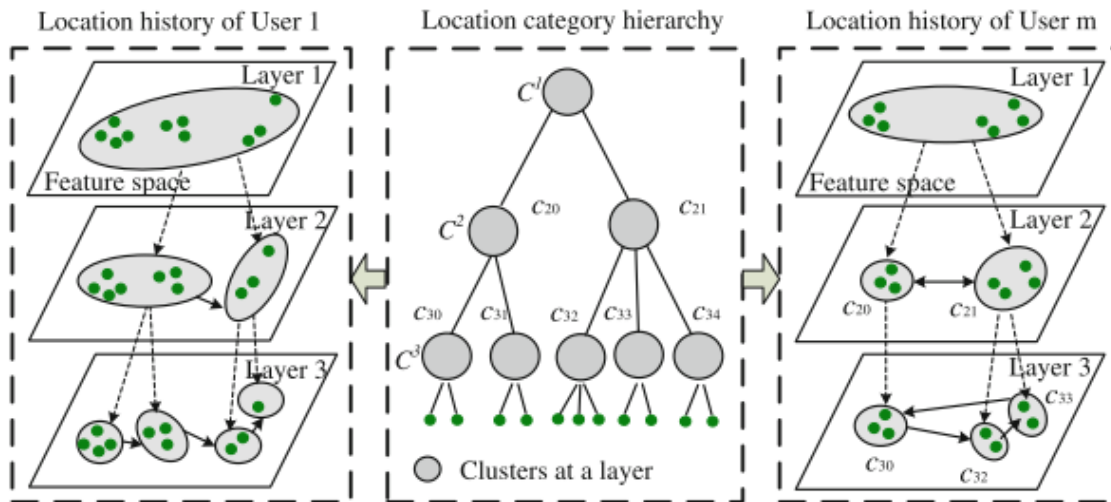


Figura 10. Relación entre el árbol de categorías y las diferentes capas. Fuente (Xiao et al., 2012)

Si bien la semántica de los sitios visitados está bien definida y es flexible, el método no considera las actividades ejecutadas en cada sitio, lo que trae consigo la posibilidad de ignorar situaciones importantes que afectan de manera significativa la similitud entre dos usuarios, e.g., no es cauto concluir que dos usuarios que visitan un centro comercial son muy similares cuando uno lo hace para trabajar y otro para realizar compras.

## 1.4 Trabajo 2: Tiakas et al.

Tiakas et al. (2009) definen una medida de similitud de trayectorias en redes espaciales que considera información espacial y temporal. El método se propone para objetos móviles en una red espacial, i.e., los objetos están sujetos a las restricciones de la red, e.g., los automóviles en una ciudad solo se pueden desplazar a través de las carreteras y según las restricciones de sentido de las vías y de los giros permitidos.

Nótese que la distancia euclidiana no puede ser usada para establecer la distancia espacial debido a las restricciones de movimiento en la red, e.g., la Figura 11 muestra una situación donde el nodo más cercano al nodo A es el nodo B si se considera la distancia euclidiana, pero debido a las restricciones de la red, para llegar al nodo A se recorre una distancia menor desde el nodo C que desde el nodo B. Debido a esta particularidad, los autores definen un método basado en el costo de viaje del camino más corto entre dos nodos del grafo. Como alternativa a este método, definen también otra medida de similitud que usa la distancia euclidiana pero que de igual forma incluye el concepto de costo entre dos nodos.

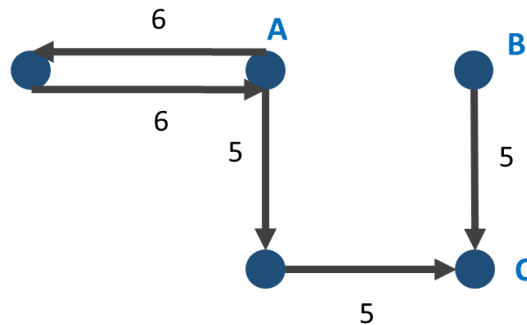


Figura 11. Ejemplo de una red espacial

La medida de similitud con respecto al tiempo considera el tiempo requerido de viaje entre un nodo y otro; y que puede ser combinada con una medida de similitud espacial. La medida de similitud temporal entre dos trayectorias  $T_a$  y  $T_b$  está dada por la Ecuación 1, que considera los tiempos de viaje entre nodos.

$$D_{time}(T_a, T_b) = \frac{1}{m-1} \sum_{i=1}^{m-1} \frac{|(T_a[i+1].t - T_a[i].t) - (T_b[i+1].t - T_b[i].t)|}{\max\{(T_a[i+1].t - T_a[i].t), (T_b[i+1].t - T_b[i].t)\}}$$

Ecuación 1

Donde  $m$  es la longitud de las trayectorias, y  $T[i].t$  es la marca de tiempo del nodo  $i$  de la trayectoria  $T$ .

La Figura 12 muestra un ejemplo de dos trayectorias con cuatro nodos cada una (un nodo es un sitio visitado); los valores sobre las aristas representan el tiempo de viaje entre dos nodos. Al aplicar la Ecuación 1 se obtiene  $D_{time}(T_a, T_b) = \frac{1}{3} \left( \frac{1}{5} + \frac{6}{9} + \frac{0}{8} \right) = 0,288$

Finalmente, para encontrar trayectorias similares considerando tanto los criterios espaciales como temporales, se proponen dos métodos:

1. Definir una distancia espaciotemporal, dada por la Ecuación 2

$$D_{total}(T_a, T_b) = W_{espacial} * D_{espacial}(T_a, T_b) + W_{temporal} * D_{temporal}(T_a, T_b)$$

Ecuación 2

Donde  $W_{espacial}$  y  $W_{temporal}$  son pesos asignados a cada criterio, tal que  $W_{espacial} + W_{temporal} = 1$

2. Hallar trayectorias similares en cuanto a distancia espacial y luego establecer su similitud con respecto a la distancia temporal.

Nótese que este método exige que las dos trayectorias tengan el mismo número de episodios (nodos), lo cual en la práctica es poco común, e.g., dos usuarios no necesariamente visitan el mismo número de sitios en un día, o dos camiones de reparto no hacen el mismo número de paradas para entregar mercancía en una ruta. Para resolver este problema, los autores proponen descomponer cada trayectoria en un conjunto de sub-trayectorias de igual longitud y luego aplicar los métodos descritos.

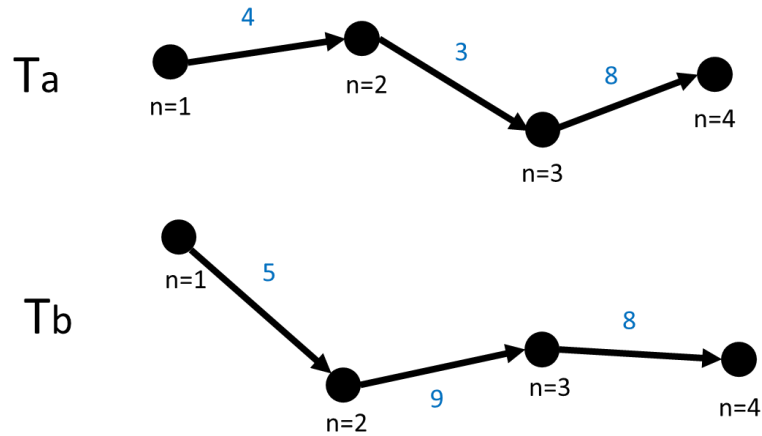


Figura 12. Ejemplo de trayectorias con 4 nodos

### 1.5 Trabajo 3: Chang et al.

Chang et al. (2007) proponen una medida para calcular la similitud entre dos trayectorias en una red de carreteras. El método calcula una distancia espacial y una distancia temporal entre dos trayectorias y las combina para obtener así un valor final de similitud. La distancia espacial es no euclidiana y la distancia temporal está compuesta por tres factores: el rango de tiempo del día (R), el día de la semana (D) y la semana (W).

La distancia temporal (TD) entre dos trayectorias Q y T está dada por la Ecuación 3.

$$TD(Q, T) = \alpha * TRD(Q, T) + \beta * TDD(Q, T) + \gamma * TWD(Q, T)$$

Ecuación 3

Donde TRD (*trajectory range distance*) es el factor rango de tiempo del día, TDD (*trajectory day distance*) es el factor día de la semana, TWD (*trajectory week distance*) es el factor semana y  $\alpha$ ,  $\beta$  y  $\gamma$  son pesos, cuyos valores adecuados son 60, 20 y 1 respectivamente según los experimentos hechos por los autores.

*TRD*: Para definir este factor el analista fragmenta el día en diferentes rangos de tiempo (R), véase un ejemplo en la Tabla 2. A su vez, cada rango de tiempo se asocia con grupos de rangos de tiempo (definidos también por el analista) como se muestra en la Tabla 3.

Tabla 2. Rangos de tiempo de un día. Fuente (Chang et al., 2007)

S.No.	Range(R)	Description	From	To
1.	MR	Morning Rush Hour	7:00 AM	9:00 AM
2.	M	Morning	9:00 AM	12:00 PM
3.	L	Lunch Time	12:00 PM	2:00 PM
4.	A	Afternoon	2:00 PM	5:00 PM
5.	ER	Evening Rush Hour	5:00 PM	7:00 PM
6.	LE	Late Evening	7:00 PM	10:00 PM
7.	N	Night (Sleeping Time)	10:00 PM	7:00 AM

Tabla 3. Grupos de rangos de tiempo. Fuente (Chang et al., 2007)

Group	Member/s
Group 1	MR, ER
Group 2	L
Group 3	M, A, LE
Group 4	N

Por ejemplo, el grupo de rango de tiempo 1 relaciona los momentos del día en que se presenta alta congestión vehicular, mientras que el grupo de rango de tiempo 4 relaciona los momentos del día para dormir.

*TDD*: Para definir este factor, se clasifican los días de la semana, e.g., en dos grupos: Grupo 1: de lunes a viernes y Grupo 2: sábado y domingo. Esto con el fin de separar el comportamiento de los días laborales de los días de fin de semana. La función propuesta por los autores considera la distancia entre días del mismo grupo y entre días de grupos diferentes.

*TWD*: Para definir este factor se considera la diferencia en valor absoluto entre semanas entre dos trayectorias, donde un año contiene las semanas de la 1 a la 52. Por ejemplo, si Q empieza en la semana 3 del año y T empieza en la semana 7 del año, la TWD de Q y T es 4.

Para calcular la distancia espacial, los autores proponen un método basado en distancia no euclidiana que calcula la distancia espacial entre dos bordes de cada trayectoria Q y T como el promedio de la distancia entre sus nodos iniciales y la distancia entre sus nodos finales. Finalmente, la similitud espacial está dada por el promedio de la distancia espacial entre cada par de bordes de las trayectorias.

La distancia espaciotemporal entre dos trayectorias está dada por la Ecuación 4.

$$STDist(Q, T) = \frac{SD + \delta * TD}{2}$$

Ecuación 4

Donde  $SD$  es la similitud espacial,  $TD$  es la similitud temporal, y  $\delta$  es un peso espaciotemporal cuyo valor adecuado es 20 según los experimentos de los autores.

Si bien el método permite que el analista defina los rangos de horas y de días, solo considera el tiempo de inicio y de fin de cada trayectoria “completa”, i.e., hace caso omiso al momento y duración de cada episodio (*moves*, *stops*). Esto se debe a que el método está propuesto para trabajar sobre rutas de objetos móviles en una red de carreteras más no sobre una trayectoria definida por *stops* y *moves*. Así, es necesario modificar el método si se desea considerar el aspecto temporal de sus episodios.

## 1.6 Trabajo 4: Kondaveeti

Para Kondaveeti (2012) una trayectoria es un conjunto de episodios: *stops* y *moves*, donde cada episodio tiene un conjunto de atributos. La Figura 13 muestra un ejemplo de dos trayectorias que se consideran similares. Nótese que, aunque sus tiempos de inicio no coinciden, el tiempo total de viaje, el modo de transporte, los sitios visitados y la secuencia (orden) de *stops* y *moves* son iguales.

El autor propone un método para agrupar (*clustering*) las trayectorias similares. En la primera fase se crea un conjunto de trayectorias sintéticas en forma aleatoria a partir de los datos reales con el fin de romper dependencias entre las variables, e.g., cuando se hace un *stop* en un mismo sitio la duración suele ser similar (e.g., el tiempo de tanqueo de un automóvil en una gasolinera). Los datos reales se etiquetan como C0 (clase 0) y los datos sintéticos como C1 (clase 1).

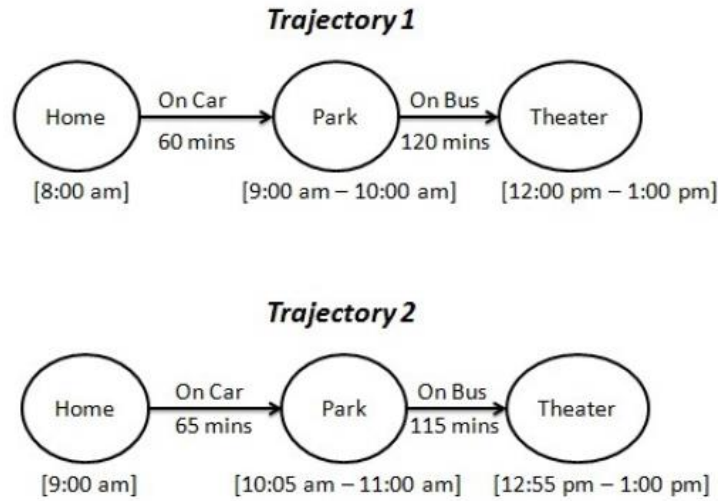


Figura 13. Ejemplo de dos trayectorias similares. Fuente (Kondaveeti, 2012)

Para generar los datos de la clase C1 se procede así: para los atributos numéricos se genera un valor aleatorio en el intervalo  $[mina - xa, maxa + xa]$  donde  $mina$ ,  $maxa$  y  $xa$  son el valor mínimo, el valor máximo y la desviación estándar respectivamente del atributo  $a$  en la clase C0; y para los atributos categóricos se escoge aleatoriamente un valor entre el conjunto de posibles valores que puede tomar el atributo en la clase C0. La Figura 14 muestra un ejemplo de la generación de la clase C1 a partir de los datos de la clase C0.

Traj ID	$s_{11}$	$s_{12}$	$m_{11}$	$s_{21}$	$s_{22}$	.	Class
1	500	Loading	70	Quality Check	80	.	0
2	90	Unloading	100	Loading	75	.	0
3	850	Loading	83	Distribution		.	0
.	.	.	.	.	.	.	.
100	.	.	.	.	.	.	.
101	98	Loading	85	Quality Check	.	.	1
102	812	Unloading	78	Distribution	.	.	1
103	534	Loading	97	Distribution	.	.	1
.	.	.	.	.	.	.	.
200	.	.	.	.	.	.	.

Random value between  
 $90 - \sigma$  and  $850 + \sigma$

Synthetic Class

Figura 14. Ejemplo de creación de clase sintética. Fuente (Kondaveeti, 2012)

Después de generar la clase C1 se usa el algoritmo de *Random Forest* (Liaw & Wiener, 2002), el cual genera una cantidad de L árboles de decisión para determinar la clasificación de cada trayectoria y generar los *clusters*.

Con el fin de analizar el método propuesto, se ejecutó el programa en lenguaje R presentado en (Kondaveeti, 2012) con 50 trayectorias, cada una conformada por tres *stops* y dos *moves*, donde cada *stop* tiene un atributo correspondiente al tiempo de permanencia en el sitio y cada *move* tiene un atributo correspondiente al tiempo de desplazamiento. A continuación, se presentan algunas desventajas del método.

	A	B	C	D	E	F	G	H	I	J	K
1	S11	S12	M11	M12	S21	S22	M21	M22	S31	S32	Class
2	10	Home	60	Car	50	Park	120	Bus	50	Theater	0
3	120	Theater	233	Bus	76	Park	179	Train	460	University	0
4	50	Park	71	Car	120	Theater	156	Bus	430	Home	0
5	450	School	249	Bus	54	Park	74	Car	56	Theater	0
6	20	Home	51	Train	105	Theater	206	Train	210	University	0
7	50	Park	32	Train	45	Home	74	Car	140	Theater	0
8	130	Theater	181	Bus	46	Park	169	Bus	240	University	0
9	25	Home	125	Car	460	University	161	Bus	30	Park	0
10	10	Home	60	Car	50	Park	120	Bus	50	Theater	0
11	22	Home	32	Train	43	Park	62	Car	560	Home	0

Figura 15. Trayectorias 1 y 9

1. *Similitud no consistente*: Por ejemplo, las trayectorias 1 y 9 son idénticas, véase la Figura 15. Por lo tanto, se espera que la similitud entre ellas sea 1 (máxima similitud), como en efecto ocurre. Se espera también que la similitud de cada una de ellas con cualquier otra trayectoria sea la misma; sin embargo, los valores de la trayectoria 1 y de la 9 con respecto a las demás trayectorias son diferentes, e.g., la similitud de la trayectoria 6 con la trayectoria 1 es de 0.8064 mientras que con la trayectoria 9 es de 0.6428, lo que evidencia una diferencia de más del 16%.
2. *Falta de etiquetas semánticas (para sitios, actividades, etc.)*: Las trayectorias 16, 17 y 18 son “casi” iguales, la única diferencia es que el sitio visitado en el stop 1 de la trayectoria 16 es una universidad (*University*), el de la trayectoria 17 es una escuela (*School*) y el de la trayectoria 18 es una casa (*Home*). Nótese que *University* y *School* son sitios de tipo educativo, mientras que *Home* no aplica en dicho tipo.

Si se consideran los tipos de los sitios, la similitud de una trayectoria T con las trayectorias 16 y 17 debería ser, en términos prácticos igual, ya que estas dos trayectorias solo difieren en un sitio, aunque del mismo tipo (educativo). Por otro lado, la similitud de T con la trayectoria 18 debería ser menor (con respecto a la 16 y a la 17) ya que el sitio que diferencia a la 18 con respecto a la 16 y a la 17 no es de tipo educativo (*Home*). Sin embargo, en los resultados esto no se cumple, e.g., para  $T = 5$ , la similitud de las trayectorias 5 y 16 es 0.5128, la similitud de las trayectorias 5 y 17 es 0.4901 y la similitud de las trayectorias 5 y 18 es 0.5769.

Una situación similar ocurriría si se consideran otros tipos de atributos propios de los episodios como las actividades. En este método no se asigna significado (semántica) a los valores de los atributos. De esta forma, actividades como trotar, correr, bailar y cantar tienen la misma connotación, y sitios como universidad y escuela no tienen relación alguna.

3. *Sensibilidad en el orden de los datos (atributos)*: Se intercambiaron las columnas que representan el sitio visitado y la duración del stop 3 y se volvió a ejecutar el método. El intercambio en el orden de los atributos de un stop no debería afectar la similitud entre las trayectorias; sin embargo, se obtuvo una matriz de proximidad diferente; e.g., antes del intercambio la similitud entre las trayectorias 3 y 8 es 0,4383 y luego del intercambio su similitud es 0,5735, i.e., una diferencia mayor que 13%. Esta situación se presenta para casi cualquier par de trayectorias. Esto indica que el método es sensible a la posición de los atributos. En este ejemplo, esto debería ser irrelevante ya que independientemente del orden de los atributos del stop, el sitio visitado y la duración no cambian.
4. *Dependencia de muestras de datos*: no es posible calcular la similitud entre dos trayectorias si no se tiene datos de otras trayectorias. Es decir, el método requiere datos de trayectorias adicionales incluso cuando solo interesa encontrar la similitud entre dos trayectorias.

## 1.7 Análisis

La Tabla 4 muestra los cuatro trabajos analizados en detalle y las desventajas detectadas en cada uno de ellos.

Tabla 4. Falencias encontradas en los trabajos analizados

Trabajo	(Xiao et al., 2012)	(Tiakas et al., 2009)	(Chang et al., 2007)	(Kondaveeti, 2012)
<b>Aspecto para establecer la similitud</b>	Temporal, semántico	Espacial, temporal	Temporal	Semántico
<b>Desventajas encontradas</b>	No considera actividades	Orientado a redes espaciales, lo que restringe su aplicación en trayectorias de objetos móviles con movimiento libre en el espacio.	No considera el momento ni la duración de cada episodio.	- Similitud no consistente - Falta de etiquetas semánticas (para sitios, actividades, etc.) - Sensibilidad en el orden de los datos (atributos) - Dependencia de muestras de datos

Algunas líneas de investigación que surgen del análisis hecho son:

- Las actividades ejecutadas por los objetos móviles en los sitios es un aspecto que se ha considerado poco en los métodos actuales. El planteamiento de un método que considere tanto los sitios visitados y las actividades ejecutadas por los usuarios en estos, ofrecería más elementos a los analistas para establecer la similitud entre los usuarios. Se puede considerar, e.g., la relación entre diferentes tipos de actividades, e.g., nadar y trotar son actividades similares ya que son de carácter deportivo. También se puede considerar la ejecución de diferentes actividades en un mismo sitio, e.g., un usuario puede visitar una universidad durante 5 horas y haber ejecutado tres actividades: estudió 2 horas, nadó una hora y trabajó dos horas. Se pueden considerar también detalles de cada actividad, e.g., el libro que leyó, el estilo de natación practicado, el género de música escuchado, entre otros.

- Aunque varios autores han considerado aspectos relacionados con el tiempo, estos se han trabajado de forma independiente. Es deseable un método que incluya diversos aspectos temporales como la duración, la cronología de los hechos, la frecuencia de los sitios visitados, el momento del día y los días de la semana en que se visitan los sitios; ya que como se presentó en la sección de motivación, cada uno de estos aspectos es relevante para establecer la similitud entre usuarios.
- El estado de ánimo y los sentimientos de los usuarios podrían ser incluidos como una característica adicional en el cálculo de la similitud. Recientemente en redes sociales como Facebook se ha implementado la opción de incluir un estado de ánimo en las publicaciones hechas por los usuarios, las cuales pueden estar acompañadas de la localización y la actividad ejecutada por el usuario, e.g., un usuario puede publicar “Juan está comiendo en el Restaurante Ribs y está feliz” o “María está viendo Harry Potter en Cinemas La 30 y se siente decepcionada”. Los sentimientos y estados de ánimo pueden establecer la diferencia entre dos usuarios ya que pueden expresar el gusto o disgusto por cierto sitio o actividad, e.g., dos usuarios que visiten la playa y gusten de ella son más similares que un tercer usuario que la visite, pero la encuentre desagradable. También se podría aplicar análisis de sentimientos (Pozzi et al., 2017) para tratar de inferir el estado de ánimo de las personas a través de sus publicaciones en redes sociales donde no lo indiquen directamente.
- Las personas con las que se hizo una actividad o se visitó un sitio pueden indicar una mayor similitud entre usuarios. Por ejemplo, en Facebook se pueden etiquetar amigos en las publicaciones: “Juan está en Playa Ipanema con Andrés y Claudia” o “Julio está tomando unos tragos con Roger”; y en Instagram se pueden etiquetar otros usuarios en las fotos publicadas. La información obtenida de la red de amigos de los usuarios puede ser usada como un criterio para establecer la similitud entre trayectorias, e.g., si Andrés fue a teatro con Juan y Andrés también fue a teatro con María (otro día), se podría decir que Juan y María tienen algún grado de similitud no solo porque visitan el mismo tipo de sitio sino porque lo hacen con un amigo en común (Andrés). Este amigo en común puede llegar a ser un elemento clave para

la detección incluso de similitudes ficticias, e.g., los datos históricos podrían mostrar que María solo asiste a teatro con Andrés, mientras que Juan suele ir a teatro solo, con Andrés y también con otros amigos. Esto podría sugerir que a Juan realmente le gusta el teatro, mientras que María quizás solo asiste a teatro por acompañar a Andrés. Esto muestra que incluso si dos usuarios visitan los mismos sitios, se requieren más elementos para determinar con mayor precisión su similitud.



## **2. Similitud de trayectorias semánticas basada en sitios y actividades**

Recientemente un gran esfuerzo se ha hecho para agregar más información a las trayectorias crudas, i.e., transformar una trayectoria cruda en una semántica (Alvares et al., 2007; Parent et al., 2013). Una trayectoria semántica tiene más información que una trayectoria cruda. Además del espacio y el tiempo, una trayectoria semántica tiene información como el nombre y tipo de sitios visitados por un objeto móvil, y las actividades ejecutadas en cada sitio (Bogorny et al., 2014). En la Figura 16, se muestra un ejemplo de dos trayectorias semánticas, considerando tanto el tipo de sitios visitados como las actividades ejecutadas allí. La trayectoria A visita el Hotel X, el Banco K y la Universidad Y, mientras que la trayectoria B visita el Hostal Z, la Escuela U, el Banco K, la Universidad Y y el Restaurante W. La trayectoria A visita un hotel, mientras que la trayectoria B visita un Hostal, los cuales son diferentes sitios, pero del mismo tipo semántico, i.e., alojamiento. Nótese que la trayectoria A visita el hotel para trabajar, mientras que la trayectoria B visita el hostal como cliente, así que, si bien ambos sitios están relacionados con alojamiento, los objetos móviles ejecutaron diferentes actividades. Ambas trayectorias también visitaron sitios educativos, una universidad y una escuela, pero con la misma actividad: enseñar. Considérese que ambas trayectorias visitan tipos de sitios similares, pero pueden ejecutar diferentes actividades allí, así que la pregunta que surge es: ¿Qué tan similares son las trayectorias A y B desde un punto de vista semántico? ¿Qué tan similares son ambas trayectorias considerando los sitios visitados? ¿Considerando las actividades? ¿Considerando tanto sitios como actividades?

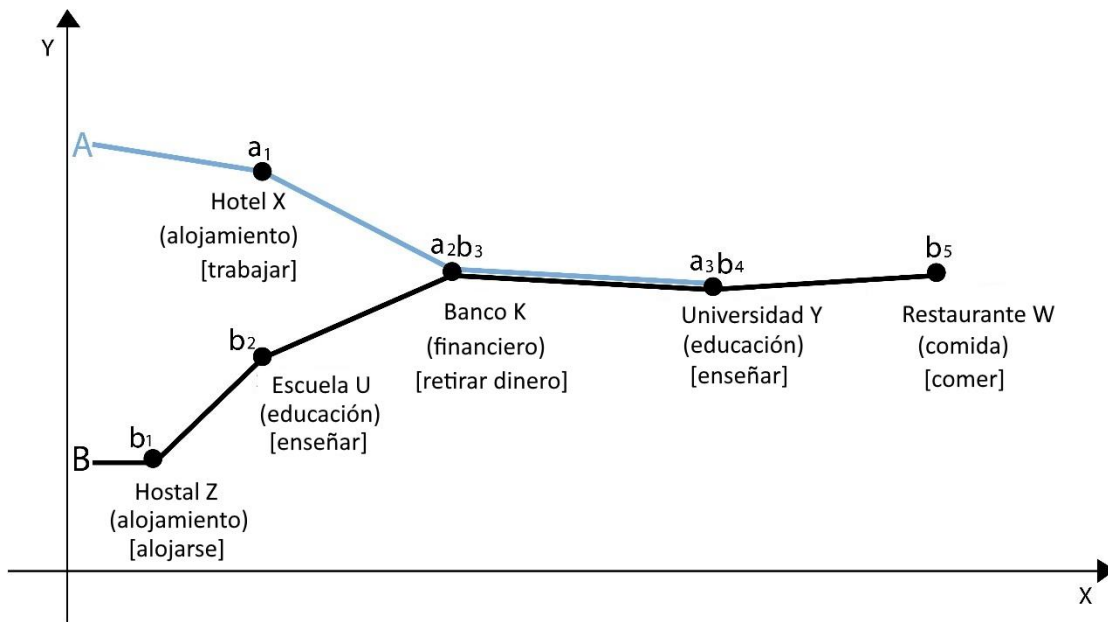


Figura 16. Ejemplo de dos trayectorias semánticas A y B

No hay una aproximación en la literatura que se enfoque en la similitud de trayectorias considerando tanto los sitios visitados como las actividades ejecutadas. Una aproximación, propuesta por Liu & Schneider (2012), divide una trayectoria semántica en sub-trayectorias y calcula la similitud semántica de dos trayectorias basado en la máxima subsecuencia común de sitios visitados. En dicha propuesta solo una coincidencia completa es considerada, i.e., 1 si hay una coincidencia en el nombre de los sitios y 0 en caso contrario (las actividades no son consideradas en este trabajo).

Xiao et al (2012) propuso una medida de similitud semántica que considera la semántica de los *stops*, la secuencia de los sitios visitados (*stops*), el tiempo de viaje entre los sitios, y la frecuencia en la que el sitio es visitado. Dos trayectorias son consideradas similares si visitan la misma secuencia de sitios, varias veces, y con tiempos de viaje similares. Nótese que esta propuesta es diferente de las medidas existentes de similitud porque considera la frecuencia de los sitios visitados, lo cual está más relacionado con patrones de trayectorias.

Más recientemente, Furtado et al. (2015) propusieron el MSM (*Multidimensional Similarity Measure*), la cual mide la similitud semántica en diferentes dimensiones, incluyendo la

semántica. Sin embargo, en esta aproximación la similitud de cada dimensión está dada por una función de distancia diferente, y la función específica para medir la similitud de cada dimensión no es el foco de ese trabajo. Los autores proponen una función que combina la similitud de las tres dimensiones en una única medida de similitud, pero los esfuerzos del trabajo no se centran en cómo medir la similitud en cada dimensión (temporal, espacial y semántica). Se proponen unas funciones de similitud básicas de cada dimensión con el fin de ejemplificar y dar mayor claridad al lector, e.g., la distancia espacial está determinada por la distancia euclidiana, mientras que la similitud semántica está dada por una coincidencia completa en el nombre del sitio (1 si hay una coincidencia completa o 0 en caso contrario).

Problemas como la clasificación y clustering de trayectorias son de especial interés debido a la información social o colectiva que pueden generar. Diferentes técnicas de clustering han sido propuestas con el fin de descubrir trayectorias similares. Por ejemplo, Lee et al. (2007) propusieron un método para agrupar trayectorias basado en la forma: dos trayectorias son consideradas similares si tienen sub-trayectorias en común (con la misma forma).

Zhao (2011) propuso un algoritmo de refinamiento progresivo, donde diferentes estrategias de clustering son definidas para descubrir trayectorias similares de acuerdo con la proximidad en tiempo y espacio. El algoritmo crea una matriz Booleana donde las columnas son los *stops* y las filas las trayectorias, y usa la *Dynamic Time Warping Distance* para medir la similitud entre trayectorias de acuerdo con la secuencia cronológica de *stops*.

Lee & Chung (2011) proponen un método para calcular la similitud entre usuarios, considerando su localización y los sitios que visitan. Se basa en un grafo de jerárquico de categorías, donde cada sitio visitado por un usuario está asociado con un nodo del grafo (llamado nodo de localización).

En (Ying et al., 2010), las trayectorias crudas se convierten en trayectorias semánticas mediante *stay cells*. Un *stay cell* representa una región geográfica donde el usuario hace un *stop* (excediendo un límite de tiempo). Subsecuentemente, asigna términos semánticos (como escuela, parque, banco, etc.) a esas celdas y define una medida de similitud entre trayectorias llamada *Maximal Semantic Trajectory Pattern Similarity (MSTP-Similarity)* basado en las *stay cells* de cada trayectoria.

Li et al. (2008) proponen un método para calcular la similitud entre usuarios basado en sus datos históricos de localización. A través de un *framework* llamado HGSM (*hierarchical-graph-based similarity measurement*) y un agrupamiento jerárquico de los sitios es posible explorar los sitios visitados por los usuarios en diferentes capas de similitud, donde la capa más fina contiene los usuarios con mayor similitud.

Yanagisawa et al. (2003) definen una medida de similitud entre dos trayectorias basado solo en características espacio temporales. Kreveld & Luo (2007) definen la disimilitud entre dos trayectorias basado en la distancia Euclidiana y sus marcas de tiempo. Similarmente, Lee et al. (2007), consideran sub-trayectorias para establecer la disimilitud.

Trajcevski et al. (2007) proponen un algoritmo que determina cuando una trayectoria es similar a una sub-trayectoria de otra trayectoria. Se basa en la distancia Euclidiana de tiempo uniforme (*Euclidean time-uniform distance*) (Cao, Wolfson, & Trajcevski, 2006), una variante de la distancia euclidiana que considera el tiempo en cual los eventos ocurren.

Tiakas et al. (2009) define dos medidas reflexivas entre dos trayectorias, una basada en el espacio, y la otra basada en tiempo; los cuales pueden ser combinadas para obtener una medida de similitud general, así, el usuario puede obtener la similitud entre dos trayectorias por esos criterios.

En este capítulo, se propone una nueva función para trayectorias semánticas, que soporte tanto la semántica de los sitios visitados por las trayectorias y las actividades ejecutadas en cada sitio, lo cual no ha sido considerado antes. Esta nueva función puede ser incorporada en trabajos previos, como MSM, para calcular la similitud de la dimensión semántica. Mientras trabajos previos solo consideran la coincidencia completa de la dimensión semántica, en este capítulo se propone una taxonomía de sitios y actividades que consideran la coincidencia parcial de los sitios y actividades.

## 2.1 Similitud de trayectorias

El árbol de categorías (conceptos) para la clasificación de sitios (CTCS) es un conjunto de nodos que tienen una relación padre-hijo y satisface que: el CTCS tiene un nodo especial  $r$  llamado "Sitio" (raíz), el cual no tiene nodo padre; y cada nodo  $ns \in CTCS$ , tal que  $ns \neq r$ , tiene un único nodo padre  $p \in CTCS$ ,  $p \neq ns$ . En la Figura 17, se muestra un ejemplo de un CTCS. La relación entre los nodos del CTCS es jerárquica, donde cada nodo hijo

representa una categoría más especializada que la categoría representada por su nodo padre.

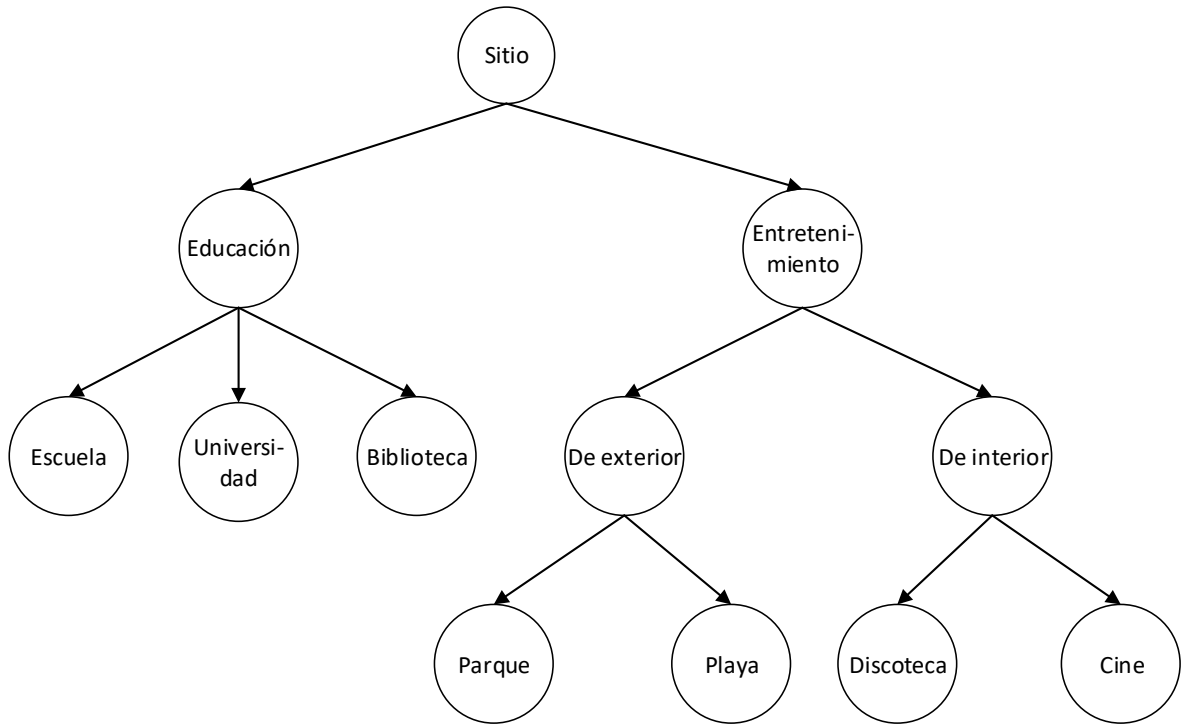


Figura 17. Ejemplo de CTCS

Similarmente, se considera un árbol de categorías para la clasificación de actividades (CTCA). Nótese que algunas combinaciones de sitios y actividades pueden no tener sentido, e.g., estudiar en una discoteca. Combinaciones validas pueden ser especificadas y controladas por el analista. En Figura 18 se muestra un ejemplo de un CTCA (adaptado de 23). Al igual que en el CTCS, el analista puede definir el CTCA como requisito para su aplicación.

Nótese que una actividad puede estar asociada con más de un nodo padre (e.g., la actividad “Bailar” puede ser también asociada al nodo “Motriz”); sin embargo, con el fin de acotar y delimitar el problema, se considera solo un nodo padre para cada actividad. A continuación, se introducen las principales definiciones de trayectorias semánticas y actividades usando como ejemplo la jerarquía mostrada en la Figura 17 y Figura 18.

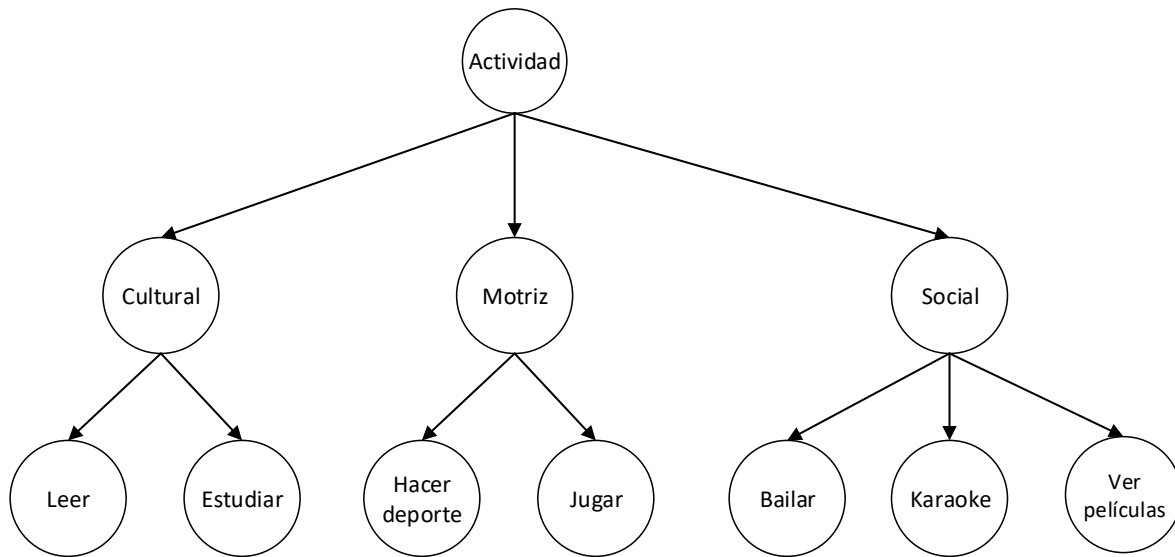


Figura 18. Ejemplo de CTCA

Sea  $S$  un conjunto de  $m$  sitios  $S = \{s_1, s_2, \dots, s_m\}$ , donde  $s_i = (s\_id, s\_name, s\_cat)$ , donde  $s\_id$  es el identificador del sitio,  $s\_name$  su nombre, y  $s\_cat$  representa la categoría del CTCS (nodo hoja) asociada con el sitio. Además, un sitio está (directamente) asociado con un nodo hoja del CTCS y (indirectamente) con todos sus nodos ancestros en el CTCS.

Ejemplo. Sea  $S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$  el conjunto de sitios, donde  $s_1 = (1, \text{Cinema Central}, \text{Cine})$ ,  $s_2 = (2, \text{Bocagrande}, \text{Playa})$ ,  $s_3 = (3, \text{Universidad de Cartagena}, \text{Universidad})$ ,  $s_4 = (4, \text{El Rosario}, \text{Playa})$ ,  $s_5 = (5, \text{Golden Disco}, \text{Discoteca})$ ,  $s_6 = (6, \text{Universidad de Bolívar}, \text{Universidad})$  y  $s_7 = (7, \text{Jardín Botánico}, \text{Parque})$ .

De manera similar, se define un conjunto de  $p$  actividades  $A = \{a_1, a_2, \dots, a_p\}$ , donde  $a_i = (a\_id, a\_name, a\_cat)$ , donde  $a\_id$  es el identificador de la actividad,  $a\_name$  su nombre, y  $a\_cat$  representa la categoría del CTCA (nodo hoja) asociada con la actividad.

Ejemplo. Sea  $A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$  el conjunto de actividades, donde  $a_1 = (1, \text{Estudiar matemáticas}, \text{estudiar})$ ,  $a_2 = (2, \text{Montar bicicleta}, \text{Hacer deporte})$ ,  $a_3 = (3, \text{Leer ciencia ficción}, \text{Leer})$ ,  $a_4 = (4, \text{Bailar electrónica}, \text{Bailar})$ ,  $a_5 = (5, \text{Estudiar español}, \text{Estudiar})$ ,  $a_6 = (6, \text{Nadar}, \text{Hacer deporte})$ ,  $a_7 = (7, \text{Cantar rock}, \text{karaoke})$  y  $a_8 = (8, \text{Ver películas de aventura}, \text{Ver películas})$ .

Por otro lado, una trayectoria  $T$  es un conjunto de  $n$  episodios  $T = \{e_1, e_2, \dots, e_n\}$ , donde  $e_i = (s_i, a_i, t_i)$ ,  $s_i \in S$  representa el sitio donde ocurrió,  $a_i \in A$  representa la actividad ejecutada en el sitio  $s_i$ , y  $t_i = (t_{ini}, t_{fin})$  representa el tiempo de inicio ( $t_{ini}$ ) y el tiempo de fin ( $t_{fin}$ ) del episodio,  $t_{ini} < t_{fin}$ .

Ejemplo. Considere la trayectoria  $T_1 = \{e_1, e_2, e_3, e_4\}$ , donde  $e_1 = (s_6, a_5, t_1)$ ,  $e_2 = (s_4, a_6, t_2)$ ,  $e_3 = (s_1, a_8, t_3)$ , y  $e_4 = (s_7, a_3, t_4)$ . La Tabla 5 muestra el detalle de los episodios de  $T_1$ .

Tabla 5. Eventos de la trayectoria  $T_i$

$e_i$	$s_i$			$a_i$			$t_i$	
	$s_{id}$	$s_{name}$	$s_{cat}$	$a_{id}$	$a_{name}$	$a_{cat}$	$t_{ini}$	$t_{fin}$
							Febrero 18, 2016	
$e_1$	6	Universidad de Bolívar	Universidad	5	Estudiar español	Estudiar	8am	12m
$e_2$	4	El Rosario	Playa	6	Nada	Hacer deporte	3pm	4pm
$e_3$	1	Cinema Central	Cine	8	Ver películas de aventura	Ver películas	4pm	5:30pm
$e_4$	7	Jardín botánico	Parque	3	Leer ciencia ficción	Leer	8pm	9pm

Sea  $POI_{ns, T_i}$  el conjunto de todos los sitios (directa o indirectamente) asociados con el nodo  $ns \in CTCS$  incluido en los episodios de la trayectoria  $T_i$ . La similitud entre dos trayectorias  $T_i$  y  $T_j$  con respecto a  $ns$ ,  $C_{ns, T_i, T_j}$ , es calculada por la Ecuación 5.

$$C_{ns, T_i, T_j} = \frac{|POI_{ns, T_i} \cap POI_{ns, T_j}|}{|POI_{ns, T_i} \cup POI_{ns, T_j}|}$$

Ecuación 5

Esto es,  $C_{ns, T_i, T_j}$  es la relación entre el número total de sitios comunes a las dos trayectorias asociados con el nodo  $ns$  y el total de número de sitios de las dos trayectorias asociados con ese nodo.  $C_{ns, T_i, T_j} = INDEF$  (Indefinido) si  $POI_{ns, T_i} \cup POI_{ns, T_j} = \emptyset$ , i.e., cuando ninguna de las dos trayectorias tiene sitios asociados con el nodo  $ns$ . Nótese que la

Ecuación 5 está basada en el índice de Jaccard, cuyo rango es el intervalo  $[0, 1]$  y su valor es 1 cuando ambos conjuntos son vacíos; sin embargo, en esta propuesta cuando eso sucede se asigna el valor de INDEF.

Nótese que en esta propuesta si el mismo sitio es incluido en diferentes episodios de la trayectoria, esta medida de similitud lo considera solo una vez. Otro aspecto para tener en cuenta es el siguiente. Suponga que una trayectoria  $T_3$  tiene un único episodio que incluye el sitio  $s_3 = (3, Universidad de Cartagena, Universidad)$  y una trayectoria  $T_4$  tiene un único episodio que incluye al sitio  $s_6 = (6, Universidad de Bolívar, Universidad)$ , i.e., ambas trayectorias incluyen una universidad en sus respectivos episodios, pero debido a que las universidades son diferentes, entonces  $C_{University, T_3, T_4} = 0$ . Nótese que esas dos trayectorias incluyen en sus episodios dos sitios diferentes pero que están asociados al mismo nodo (Universidad). A estos sitios se denominarán *non-matching*. Esta situación sugiere una similitud mayor que cero. Así, con el fin de incorporar estos sitios en la medida de similitud, se propone un parámetro llamado peso de sitios tipo *non-matching*  $nmsw \in [0, 1]$ . Este parámetro actúa como un peso mediante el cual el usuario establece el grado de contribución de los sitios tipo *non-matching* a la similitud. Así, la fórmula para la similitud es modificada de acuerdo con la Ecuación 6 (el parámetro  $nnms$  llamada número de sitios tipo *non-matching* se explica luego).

$$C_{ns, T_i, T_j, nmsw} = \frac{|POI_{ns, T_i} \cap POI_{ns, T_j}| + nmsw * nnms}{|POI_{ns, T_i} \cup POI_{ns, T_j}| - nmsw * nnms}$$

Ecuación 6

De manera similar a la Ecuación 5, la Ecuación 6 tiene un rango en el intervalo  $[0, 1]$  y es INDEF cuando  $SA \neq T_i POI_{ns, T_i} \cup POI_{ns, T_j} = \emptyset$ . Nótese que cuando  $nmsw = 0$ , entonces la Ecuación 6 es igual a la Ecuación 5. Por ejemplo, considerando de nuevo las trayectorias  $T_3$  y  $T_4$  se obtiene  $C_{University, T_3, T_4, 0} = 0$ . Además, cuando  $nmsw = 1$ ,  $C_{University, T_3, T_4, 1} = 1$ , i.e., se considera que las trayectorias  $T_3$  y  $T_4$  son 100% similares respecto al nodo Universidad porque, aunque visitaron sitios diferentes ( $s_3$  and  $s_6$ ), ambas pertenecen a la misma categoría (Universidad).

Ahora se explicará el parámetro  $nnms$ . Considérese las trayectorias  $T_5$  y  $T_6$ .  $T_5$  incluye en sus episodios los siguientes sitios asociados con el nodo universidad:  $POI_{Universidad, T_5} =$

$\{s_{10}, s_{11}, s_{12}, s_{13}\}$  donde  $s_{10} = (10, \text{Universidad A}, \text{Universidad})$ ,  $s_{11} = (11, \text{Universidad B}, \text{Universidad})$ ,  $s_{12} = (12, \text{Universidad C}, \text{Universidad})$  y  $s_{13} = (13, \text{Universidad D}, \text{Universidad})$ .  $T_6$  incluye en sus episodios los siguientes sitios asociados con el nodo universidad:  $POI_{\text{Universidad}, T_6} = \{s_{10}, s_{14}, s_{15}\}$  donde  $s_{14} = (14, \text{Universidad E}, \text{Universidad})$ ,  $s_{15} = (15, \text{Universidad F}, \text{Universidad})$ .

Además, trayectorias  $T_5$  y  $T_6$  tienen un sitio en común (sitio  $s_{10}$ , Universidad A), i.e.,  $|(POI_{ns, T_5} \cap POI_{ns, T_6})| = 1$ . Por otro lado,  $T_5$  tiene en sus episodios tres diferentes universidades en comparación con  $T_6$ , mientras  $T_6$  tiene en sus episodios dos diferentes universidades en comparación con  $T_5$ . Así, aunque  $T_5$  visitó más universidades en comparación con  $T_6$ , se puede concluir que cada trayectoria tiene en sus episodios al menos dos universidades (además de la que tienen en común, Universidad A). Esto es el valor de  $nnms$ . Formalmente,  $nnms$  es calculado de acuerdo con la Ecuación 7

$$nnms = \text{Min}(|POI_{ns, T_i} - POI_{ns, T_j}|, |POI_{ns, T_j} - POI_{ns, T_i}|)$$

Ecuación 7

Considérese de nuevo las trayectorias  $T_5$  y  $T_6$ ,  $nmsw = 1$  y  $ns = \text{Universidad}$ . Nótese que si el término  $nmsw * nnms$  no es considerado en el denominador de la Ecuación 6, la similitud estaría dada por  $C_{ns, T_5, T_6, 1} = \frac{1+1*2}{6} = \frac{3}{6} = 0.5$ , puesto que es considerado que las dos universidades visitadas por  $T_6$  ( $s_{14}$ ,  $s_{15}$ ) son “iguales” a las dos universidades visitadas por  $T_5$  (dos de  $s_{11}$ ,  $s_{12}$ ,  $s_{13}$ ), i.e., que las trayectorias tengan en común dos universidades más (además de  $s_{10}$ , así el numerador es 3), entonces el total de “diferentes” universidades entre las dos trayectorias debe ser cuatro y no seis, tal como la intersección de los sitios visitados aumenta de uno a tres. Luego, el término  $nmsw * nnms$  es restado en el denominador y  $C_{ns, T_5, T_6, 1} = \frac{1+1*2}{6-1*2} = \frac{3}{4} = 0.75$ , (en términos prácticos eso significa que  $T_5$  y  $T_6$  tienen tres de cuatro universidades en común). Además,  $C_{ns, T_5, T_6, 0} = \frac{1}{6} = 0.167$  (i.e.,  $T_5$  y  $T_6$  tiene estrictamente solo una de seis universidades en común) y  $C_{ns, T_5, T_6, 0.5} = \frac{1+0.5*2}{5} = \frac{2}{5} = 0.4$  (con  $nmsw = 0.5$  significa, en términos prácticos, que  $T_5$  y  $T_6$  tienen dos de cinco universidades en común).

Inicialmente, se proponen dos métodos para calcular la similitud entre dos trayectorias considerando solo los sitios incluidos en sus episodios, i.e., basado en el CTCS. Luego se consideran las actividades ejecutadas en cada sitio para establecer la similitud.

## 2.2 Método 1

Considérese dos trayectorias  $T_i$  y  $T_j$ . En este método se calcula la similitud de cada nodo  $ns \in CTCS$  a través de la Ecuación 6, i.e.,  $SIM_{ns} = C_{ns,T_i,T_j,nmsc}$ . De esta forma, el usuario puede analizar la similitud de las trayectorias respecto a cada nodo del CTCS. Por ejemplo, si  $ns$  es la raíz de CTCS, entonces  $C_{ns,T_i,T_j,nmsc}$  indica la similitud de las trayectorias desde un punto de vista general (nodo "Sitio"). El usuario puede analizar la similitud desde un punto de vista más específico a medida que desciende a través de los niveles del CTCS (un "drill-down").

Nótese que, en este método, para calcular la similitud de un nodo no hoja, no es necesario calcular la similitud de sus nodos hijos (confrontar con el método 2).

Ejemplo: Considérese la trayectoria  $T_2 = \{e_1, e_2, e_3, e_4, e_5\}$ , donde  $e_1 = (s_3, a_1, t_1)$ ,  $e_2 = (s_2, a_4, t_2)$ ,  $e_3 = (s_4, a_2, t_3)$ ,  $e_4 = (s_1, a_8, t_3)$  y  $e_5 = (s_5, a_7, t_3)$ . La Tabla 6 muestra en detalle los episodios de T2.

Tabla 6. Episodios de la trayectoria T2

$e_i$	$s_i$			$a_i$			$t_i$	
	$s\_id$	$s\_name$	$s\_cat$	$a\_id$	$a\_name$	$a\_cat$	$t_{ini}$	$t_{fin}$
							Febrero 18, 2016	
$e_1$	3	Universidad de Cartagena	Universidad	1	Estudiar matemáticas	Estudiar	7am	10am
$e_2$	2	Bocagrande	Playa	4	Bailar electrónica	Bailar	11am	1pm
$e_3$	4	El Rosario	Playa	2	Montar bicicleta	Hacer deporte	2pm	3pm
$e_4$	1	Cinema Central	Cine	8	Ver películas de aventura	Ver películas	9pm	11pm
$e_5$	5	Golden Disco	Discoteca	7	Cantar rock	Karaoke	10:30pm	11:30pm

Con  $nmsw = 0.5$  y considerando las trayectorias  $T_1$  y  $T_2$ , el CTCS con la similitud de cada nodo es mostrando en la Figura 19. Por ejemplo, el cálculo de  $SIM_{Universidad}$  (un nodo hoja) es obtenido de esta forma: las trayectorias no tienen sitios en común con respecto a este nodo, i.e.,  $|POI_{ns,T_1} \cap POI_{ns,T_2}| = 0$ , donde  $ns = Universidad$ . Además, cada trayectoria incluye en sus episodios una universidad, i.e.,  $nnms = 1$ ; luego,  $Sim_{Universidad} = \frac{(0 + 0.5 * 1)}{(2 - 0.5 * 1)} = 0.33$ .

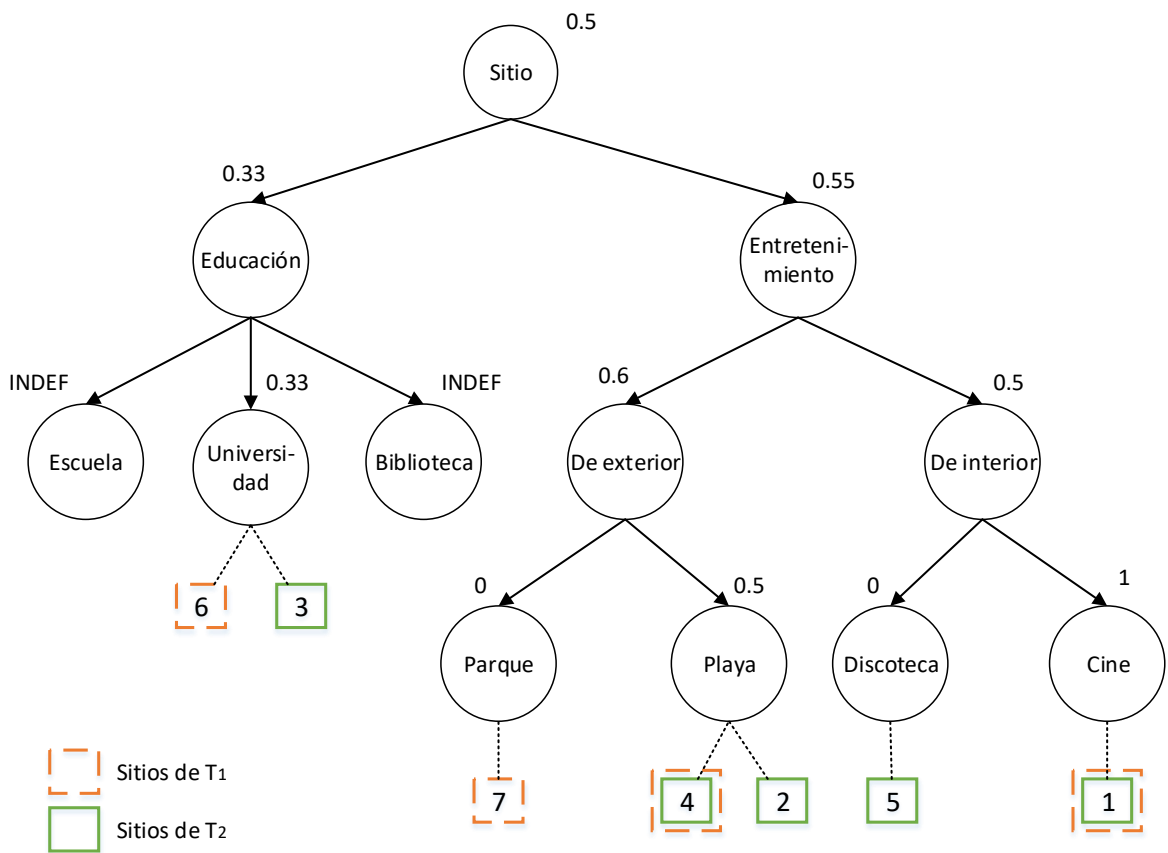


Figura 19. CTCS con valores de similitud para T1 y T2 usando el método 1

Para calcular  $SIM_{Entretenimiento}$  (un nodo no hoja) sus nodos hojas son considerados (Parque, Playa, Discoteca, Cine). Las trayectorias tienen dos sitios en común (s1 y s4),  $nnms = 1$ , y  $|POI_{ns,T_1} \cup POI_{ns,T_2}| = 5$ , donde  $ns = Entretenimiento$ . Así,  $SIM_{Entretenimiento} = \frac{(2 + 0.5 * 1)}{(5 - 0.5 * 1)} = 0.55$ .

## 2.3 Método 2

En este método, cada nodo  $ns \in CTCS$  tiene una similitud  $SIM_{ns}$ , donde  $SIM_{ns} = C_{ns,T_i,T_j,nmsw}$  para un nodo hoja, i.e., Ecuación 6. Para un nodo no hoja,  $SIM_{ns}$  es calculado por la Ecuación 8.

$$SIM_{ns} = \sum(\text{weight}_{nl} * SIM_{nl}), \forall nl \in H \wedge SIM_{nl} \neq INDEF$$

Ecuación 8

Donde  $H$  es el conjunto de nodos hijo del nodo  $ns$  y  $\text{weight}_{nl}$ , denominado peso del nodo  $nl$ , es el peso asignado por el analista al nodo  $nl$ , i.e., el analista puede especificar el peso con el que cada nodo hijo  $nl$  contribuye a la similitud de su nodo padre  $ns$ . Por ejemplo, un usuario puede considerar para el nodo “De exterior” que las playas deben pesar (contribuir) más en la similitud que los parques. Para esto, él podría especificar que  $\text{weight}_{Beach} = 0.8$  y  $\text{weight}_{Park} = 0.2$ . Nótese que la suma de los pesos de los hijos de un nodo debe ser igual a 1, i.e.,  $\sum \text{weight}_{nl} = 1, \forall nl \in H$ .

Ejemplo: Considere las trayectorias  $T_1$  y  $T_2$ . El CTCS con la similitud de cada nodo es mostrado en la Figura 20. El mismo peso fue considerado para los nodos hijos de un nodo. Por ejemplo, para  $nmsw = 0$ ,  $SIM_{Playa}$  (un nodo hoja) se obtiene de esta forma: puesto que ambas trayectorias incluyen en sus episodios el sitio  $s_4$  y  $T_2$  también incluye el sitio  $s_2$ , entonces  $SIM_{playa} = C_{Beach,T_1,T_2,0} = \frac{1+0*0}{2} = 0.5$ .

Para calcular  $SIM_{De\ exterior}$  (un nodo no hoja), se considera la similitud de los nodos hoja Parque ( $SIM_{Parque} = 0$ ) y Playa ( $SIM_{Playa} = 0.5$ ); aplicando la Ecuación 8 con  $\text{weight}_{Playa} = \text{weight}_{Parque} = 0.5$  se obtiene:  $(0.5 * 0 + 0.5 * 0.5) = 0.25$ . Para calcular  $SIM_{Educación}$  (otro nodo no hoja) se considera solo la similitud del nodo universidad, puesto que la similitud de los nodos escuela y biblioteca es INDEF.

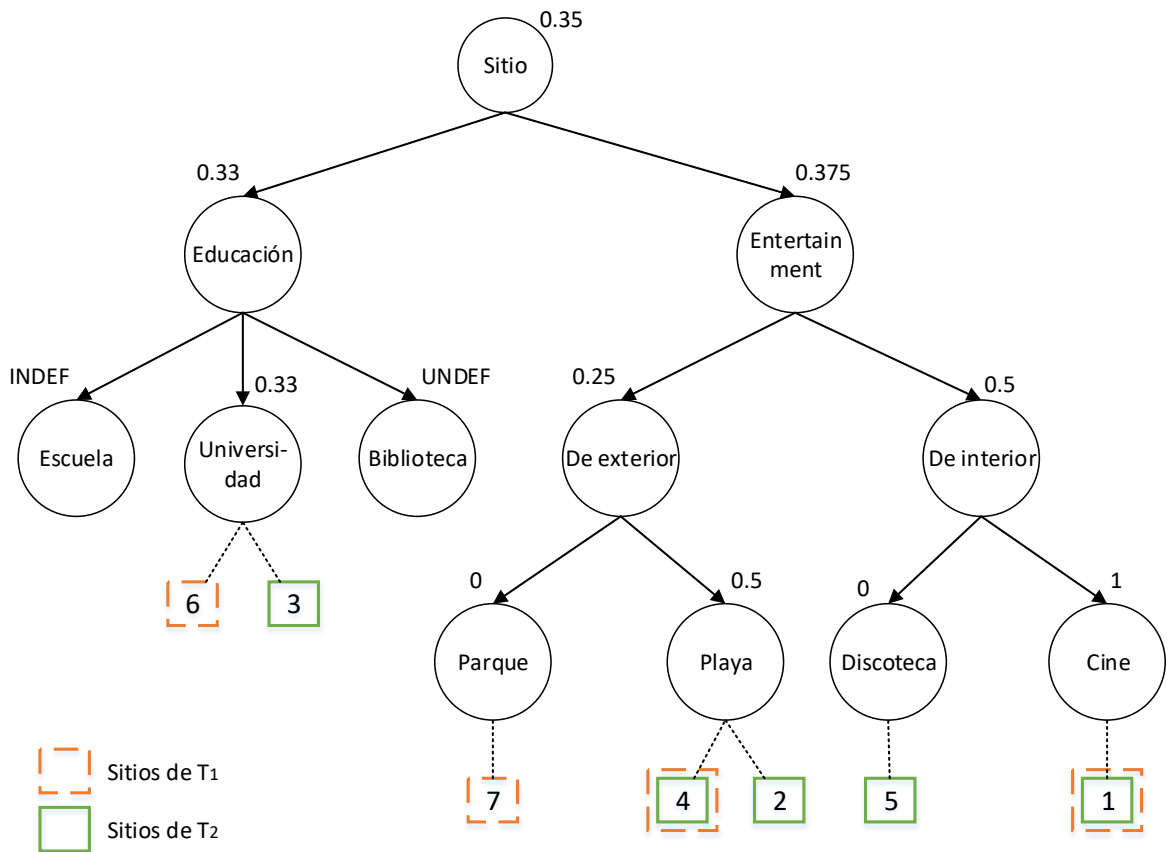


Figura 20. CTCS con valores de similitud para  $T_1$  y  $T_2$  usando el método 2

## 2.4 Diferencias e interpretación de los dos métodos

La Figura 19 y la Figura 20 muestran que la similitud de dos trayectorias con respecto a nodos no hoja puede ser diferentes dependiendo del método que sea aplicado (en ambos métodos la similitud respecto a nodos hoja es igual). Por ejemplo, la similitud de las trayectorias  $T_1$  y  $T_2$  con respecto al nodo hoja (Sitio) es 0.5 con el método 1 y 0.35 con el método 2. Esta diferencia ocurre debido a la asignación de pesos de los nodos hijos y al número de sitios de la trayectoria asociados con los nodos hoja. Por ejemplo, si se considera el mismo peso  $w$  para los nodos hijos de un nodo  $ns$ , la diferencia de la similitud obtenida en los dos métodos con respecto a  $ns$  se vuelve más grande a medida que el conjunto de sitios de la trayectoria  $T_i$  asociados con los nodos hoja descendientes de  $ns$  se vuelve más grande con respecto al correspondiente conjunto de sitios de la trayectoria  $T_j$ . Esto es porque el método 1 considera para cada nodo (bien sea hoja o no) todos los sitios asociados con él (directa o indirectamente), mientras que en el método 2 después de

calcular la similitud de cada nodo, la similitud de  $ns$  es calculada considerando únicamente la similitud de sus hijos y el peso  $w$  asignado a cada uno de ellos.

Considere el CTCS de la Figura 21 y dos trayectorias  $T_7$  y  $T_8$ . Considérese el nodo discoteca (un nodo hoja),  $nnms = 0$ , y supóngase que  $|(POI_{ns,T_7} \cap POI_{ns,T_8})| = 0$  y  $|(POI_{ns,T_7} \cup POI_{ns,T_8})| = 1$ , donde  $ns = \text{Discoteca}$ . Además, considérese el nodo cine (un nodo hoja),  $nnms = 0$ , y supóngase que  $|(POI_{ns,T_7} \cap POI_{ns,T_8})| = 7$  y  $|(POI_{ns,T_7} \cup POI_{ns,T_8})| = 10$ , donde  $ns = \text{Cine}$ . Considérese  $weight = 0.5$ , con el método 1 se obtiene  $SIM_{\text{Discoteca}} = 0$ ,  $SIM_{\text{Cine}} = 0.7$ , and  $SIM_{\text{De interior}} = 0.6363$ . Con el método 2 se obtiene  $SIM_{\text{Discoteca}} = 0$ ,  $SIM_{\text{Cine}} = 0.7$ , and  $SIM_{\text{De interior}} = 0.35$ . Nótese que la diferencia en las similitudes respecto al nodo De interior es 0.28.

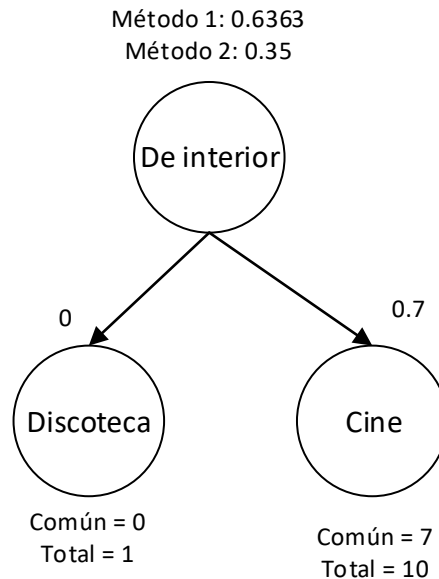


Figura 21. Diferentes valores de similitud obtenidos con el método 1 y método 2

Sea  $U_1$  y  $U_2$  dos usuarios con trayectorias  $T_i$  y  $T_j$ , respectivamente. Tomando como ejemplo el nodo  $ns = \text{“Entretenimiento”}$ , y asumiendo que  $U_1$  visitó más sitios de entretenimiento de exterior y  $U_2$  visitó más sitios de entretenimiento de interior, se determinará cuando es apropiado aplicar el método 1 o el método 2. Para esto, considérese la pregunta: ¿es importante considerar el tipo de entretenimiento experimentado por el usuario o solo es importante que el usuario se haya entretenido (sin importar el tipo de entretenimiento)? Si en el dominio de aplicación es importante diferenciar el tipo de entretenimiento experimentado por los usuarios, i.e., que la medida de similitud es afectada porque cada

usuario visitó más sitios en diferentes categorías, es apropiado usar el método 2 puesto que este considera todas las subcategorías (e incluso es posible asignar pesos diferentes a cada tipo de entretenimiento); sin embargo, si solo se quiere obtener una medida de similitud respecto al tipo de entretenimiento es apropiado usar el método 1.

Nótese que en el método 1 todos los sitios permanecen con el mismo nivel de importancia, e.g., en la Figura 21 una discoteca es igual de importante que un cine puesto que el tipo específico de sitio no es de interés; mientras que en el método 2, una discoteca se vuelve más relevante (pesa más en el cálculo de la similitud). Consecuentemente, el valor de la similitud decrece con respecto al método 1.

## 2.5 Algoritmos de similitud entre trayectorias

A continuación, se proponen dos algoritmos para encontrar la similitud entre dos trayectorias, correspondientes a los métodos explicados anteriormente.

### Algoritmo 1. Algoritmo para el método 1

SimMethod1(T1, T2, nmsw, G, ns)

**Input:** T1, T2: Trayectorias, nmsw, G: CTCS, ns: Nodo  $\in$  G

**Output:** Nodo ns con su similitud

**BEGIN**

1. ST = G.subTree(ns); //Extrae el subárbol con ns como raíz
  2. L = leafNodes(ST); //Extrae el conjunto de nodos hoja de ST
  3. S1 = {}; //Conjunto de sitios de T1 relacionados con los nodos de interés para calcular la similitud
  4. S2 = {}; //Conjunto de sitios de T2 relacionados con los nodos de interés para calcular la similitud
  5. **FOREACH** nsAux  $\in$  L
  6.     Agregar a S1 los sitios de T1 relacionados con el nodo nsAux
  7.     Agregar a S2 los sitios de T2 relacionados con el nodo nsAux
  8. **END FOR**
  9. **IF** |S1| = 0 **AND** |S2| = 0 **THEN**
  10.    ns.sim = INDEF;
  11. **ELSE**
  12.    nms = MIN(|S1 - S2|, |S2 - S1|);
  13.    ns.sim = (|S1  $\cap$  S2| + nmsw \* nms) / (|S1  $\cup$  S2| - nmsw \* nms);
  14. **END IF**
- END** SimMethod1

---

 Algoritmo 2. Algoritmo para el método 2
 

---

SimMethod2( $T_1, T_2, nmsw, G, ns$ )

**Input:**  $T_1, T_2$ : Trayectorias,  $nmsw, G$ : CTCS,  $ns$ : Nodo  $\in G$

**Output:** Nodo  $ns$  con su similitud

**BEGIN**

1. **IF**  $ns.isLeaf$  **THEN**

2.     SimMethod1( $T_1, T_2, nmsw, G, ns$ );

3. **ELSE**

4.      $H = G.children(ns)$ ; //Extrae el conjunto de hijos de un nodo  $ns$

5.      $sum = 0$ ;

6.     **FOREACH**  $nsAux \in H$

7.         Sim2( $T_1, T_2, nmsw, G, nsAux$ );

8.         **IF**  $nsAux.sim \neq INDEF$  **THEN**

9.              $sum += nsAux.weight * nsAux.sim$ ;

10.         **END IF**

11.     **END FOR**

12.      $ns.sim = sum$ ;

13. **END IF**

**END** SimMethod2

---

Nótese que el algoritmo SimMethod2 calcula la similitud de todos los descendientes del nodo de interés  $ns$ , lo cual permite acceder al valor de similitud de cualquiera de esos nodos sin tener que calcularla de nuevo. Esto también permite encontrar la similitud de cada nodo del CTCS cuando es invocado con el nodo raíz.

También se puede usar los algoritmos 1 y 2 para encontrar la similitud entre dos trayectorias con respecto a las actividades ejecutadas usando un CTCA en vez de un CTCS. Por ejemplo, si SimMethod1 es invocado con el CTCA de la Figura 18, i.e.,  $SimMethod1(T_1, T_2, 0.5, CTCA, Actividad)$ , los resultados son mostrados en la Figura 22.

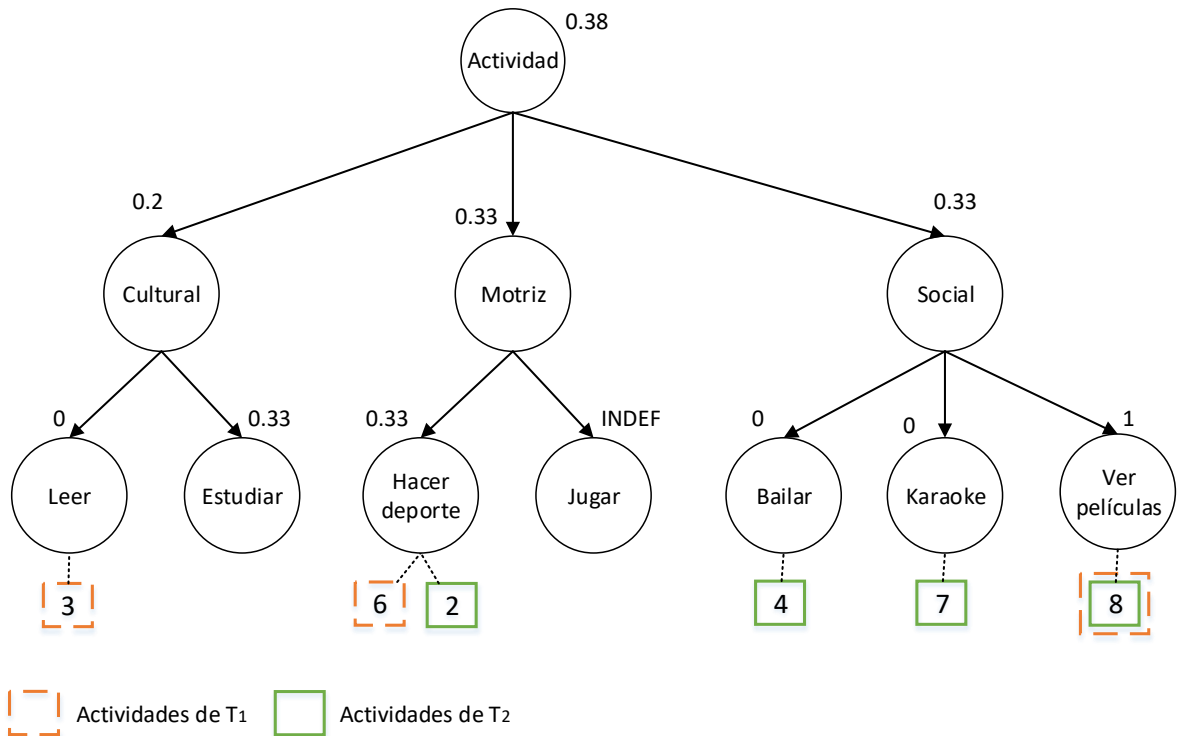


Figura 22. CTCA con valores de similitud para  $T_1$  y  $T_2$  usando el método 1

## 2.6 Similitud combinada: Sitios y actividades

Hasta el momento se ha calculado la similitud basada en los sitios visitados o en las actividades ejecutadas en esos sitios, pero ambos criterios no han sido considerados simultáneamente. A continuación, se propone un método para esto.

Sea  $ns \in CTCS$  el nodo de interés  $T_1$  el subconjunto de episodios correspondientes a  $T_1$  cuyos sitios están asociados al nodo  $ns$  o a los descendientes de  $ns$ . La similitud con respecto a las actividades es obtenida aplicando los métodos 1 o 2 enviando como  $ns = Actividad$  como parámetro. Entonces es obtenido un CTCA con un valor para cada uno de sus nodos, el cual representa la similitud entre las dos trayectorias basado en las actividades ejecutadas en el sitio  $ns$ ; el valor del nodo *Actividad* representa la similitud con respecto a todas las actividades ejecutadas por los usuarios en el sitio  $ns$ .

Nótese que para cada nodo  $ns \in CTCS$ , un CTCA es generado con valores de similitud para cada nodo CTCA, el cual indica la similitud de cada actividad ejecutada en el sitio  $ns$ .

Si  $ns$  es la raíz del CTCS, entonces el CTCA generado representa la similitud de todas las actividades ejecutadas respecto al sitio donde son ejecutadas.

Ejemplo: En la Figura 23, se muestra el CTCA con los valores de similitud cuando se aplica el método 1 para el nodo  $ns = \text{Entretenimiento}$  y  $nmsw = 0.5$ . Para calcular la similitud en el nodo cultural, solo se usa  $a_3$  puesto que  $a_1$  y  $a_5$  no fueron ejecutadas en sitios de entretenimiento; entonces,  $SIM_{Cultural} = 0$ . Se concluye que las trayectorias  $T_1$  y  $T_2$  son similares en 0.4 respecto a sitios de entretenimiento, y 0.25 respecto a las actividades ejecutadas en ese tipo de sitio.

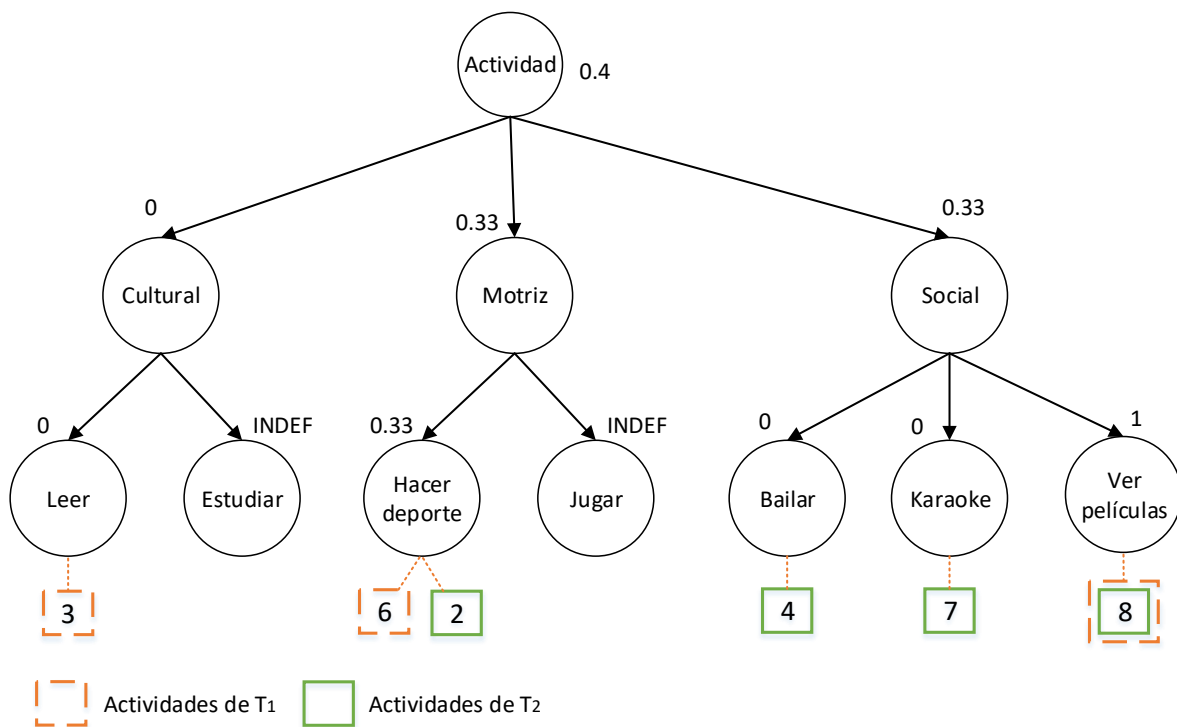


Figura 23. CTCA con valores de similitud para el nodo  $ns = \text{Entretenimiento}$  para  $T_1$  y  $T_2$  usando el método 1

## 2.7 Experimentos y resultados

En esta sección se presentan los resultados de experimentos y se comparan con la propuesta de Zhao et al. (2009). Para los experimentos, se usó una base de datos que guarda el registro de usuario de *Foursquare* en New York entre octubre 24 de 2011 y febrero 20 de 2012. Se realizó una implementación de los algoritmos en lenguaje de

programación JAVA y se usó una base de datos relacional en PostgreSQL. Los sitios fueron clasificados de acuerdo con las etiquetas indicadas por los usuarios y el CTCS es mostrado en la Figura 24. De manera similar, las actividades fueron clasificadas de acuerdo con los comentarios dejados por los usuarios cuando hicieron check-in; el CTCA usado es mostrado en la Figura 25. Puesto que no todos los usuarios tienen registros de check-in en los comentarios, fue necesario asumir algunas actividades de acuerdo con la actividad más probable que el usuario hizo en cada sitio. Para el análisis se escogieron los 51 pares de trayectorias que tienen más sitios en común.

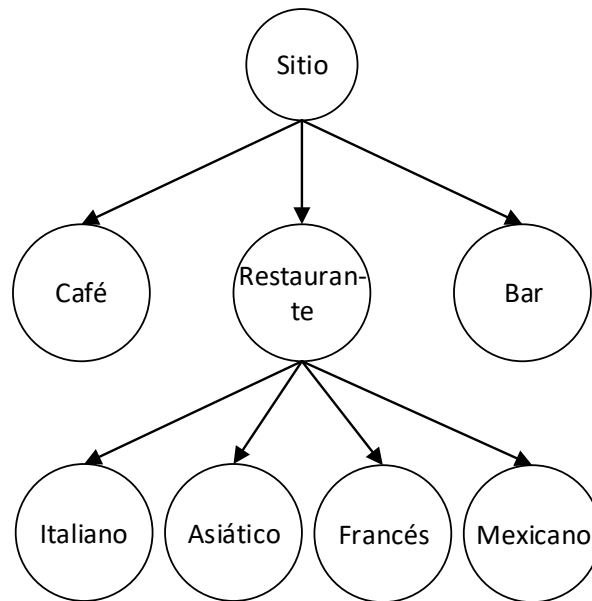


Figura 24. CTCS usado para los experimentos

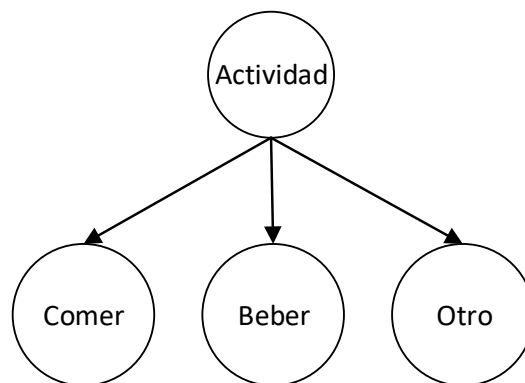


Figura 25. CTCA usado para los experimentos

Inicialmente, los métodos 1 y 2 fueron aplicados para encontrar la similitud basada en los sitios y las actividades de manera separada, con  $nmsw = 0.5$ , y el mismo peso para cada nodo en el mismo nivel en el método 2. La Tabla 7 y la Tabla 8 muestran los resultados obtenidos en cada método.

Tabla 7. Resultados de los métodos 1 y 2 respecto a los sitios

Nodo	Similitud del método 1			Similitud del método 2		
	Promedio	Max	Min	Promedio	Max	Min
Sitio	0.35	0.87	0.07	0.29	0.65	0.09
Bar	0.31	0.92	0	0.31	0.92	0
Café	0.28	0.85	0	0.28	0.85	0
Restaurante	0.34	0.92	0.05	0.28	0.54	0.06
Italiano	0.27	0.73	0	0.27	0.73	0
Asiático	0.32	0.75	0	0.32	0.75	0
Francés	0.22	0.8	0	0.22	0.8	0
Mexicano	0.3	1	0	0.3	1	0

Tabla 8. Resultados de los métodos 1 y 2 respecto a las actividades

Nodo	Similitud método 1			Similitud método 2		
	Promedio	Max	Min	Promedio	Max	Min
Actividad	0.41	0.91	0.07	0.41	0.83	0.07
Comer	0.39	0.9	0.07	0.39	0.9	0.07
Beber	0.42	0.98	0.09	0.42	0.98	0.09
Otro	0.42	1	0	0.42	1	0

La Figura 26 muestra los resultados obtenidos por el par de trayectorias que obtuvieron la mayor similitud en el nodo Sitio y Actividad. Nótese que, aunque los resultados en los dos métodos fueron diferentes, la pareja que obtuvo la mayor similitud en ambos casos fue la misma.

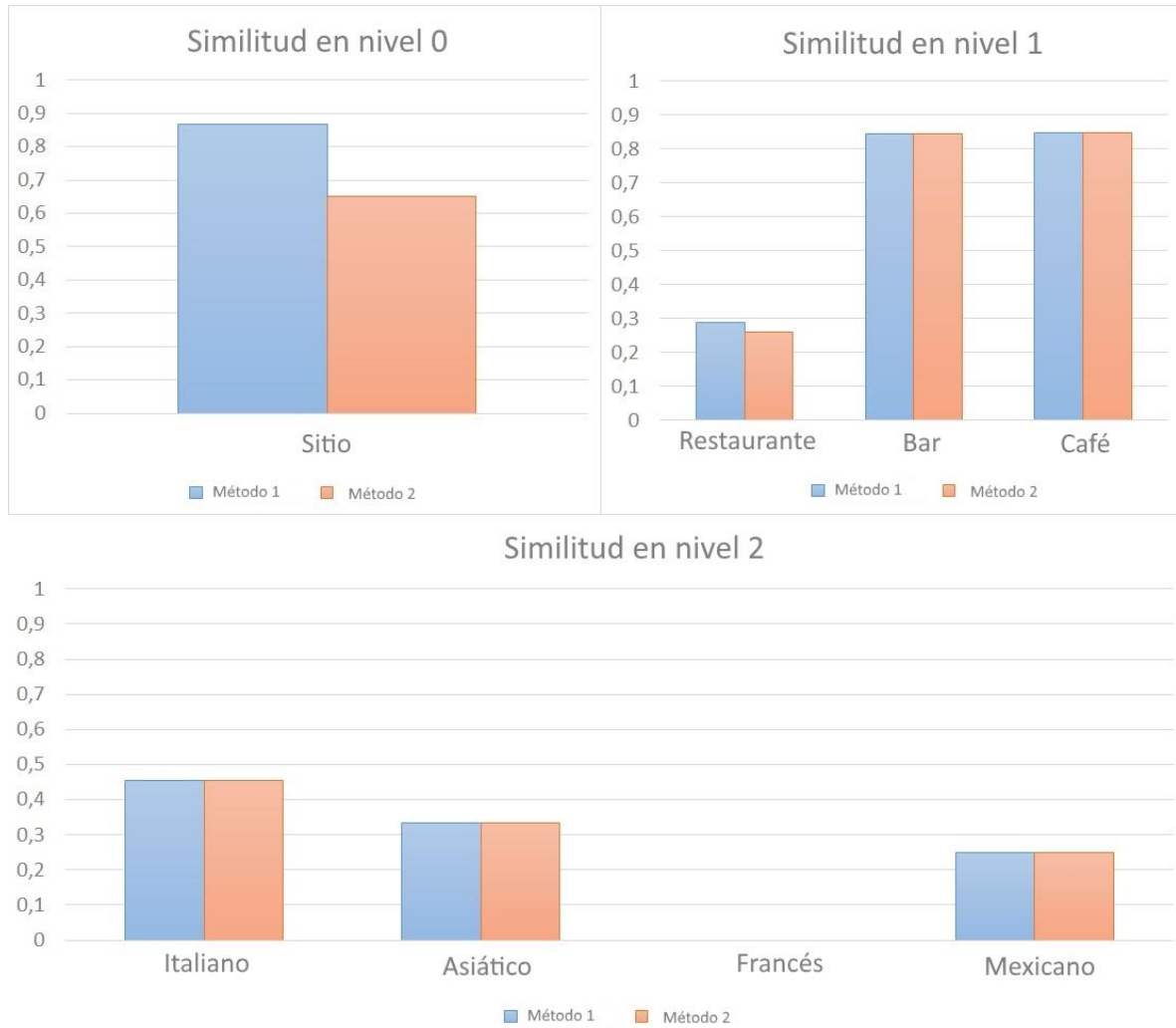


Figura 26. Resultados para los pares de trayectorias que obtuvieron la mayor similitud respecto a los sitios.

Nótese que, debido a la naturaleza de los métodos, en los nodos hoja el mismo valor de similitud se obtendrá sin importar cual método se use. Subsecuentemente, la similitud basada en sitios y actividades fue aplicada. La similitud de cada nodo perteneciente al CTCs fue calculada y también la similitud obtenida en el nodo Actividad es mostrada. La Tabla 9 y la Tabla 10 muestran la similitud promedio de cada nodo sitio y la similitud promedio de los usuarios que ejecutaron actividades en ese tipo de sitio.

Tabla 9. Resultados de similitud combinada con el método 1

Nodo	Similitud combinada con el método 1					
	Sitio			Nodo Actividad		
	Promedio	Max	Min	Promedio	Max	Min
Sitio	0.35	0.87	0.07	0.41	0.91	0.07
Bar	0.31	0.92	0	0.19	0.5	0
Café	0.28	0.85	0	0.22	0.79	0
Restaurante	0.34	0.92	0.05	0.3	0.65	0.03
Italiano	0.27	0.73	0	0.15	0.36	0
Asiático	0.32	0.75	0	0.22	0.47	0
Francés	0.22	0.8	0	0.13	0.5	0
Mexicano	0.3	1	0	0.18	0.45	0

Tabla 10. Resultados de similitud combinada con el método 2

Nodo	Similitud combinada con el método 2					
	Sitio			Nodo Actividad		
	Promedio	Max	Min	Promedio	Max	Min
Sitio	0.29	0.65	0.09	0.41	0.83	0.07
Bar	0.31	0.92	0	0.13	0.43	0
Café	0.28	0.85	0	0.2	0.64	0
Restaurante	0.28	0.54	0.06	0.26	0.58	0.02
Italiano	0.27	0.73	0	0.11	0.44	0
Asiático	0.32	0.75	0	0.16	0.37	0
Francés	0.22	0.8	0	0.09	0.63	0
Mexicano	0.3	1	0	0.12	0.53	0

A continuación, se usaron los mismos 51 pares de trayectorias para la comparación con la propuesta de Zhao, y debido a que esa propuesta no considera actividades, se aplican los métodos 1 y 2 considerando únicamente los sitios. Los experimentos para diferentes valores de *nmsw* fueron usados y se observó como la similitud cambió.

La Figura 27 muestra la medida de similitud obtenida en cada uno de los tres métodos cuando  $nmsw = 0$ , se puede ver que tanto el método 1 como la propuesta de Zhao tienen valores iguales en todos los casos puesto que  $nmsw = 0$ , las ecuaciones de similitud son iguales en ambos casos. El método 2 presenta diferentes valores porque usa el CTCS (su estructura jerárquica) y los pesos asignados a cada nodo para determinar la similitud.

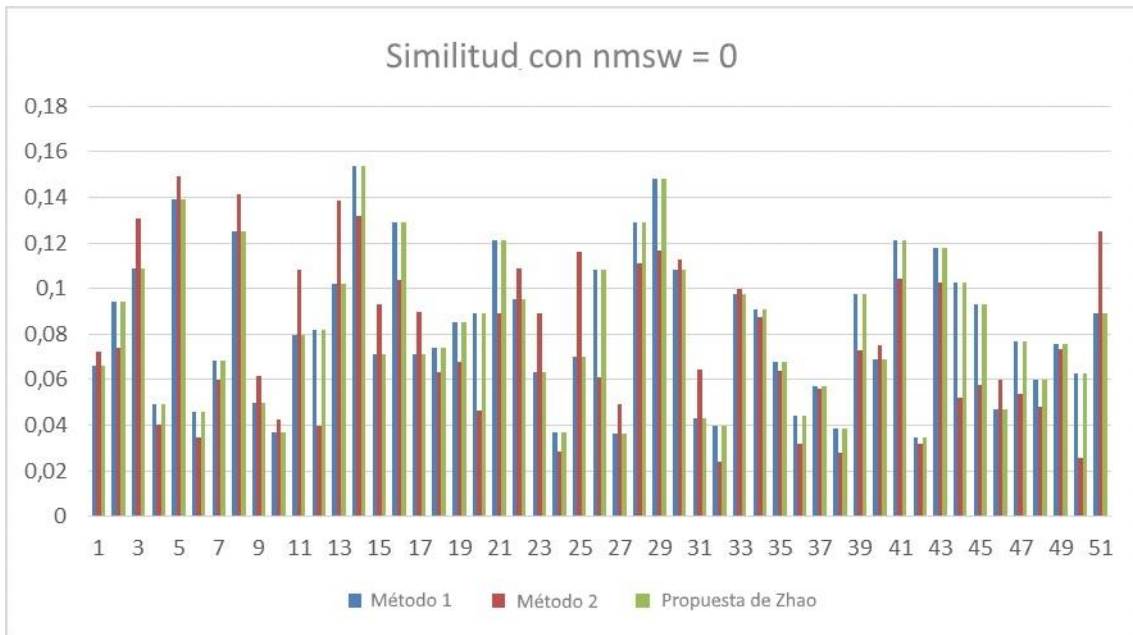
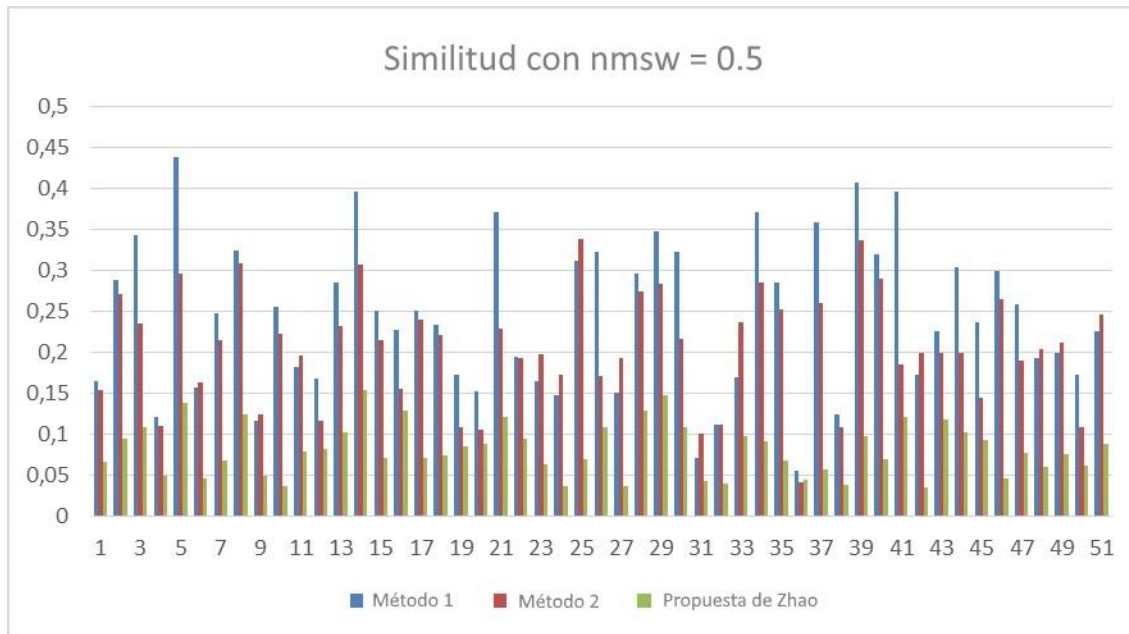
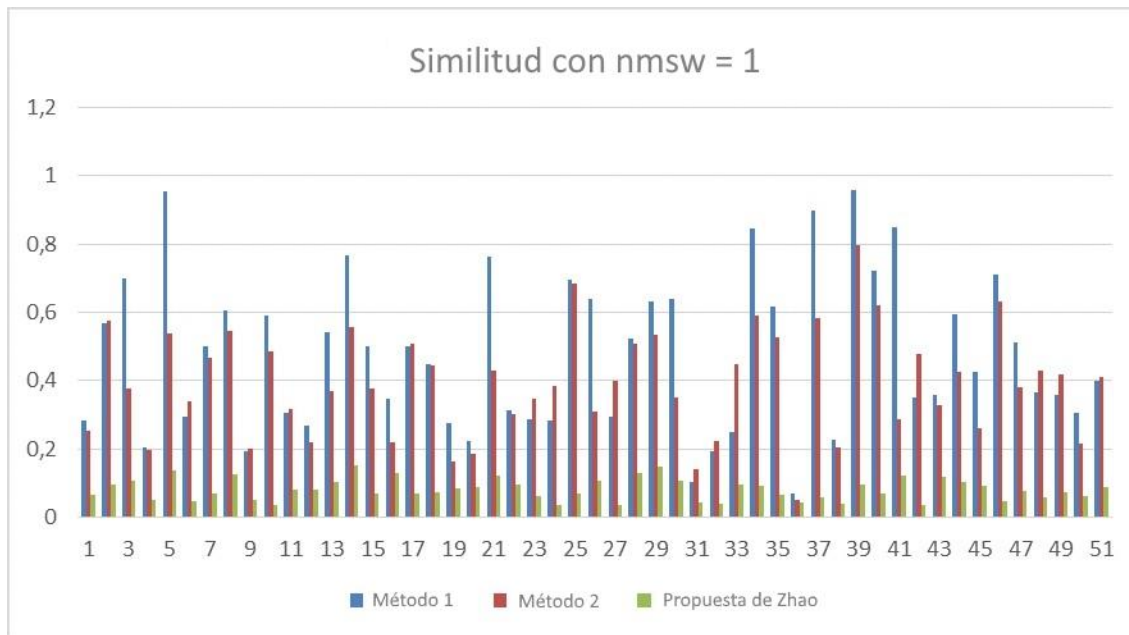


Figura 27. Comparación de resultados cuando  $nmsw = 0$

La Figura 28 muestra la similitud obtenida cuando  $nmsw = 0.5$  y la Figura 29 cuando  $nmsw = 1$ . Aquí, la similitud obtenida por los métodos 1 y 2 es mayor que la obtenida por la propuesta de Zhao puesto que esos valores afectan positivamente la similitud al considerar sitios que son de categorías similares, y como es de esperarse, cuando  $nmsw = 1$ , el valor de la similitud es mayor respecto a la propuesta de Zhao. En estos métodos la similitud es mayor cuando las trayectorias son más similares, debido a la relación entre los sitios (a través de las categorías), lo cual va de acuerdo con lo esperado. También es importante notar que, en la mayoría de los casos, el valor de la similitud del método 2 es inferior al del método 1, esto es determinado por la estructura y el peso de CTCA del método 2, el cual puede incrementar o disminuir la similitud del nodo raíz (nodo Sitio) mientras que en el método 1 todos los sitios tienen el mismo nivel de importancia para calcular la similitud.

Figura 28. Comparación de resultados cuando  $nmsw = 0.5$ Figura 29. Comparación de resultados cuando  $nmsw = 1$ 

El código y los datos utilizados para los experimentos puede ser accedidos a través del link: <https://www.dropbox.com/sh/q586exkc2cpdtgg/AADXSMJJvZ56ITCkI4IEKJjKa?dl=0>

### **3. Similitud de trayectorias basada en la cronología de los episodios**

La similitud entre trayectorias semánticas puede estar determinada por diversos criterios, tales como los sitios visitados, las actividades ejecutadas durante un *stop* (e.g., comer en un restaurante) o durante un *move* (e.g., escuchar música), el tiempo (e.g., mañana, tarde o noche) de ocurrencia de los eventos, la duración o el orden en que estos ocurren, entre otros. En particular, el orden (cronología) de los eventos puede indicar la secuencia de los sitios visitados (y de las actividades ejecutadas en cada uno) por el objeto móvil, e.g., una persona estuvo en una oficina trabajando, luego visitó un centro comercial donde hizo compras y al final del día fue a la casa a comer y a dormir.

El orden en que ocurren los eventos puede ser determinante a la hora de establecer la similitud entre las trayectorias de las personas, ya que estos pueden llegar a revelar sus gustos o su estilo de vida. Por ejemplo, supóngase que una persona X prefiere ir al gimnasio (a hacer deporte) antes de ir al trabajo, mientras que dos personas Y y Z prefieren ir al gimnasio (también a hacer deporte) después de ir al trabajo; considerando el orden de estos eventos, Y y Z deberían ser más similares que X con cualquiera de las otras dos, aun cuando las tres ejecutan las mismas actividades.

El tiempo cuando ocurren los eventos también puede ser considerado, ya que, aunque dos trayectorias pueden visitar un mismo conjunto de sitios (incluso en el mismo orden, i.e., una secuencia de sitios), uno lo puede hacer en la mañana y la otra en la noche. Esto los haría menos similares a si ambos los visitaran en la mañana o en la noche. También se puede considerar la duración de los eventos, ya que, si bien dos trayectorias pueden visitar una misma secuencia de sitios, uno pudo haber estado más tiempo en los sitios que la otra, lo cual disminuiría su similitud a si ambos se demorasen lo mismo en los sitios.

### 3.1 Motivación



Figura 30. Ejemplo de dos trayectorias

La Figura 30 muestra la secuencia de los sitios visitados (para efectos de este ejemplo, no es de interés diferenciar entre el tipo de sitio y un sitio en concreto, e.g., “Restaurante” y “Restaurante El Carmen”; este aspecto se analiza más adelante) por dos personas A y B. Como se observa, las dos personas visitaron los mismos (tipos de) sitios; sin embargo, B prefiere ejecutar actividades deportivas (ir al gimnasio y a la piscina) al inicio del día, mientras que A las ejecuta al final del día. Nótese que las dos personas tienen dos subsecuencias de (tipos de) sitios que visitaron en el mismo orden: <Universidad, Supermercado, Restaurante> y <Gimnasio, Piscina>, tal como se resalta en la Figura 31.



Figura 31. Subsecuencias en común de dos personas A y B

Lo anterior sugiere el establecimiento de una medida de similitud que considere la cronología de los eventos, i.e., que considere el orden en que se visitaron los sitios o el orden en que se ejecutaron las actividades (o ambos aspectos simultáneamente).

Una primera aproximación es identificar las subsecuencias de sitios en común de las trayectorias y dar mayor importancia (peso) a las subsecuencias de mayor longitud. La idea es que las trayectorias que tengan subsecuencias en común más largas deberían ser más similares que las que tengan subsecuencias en común más cortas, e.g., la subsecuencia <Hotel, Gimnasio, Playa, Restaurante, Bar> común entre dos personas puede indicar una gran similitud entre ellas, e.g., las dos podrían estar de vacaciones (e incluso se podría pensar que hacen parte de un mismo grupo de turistas) o ser empleados de una misma empresa que tienen en su jornada laboral asignada esta secuencia de sitios.

Por ejemplo, en la Figura 31, la subsecuencia de sitios en común más larga corresponde a la del rectángulo rojo, i.e., <Universidad, Supermercado, Restaurante>. Luego, se identifica la siguiente subsecuencia más larga de sitios en común (que no compartan sitios con la anterior); la cual corresponde a la subsecuencia del rectángulo verde, i.e., <Gimnasio, Piscina>. De esta forma, se tiene una subsecuencia de tres sitios y una de dos. Se procede a asignar un peso a cada subsecuencia para establecer una similitud entre las dos trayectorias semánticas. Sin embargo, esto no es una tarea trivial, ya que se deben considerar diversos factores:

- Entre más larga es una subsecuencia en común más peso debería tener. Sin embargo, no es lo mismo una subsecuencia en común de cuatro sitios cuando las trayectorias visitaron cuatro sitios en total cada una, que cuando las trayectorias visitaron, e.g., diez sitios en total cada una. Esto sugiere que el peso debe ser relativo al número de sitios en total visitados por las dos trayectorias. Además, se debe considerar que las dos trayectorias no necesariamente visitaron el mismo número de sitios, e.g., se puede presentar una subsecuencia en común de cuatro sitios cuando una trayectoria visitó diez sitios en total y la otra visitó solo cuatro.

Considerando lo anterior, en el siguiente ejemplo, la pareja de trayectorias (Ta, Tb) debería ser más similar que la pareja (Tc, Td); a pesar de que las dos parejas de trayectorias tienen la misma subsecuencia de sitios en común. Nota: Las secuencias de letras representan los sitios visitados por cada trayectoria.

Ta: <A, Z, G, H>

Tb: <A, Z, G, H>

Tc: <G, H, Z, Y, A, Z, G, H, Y, A, Q, L, Z, A, H>

Td: <S, Y, Q, H, L, P, B, A, Z, G, H, X, J, K, L>

- Puede ocurrir que dos trayectorias tengan varias subsecuencias en común de longitud corta (e.g., de dos sitios); se podría considerar que su similitud debería ser mayor que la de dos trayectorias que solo tienen una subsecuencia en común de longitud mayor (e.g., de cuatro sitios). En este tipo de situaciones la asignación de los pesos (por parte de los analistas) es determinante para establecer la similitud.

En el siguiente ejemplo, no es claro cuál pareja de trayectorias (Ta, Tb) y (Tc, Td) es más similar, ya que, aunque Ta y Tb tienen una subsecuencia en común larga (de longitud seis), Tc y Td tienen cinco subsecuencias en común más cortas (cada una de longitud dos).

Ta: <A, H, A, F, Q, A, G, D, A, Y>

Tb: <G, P, H, A, F, Q, A, G, D, S, G>

Tc: <Y, G, D, I, F, S, Y, H, L, S, R, Q, S, Y, T>

Td: <S, Y, G, T, P, I, F, R, Y, H, E, R, Q, P, I, Q, Y, T>

- Si hay varias subsecuencias de la misma longitud, la similitud debería ser mayor si estas subsecuencias se presentan en el mismo orden cronológico, e.g., si en una pareja de trayectorias (Ta, Tb) se presenta, en cada una, las subsecuencias: ...Sa...Sb... (donde Sa y Sb son dos subsecuencias de sitios en común), Ta y Tb deberían ser más similares que una pareja de trayectorias (Tc, Td) donde las subsecuencias ...Sa...Sb se presentan (en ese orden) en Tc y las subsecuencias ...Sb...Sa... se presentan (en ese orden) en Td. Según lo anterior, en el siguiente ejemplo, Ta y Tb deberían ser más similares que Tc y Td:

Ta: <Q, H, I, K, S, O, P, L, S>

Tb: <F, Q, H, I, M, O, P, L, A, S>

Tc: <K, W, O, P, L, G, K, Q, H, I>

Td: <Q, H, I, C, B, G, O, P, L, O, A>

- El número de sitios no comunes que separan a las subsecuencias en común también puede indicar una mayor o menor similitud entre dos trayectorias, puesto que entre menor sea el número de sitios no comunes visitados por los usuarios, más similares deberían ser. Se debería tener presente no solamente el número de sitios visitados entre las subsecuencias en común, sino también el tiempo transcurrido en las visitas.

En el ejemplo siguiente, Ta y Tb deberían ser menos similares que Tc y Td:

Ta: <Q, H, I, K, S, O, P, L, S>

Tb: <F, Q, H, I, M, W, X, Y, H, O, P, L, A, S>

Tc: <Q, H, I, K, S, O, P, L, S>

Td: <F, Q, H, I, M, O, P, L, A, S>

## 3.2 Métodos para identificar subsecuencias comunes

El proceso para identificar las subsecuencias en común también puede considerar diferentes métodos. Algunos métodos se proponen a continuación.

- Método 1: Sin eliminar subsecuencias: Se identifica la subsecuencia de sitios en común más larga. Por ejemplo, sean las trayectorias:

T1: <L, X, A, Z, Q, S, P, M>

T2: <A, Z, S, L, X, A, P, M>

Inicialmente, se identifica la subsecuencia de sitios más larga <L, X, A>. Se continúa el proceso, pero sin eliminar esta subsecuencia de las trayectorias.

Ahora, se busca la siguiente subsecuencia de sitios en común más larga. Aquí, se tienen dos opciones: a) sin permitir “solapes”, i.e., no permitir que una parte de la primera subsecuencia haga parte de la siguiente subsecuencia y b) permitir

“solapes”, i.e., permitir que una parte de la primera subsecuencia haga parte de la siguiente subsecuencia. A continuación se ejemplifican las dos opciones.

- a) Sin permitir “solapes”: Se identifica la subsecuencia <P, M> de longitud dos. Nótese que esta subsecuencia no comparte elementos con la subsecuencia <L, X, A>. En resumen, se obtienen dos subsecuencias en común: una subsecuencia de tres sitios y una de dos sitios.
- b) Permitir “solapes”: Se identifican dos subsecuencias en común de longitud dos: <A, Z> y <P, M>:

T1: <L, X, A, Z, Q, S, P, M>

T2: <A, Z, S, L, X, A, P, M>

Como ya no hay más subsecuencias en común que contengan al menos dos sitios, se finaliza el proceso. En resumen, se obtienen tres subsecuencias en común: una subsecuencia de tres sitios y dos subsecuencias de dos sitios.

El ejemplo anterior muestra que los dos métodos pueden generar diferentes resultados. Nótese además, que en la opción b), un mismo sitio puede hacer parte de varias subsecuencias, e.g., el tercer sitio de T1 (i.e., A) en <L, X, A> y <A, Z>.

- Método 2: Eliminar subsecuencias: Inicialmente, se identifica la subsecuencia de sitios en común más larga. Por ejemplo, sean las mismas trayectorias del ejemplo anterior:

T1: <L, X, A, Z, Q, S, P, M>

T2: <A, Z, S, L, X, A, P, M>

Inicialmente, se identifica la subsecuencia de sitios en común más larga (L, X, A). A continuación, se continúa el proceso, pero se elimina esta subsecuencia de las trayectorias:

T1: <L, X, A, Z, Q, S, P, M>      →      T1': <Z, Q, S, P, M>  
 T2: <A, Z, S, L, X, A, P, M>      →      T2': <A, Z, S, P, M>

Ahora, se busca de nuevo la subsecuencia de sitios en común más larga, i.e., <S, P, M> y se elimina de las trayectorias:

T1': <Z, Q, S, P, M>      →      T1'': <Z, Q>  
 T2': <A, Z, S, P, M>      →      T2'': <A, Z>

Como ya no hay más subsecuencias en común que contengan al menos dos sitios, se finaliza el proceso. En resumen, se obtienen dos subsecuencias de tres sitios cada una.

- Método 3: Permitir “huecos” (sitios en el medio): Esta idea se basa en el algoritmo de la subsecuencia en común más larga (LCS, *Longest Common Subsequence*), el cual encuentra la subsecuencia en común más larga a dos secuencias, no necesariamente compuesta por elementos consecutivos, pero sí en el mismo orden (Bergroth et al., 2000). Es decir, se encuentra una subsecuencia de sitios en común a ambas trayectorias, pero en la que cada objeto móvil pudo haber visitado sitios distintos (al del otro objeto móvil), i.e., “en el medio” de los sitios en común. A diferencia de los anteriores, este método genera una única subsecuencia en común. Por ejemplo, sean las mismas trayectorias del ejemplo anterior:

T1: <L, X, A, Z, Q, S, P, M>  
 T2: <A, Z, S, L, X, A, P, M>

Se busca la máxima subsecuencia formada por sitios en común, en el mismo orden, pero no necesariamente consecutivos, lo cual corresponde a la subsecuencia <L, X, A, P, M>:

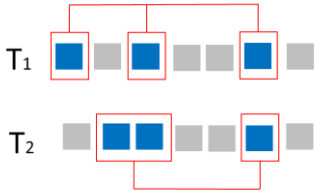
T1: <L, X, A, Z, Q, S, P, M>  
 T2: <A, Z, S, L, X, A, P, M>

Así, se tiene una subsecuencia de cinco sitios en común.

La Tabla 11 muestra los diferentes métodos ordenados según el nivel de exigencia en cuanto a la contigüidad de los sitios de las subsecuencias en común, donde el método más exigente (grado 1) requiere que ambas trayectorias tengan los mismos sitios en el mismo orden y sin otros sitios (“huecos”) entre ellos, mientras que el método más flexible (grado 4) permite “huecos” (sitios en el medio). La Tabla 11 ilustra subsecuencias en común entre dos trayectorias, donde los cuadros de un mismo color son sitios en común y los cuadros grises son sitios no comunes.

Tabla 11. Grado de dureza de los diferentes métodos

Método	Grado de exigencia de contigüidad	Descripción	Gráfico
1a: Sin eliminar secuencias y sin permitir solapes	1	Es el método más exigente ya que las subsecuencias en común se generan solo cuando las trayectorias visitan los mismos sitios y en el mismo orden. No se permite que las subsecuencias compartan sitios.	
1b: Sin eliminar secuencias y permitir solapes	2	Solo genera subsecuencias en común cuando las trayectorias visitan los mismos sitios y en el mismo orden. Se permite que las subsecuencias compartan sitios.	
2: Eliminar secuencias	3	Al eliminar las subsecuencias en común que se van generando, permite que se formen subsecuencias en común conformadas por sitios que estaban separados por las subsecuencias eliminadas.	

<p>3: Permitir "huecos"</p>	<p>4</p>	<p>Es el método menos exigente ya que permite que las trayectorias tengan sitios no comunes en medio de la subsecuencia en común, i.e., la subsecuencia en común se puede formar con sitios no necesariamente contiguos.</p>	
-----------------------------	----------	--	---

Nótese que, en algunos métodos, la cantidad de subsecuencias en común halladas entre dos trayectorias puede depender de la subsecuencia seleccionada en uno de los pasos del proceso iterativo cuando es posible seleccionar entre varias (i.e., cuando existen varias subsecuencias en común de la misma longitud). Considérese las siguientes dos trayectorias y el método 1a:

T3: <A, M, F, Q, R, C, I, C, X, C, A>

T4: <M, F, C, A, M, S, T, C, I, C>

En la primera iteración se encuentra la subsecuencia <C, I, C>

T3: <A, M, F, Q, R, C, I, C, X, C, A>

T4: <M, F, C, A, M, S, T, C, I, C>

En la segunda iteración se puede seleccionar cualquiera de las subsecuencias <A, M>, <M, F> y <C, A>. Por lo cual se pueden presentar los siguientes escenarios:

- Se selecciona <A, M>:

T3: <A, M, F, Q, R, C, I, C, X, C, A>

T4: <M, F, C, A, M, S, T, C, I, C>

Y el proceso finaliza acá puesto que no se encuentran más subsecuencias en común. Se tiene un total de una subsecuencia de longitud uno y una subsecuencia de longitud dos.

- Se selecciona <M, F>:

T3: <A, M, F, Q, R, C, I, C, X, C, A>

T4: <M, F, C, A, M, S, T, C, I, C>

En la tercera iteración se encuentra la subsecuencia <C, A>

T3: <A, M, F, Q, R, C, I, C, X, C, A>

T4: <M, F, C, A, M, S, T, C, I, C>

El proceso finaliza con una subsecuencia de longitud tres y dos subsecuencias de longitud dos.

- Se selecciona <C, A>:

T3: <A, M, F, Q, R, C, I, C, X, C, A>

T4: <M, F, C, A, M, S, T, C, I, C>

En la tercera iteración se encuentra la subsecuencia <M, F>

T3: <A, M, F, Q, R, C, I, C, X, C, A>

T4: <M, F, C, A, M, S, T, C, I, C>

El proceso finaliza con una subsecuencia de longitud tres y dos subsecuencias de longitud dos.

Como se puede observar en el ejemplo anterior, el resultado final del proceso puede variar según se escoja determinada subsecuencia. Por lo tanto, se sugiere utilizar como resultado el camino que más subsecuencias arroje. Nótese además que esta misma situación se puede presentar con los métodos 1b y 2.

### 3.3 Cálculo de la similitud temporal

Una vez halladas las subsecuencias en común de acuerdo con el método de preferencia elegido, surge el reto de definir una medida de similitud temporal, que permita establecer que tan semejantes son dos trayectorias respecto a la cronología de sus episodios. Por simplicidad, y con el objetivo de evaluar los métodos, se propone la siguiente medida de similitud. Como trabajo futuro, dicha medida puede ser expandida o redefinida con el fin de considerar todos los casos mencionados en la sección 3.1.

Sea

$$simT(T_1, T_2) = \sum w_i * (|s\_comun_i| / \min(|T_1|, |T_2|))$$

Ecuación 9

La similitud temporal entre las trayectorias  $T_1$  y  $T_2$ . Donde  $s\_comun_i$  es la secuencia de sitios en común en la iteración  $i$ , y  $w_i$  es un peso que depende de la iteración  $i$  así:

- Inicialmente  $w_i = 1$
- Cada vez que la cardinalidad de  $s\_comun$  disminuye,  $w_i$  disminuye en  $1/\min(|T_1|, |T_2|)$ , hasta un valor mínimo de  $w_i = 0$ .

Es decir, cada vez que se reduce la cantidad de sitios en común encontrados, se disminuye el peso  $w$ .

La definición de la medida de similitud propuesta y el factor de reducción del peso  $w$  son una propuesta base inicial que no tiene en cuenta todos los casos mencionados en este trabajo. La parametrización y calibración de estos es un trabajo futuro que requiere más exploración, ya que no es una tarea trivial y va más allá del alcance de este trabajo.

Retomando como ejemplo las trayectorias  $T_3$  y  $T_4$ , la Figura 32 muestra la similitud por cada uno de los posibles caminos.

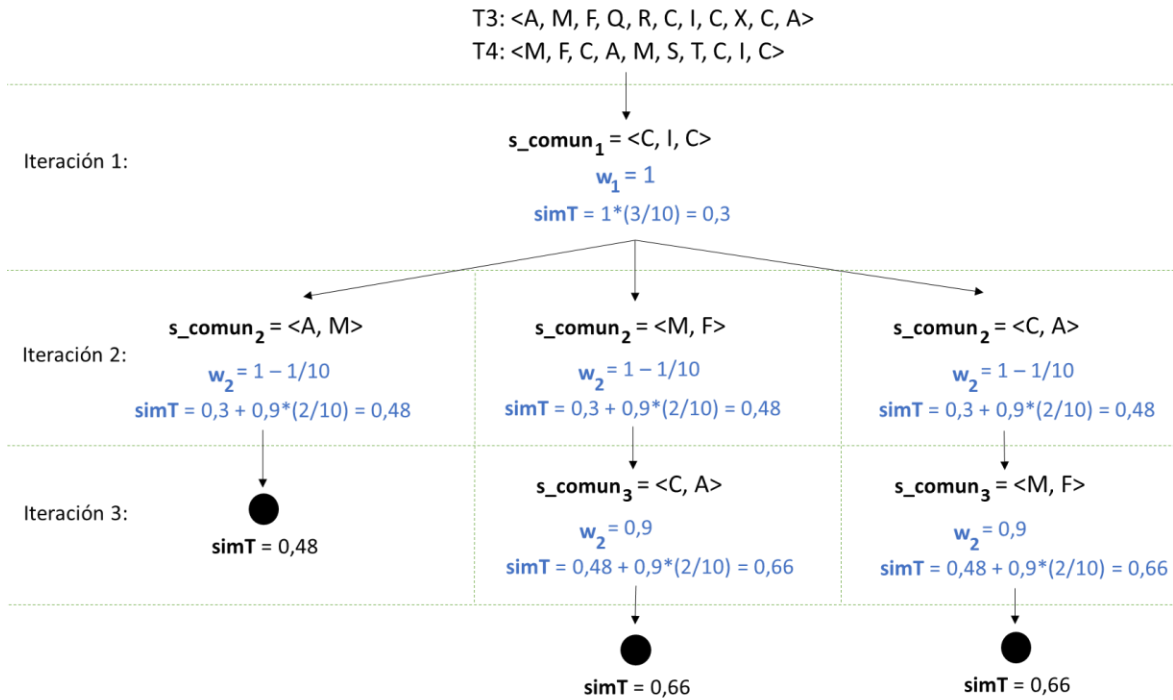


Figura 32. Similitud temporal por cada uno de los posibles caminos según el método 1a para T3 y T4

Como se puede observar, la similitud parcial en cada iteración depende del tamaño de la subsecuencia en común en relación con el tamaño de la trayectoria más pequeña (i.e., la cantidad de sitios comunes respecto a los sitios originales), y del peso  $w$ , el cual se reduce cada vez que el tamaño de la subsecuencia común se reduce. La similitud final está determinada por el máximo de las similitudes de cada uno de los posibles caminos, es decir,

$$simT(T_1, T_2) = 0,66$$

Nótese que cuando las trayectorias son iguales, o cuando una trayectoria es subconjunto de otra,  $simT = 1$ , lo cual es de esperarse ya que tienen todos los sitios posibles en común.

La definición anterior se puede aplicar a los métodos 1a, 1b y 2; sin embargo, el método 3 no es un proceso iterativo, ya que arroja una única subsecuencia común en su primera llamada, por lo cual se define la similitud así:

$$simT(T_1, T_2) = |s\_comun| / \min(|T_1|, |T_2|)$$

Utilizando la fórmula de ponderación actual propuesta, puede suceder que la similitud final por un camino sea mayor que 1 en el caso del método 1b, puesto que, al permitir solapes, se pueden generar secuencias de longitud mayor al total de sitios de la trayectoria más pequeña. Este caso se presenta de manera atípica y se debe establecer la similitud en un máximo de 1.

### 3.4 Algoritmo

A continuación, se muestran los algoritmos correspondientes a cada uno de los métodos de acuerdo con la medida de similitud definida en la sección anterior. El método 1a y 1b son en esencia los mismos, por lo que se presenta el método a y se indica las llamadas alternativas en el caso del método b.

#### Algoritmo 3. Algoritmo para los métodos 1a y 1b

---

metodo1a (T1, T2) (Alternativo metodo1b)

**Input:** T1, T2: Trayectorias

**Output:** Similitud entre T1 y T2 según el método 1a (Alternativo 1b)

**BEGIN**

1. return subsecuencia1a(T1, T2, null, 0, 1); //Inicialización de parámetros. Llamada alternativa a subsecuencia1b

**END** metodo1a

subsecuencia1a (T1, T2, Sa, sum, w) (Alternativo subsecuencia1b)

**Input:** T1, T2: Trayectorias, Sa: secuencia común anterior, sum: sumatoria acumulada, w: peso w de la iteración anterior

**Output:** Similitud entre T1 y T2 según el método 1a (Alternativo 1b)

**BEGIN**

1. S = maxSubsecuencias1a(T1, T2); //Ver nota. Llamada alternativa a maxSubsecuencias1b

2. maxim = sum; //Almacena el maximo de cada uno de los caminos

3. **FOREACH** s ∈ S

4. **IF** |s| < |Sa| **THEN**

5. w = w - 1 / min(|T1|, |T2|);

6. **END IF**

7. maxim = max(maxim, subsecuencia1a(T1, T2, s, sum + w \* (|s| / min(|T1|, |T2|)), w); //Llamada recursiva. Llamada alternativa a subsecuencia1b

8. **END FOR**

9. return maxim;

**END** subsecuencia1a

---

Nota: La función maxSubsecuencias1a retorna las subsecuencias en común más largas entre T1 y T2 que no hayan sido retornadas anteriormente, sin permitir que una subsecuencia sea un subconjunto de una ya retornada. La función maxSubsecuencias1b se comporta de manera similar, pero permite que algunos elementos de la subsecuencia común sean parte de una ya retornada, sin que sea un subconjunto de alguna de estas últimas.

#### Algoritmo 4. Algoritmo para el método 2

metodo2 (T1, T2)

**Input:** T1, T2: Trayectorias

**Output:** Similitud entre T1 y T2 según el método 2

**BEGIN**

1. return subsecuencia3(T1, T2, null, 0, 1);

**END** metodo2

subsecuencia2 (T1, T2, Sa, sum, w)

**Input:** T1, T2: Trayectorias, Sa: secuencia común anterior, sum: sumatoria acumulada, w: peso w de la iteración anterior

**Output:** Similitud entre T1 y T2 según el método 2

**BEGIN**

1. T1.eliminar(Sa); //Elimina los sitios de la subsecuencia Sa de la trayectoria T1

2. T2.eliminar(Sa);

3. S = maxSubsecuencias2(T1, T2); //Ver nota

4. maxim = sum;

5. **FOREACH** s ∈ S

6. **IF** |s| < |Sa| **THEN**

7. w = w - 1 / min(|T1|, |T2|);

8. **END IF**

9. maxim = max(maxim, subsecuencia2(T1, T2, s, sum + w \* (|s| / min(|T1|, |T2|)), w);

10. **END FOR**

11. return maxim;

**END** subsecuencia2

Nota: La función maxSubsecuencia2 retorna las subsecuencias en común más largas entre T1 y T2.

---

**Algoritmo 5. Algoritmo para el método 3**

---

metodo3 (T1, T2)

**Input:** T1, T2: Trayectorias

**Output:** Similitud entre T1 y T2 según el método 3

**BEGIN**

1.  $s = \text{lcs}(T1, T2)$ ; //Retorna la subsecuencia común más larga entre T1 y T2
2.  $\text{return } |s| / \min(|T1|, |T2|)$ ;

**END** metodo3

---

Las funciones `maxSubsecuencia1a`, `maxSubsecuencia1b` y `maxSubsecuencia2` vienen determinadas por el problema del substring común más largo, el cual es un algoritmo ya trabajado con solución óptima por programación dinámica. Las llamadas a esta función en los métodos 1a y 1b suponen variaciones menores en el algoritmo para obtener el comportamiento deseado. De manera similar, la función `lcs` viene determinada por el problema de la subsecuencia común más larga.

### 3.5 Niveles de análisis

Cada uno de los episodios de una trayectoria está conformado por el sitio visitado y la actividad ejecutada por el usuario, además de la marca de tiempo. Tanto los sitios como las actividades están directamente relacionadas con un nodo hoja del CTCS y CTCA, como se muestra en la Figura 33. Los métodos presentados en este capítulo pueden ser ejecutados para hallar la similitud tanto respecto a sitios como a actividades; además pueden ser llamados al nivel deseado del árbol de categorías, i.e., se puede encontrar la similitud temporal respecto a sitios (o actividades) a nivel de sitios concretos (e.g., Restaurante La sazón, Universidad Nacional) o a nivel de categorías, en cualquiera de los niveles del árbol (e.g., Universidad, Educación).

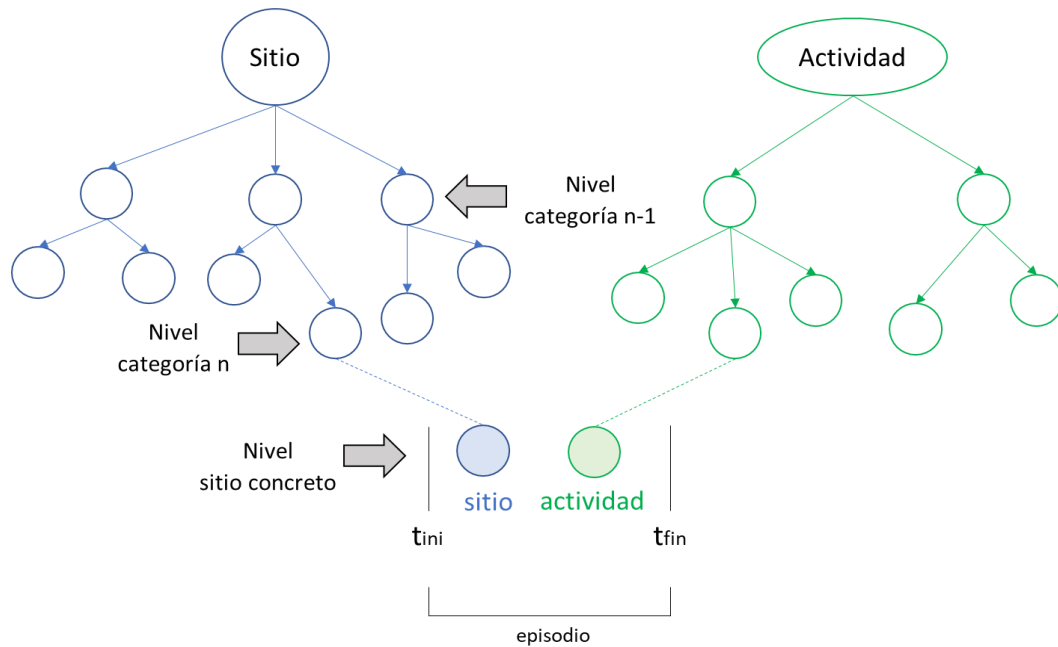


Figura 33. Relación de un episodio con el CTCS y CTCA

Por ejemplo, sean las trayectorias:

T5: <Coffe House, The 90's, Escuela Las Lomitas, El Sazón Paisa>

T6: <BodyGym, Café del suroeste, Rock and Roll Bar, Universidad Nacional>

A nivel de sitios concretos T5 y T6 no poseen secuencias en común, pero si subimos un nivel en el árbol de categorías, se podría presentar que:

T5: <Café, Bar, Escuela, Restaurante>

T6: <Gimnasio, Café, Bar, Universidad>

Lo cual nos arroja una secuencia común de longitud dos <Café, Bar>. Y si subimos aún más en el árbol se podría presentar que:

T5: <Entretenimiento, Entretenimiento, Educación, Restaurante>

T6: <Deporte, Entretenimiento, Entretenimiento, Educación>

Arrojando una secuencia en común de longitud tres <Entretención, Entretención, Educación>.

De esta forma, es posible encontrar la similitud temporal entre dos trayectorias de acuerdo con los intereses del analista y el dominio de aplicación. Es evidente que entre más “se suba” en el árbol, mayor será la similitud entre las trayectorias, puesto que se promueve la generación de nuevas secuencias en común. Nótese que si se llama el algoritmo a nivel raíz del árbol (Sitio o Actividad) la similitud entre cualquier par de trayectorias sería 1.

### 3.6 Similitud multidimensional

Furtado et al. (2015) define una medida de similitud multidimensional que considera la distancia entre dos conjuntos de elementos en diferentes dimensiones, basado en un puntaje entre dos elementos a y b como se define en la Ecuación 10. En el caso particular de las trayectorias, esos elementos son episodios.

$$score(a, b) = \sum_{k=1}^{|D|} (match_k(a, b) * w_k)$$

Ecuación 10

Donde D es un conjunto de dimensiones,  $w_k$  es el peso asignado a cada dimensión, y  $match_k(a, b)$  está dado por la Ecuación 11.

$$match_k(a, b) = \begin{cases} 1, & \text{si } dist_k(a, b) \leq maxDist_k \\ 0, & \text{en caso contrario} \end{cases}$$

Ecuación 11

Donde  $maxDist_k$  es un umbral de distancia para la dimensión k. En este sentido, diferentes dimensiones pueden ser consideradas, como tiempo, espacio y semántica, para calcular la similitud entre trayectorias como se muestra en la Figura 34. El foco del trabajo de Furtado es definir una medida de similitud multidimensional, pero no la función de similitud de cada dimensión; sin embargo, a manera de ejemplo, ellos definen una medida de similitud muy simple para la dimensión semántica, la cual está dado por la Ecuación 12.

$$dist_k(a, b) = \begin{cases} 0, & \text{si } a.tipo = b.tipo \\ 1, & \text{en caso contrario} \end{cases}$$

Ecuación 12

Esto es, la similitud entre dos episodios  $a$  y  $b$  es 0 (coincidencia completa) si los episodios tienen el mismo tipo de sitio, o 1 en caso contrario. En esta tesis se propone una nueva medida para la dimensión semántica basada en los sitios visitados y las actividades ejecutadas allí, definiendo la medida de similitud para la dimensión semántica entre dos trayectorias, i.e.,  $dist_k$  para  $k = \text{semántica}$ . También se define una medida de similitud temporal basada en el orden de los episodios para  $k = \text{temporal}$ , como se resalta en la Figura 34.

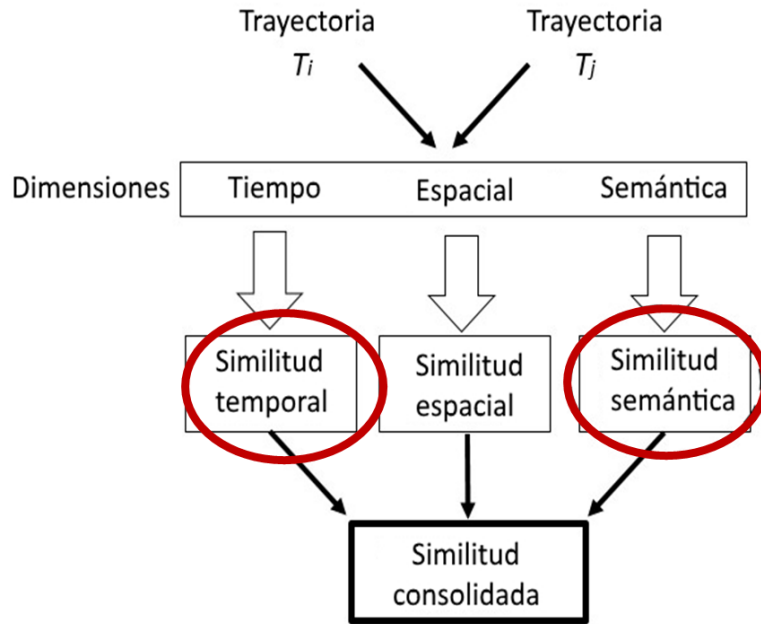


Figura 34. Foco de esta tesis

La paridad entre dos trayectorias  $A$  y  $B$  se define según la Ecuación 13 como el máximo *score* de todos los episodios de  $A$  comparados con todos los episodios de  $B$ .

$$parity(A, B) = \sum_{a \in A} \max\{score(a, b) : b \in B\}$$

Ecuación 13

La similitud multidimensional se define como el promedio de la paridad de A con B y B con A, según la Ecuación 14.

$$MSM(A, B) = \begin{cases} 0, & \text{si } |A| = 0 \text{ o } |B| = 0 \\ \frac{parity(A, B) + parity(B, A)}{|A| + |B|}, & \text{en caso contrario} \end{cases}$$

Ecuación 14

De esta forma, se puede utilizar como medida de similitud para la dimensión semántica el valor obtenido por alguno de los métodos respecto a sitios o actividades del capítulo 2, y como valor para la dimensión temporal el valor obtenido por la medida de similitud de cronología de este capítulo.

### 3.7 Experimentos

Para los experimentos, se utilizaron de nuevo los mismos datos y el mismo prototipo del capítulo 2, i.e., el registro de usuarios de Foursquare en New York entre octubre 24 de 2011 y febrero 20 de 2012. Se seleccionaron 50 usuarios y ejecutaron cada uno de los cuatro métodos para cada posible par de trayectorias. Se realiza la suposición que el orden en que aparecen los episodios en dichos datos es el orden en que los usuarios visitaron los sitios, ya que los mismos no cuentan con marcas de tiempo.

La matriz presentada en la Figura 35 muestra la similitud para los 10 primeros pares de trayectorias según el método 1a. La matriz es simétrica puesto que  $simT(T1, T2) = simT(T2, T1)$  y la diagonal es igual a 1 ya que  $simT(T1, T1) = 1$ . De manera similar, la Figura 36 presenta la matriz de similitud para el método 1b, la Figura 37 para el método 2 y la Figura 38 para el método 3.

Método 1a										
	570	14783	18179	20423	24915	29921	36865	39619	41376	45010
570	1	0,21	0,57	0,48	0,64	0,43	0,43	0,68	0,57	0,42
14783		1	0,29	0,35	0,74	0,14	0,43	0,35	0	0,29
18179			1	0,4	0,68	0,42	0,38	0,8	0,42	0,42
20423				1	0,57	0,65	0,39	0,65	0,57	0,48
24915					1	0,53	0,61	0,58	0,5	0,48
29921						1	0,42	0,48	0,43	0,49
36865							1	0,46	0,36	0,14
39619								1	0,34	0,68
41376									1	0,35
45010										1

Figura 35. Resultados del método 1a

Método 1b										
	570	14783	18179	20423	24915	29921	36865	39619	41376	45010
570	1	0,35	0,57	0,68	0,99	0,57	0,56	0,92	0,71	0,42
14783		1	0,43	0,35	1	0,14	0,57	0,69	0	0,29
18179			1	0,4	0,98	0,66	0,38	0,87	0,42	0,61
20423				1	0,7	0,9	0,39	0,72	0,57	0,48
24915					1	0,81	0,83	0,73	0,77	0,69
29921						1	0,42	0,67	0,77	0,62
36865							1	0,56	0,42	0,14
39619								1	0,34	0,81
41376									1	0,48
45010										1

Figura 36. Resultados del método 1b

Método 2										
	570	14783	18179	20423	24915	29921	36865	39619	41376	45010
570	1	0,35	0,62	0,61	0,78	0,57	0,43	0,8	0,57	0,42
14783		1	0,43	0,35	0,94	0,14	0,5	0,35	0	0,29
18179			1	0,59	0,74	0,54	0,55	0,88	0,54	0,55
20423				1	0,91	0,77	0,52	0,65	0,57	0,48
24915					1	0,58	0,67	0,79	0,59	0,61
29921						1	0,42	0,48	0,67	0,76
36865							1	0,66	0,36	0,14
39619								1	0,34	0,68
41376									1	0,35
45010										1

Figura 37. Resultados del método 2

Método 3										
	570	14783	18179	20423	24915	29921	36865	39619	41376	45010
570	1	0,5	0,57	0,57	0,79	0,57	0,71	0,79	0,64	0,5
14783		1	0,43	0,36	0,86	0,43	0,64	0,57	0,43	0,29
18179			1	0,47	0,69	0,56	0,62	0,69	0,56	0,57
20423				1	0,73	0,6	0,47	0,67	0,47	0,43
24915					1	0,55	0,58	0,55	0,48	0,79
29921						1	0,47	0,55	0,45	0,57
36865							1	0,53	0,53	0,43
39619								1	0,45	0,64
41376									1	0,57
45010										1

Figura 38. Resultados del método 3

Nótese que los resultados del método 1b y 2 son iguales o mayores a los resultados del método 1a para todos los casos, lo cual es un comportamiento esperado puesto que los métodos 1b y 2 pueden promover la generación de nuevas subsecuencias comunes al eliminar sitios en cada iteración y al permitir solapes respectivamente.

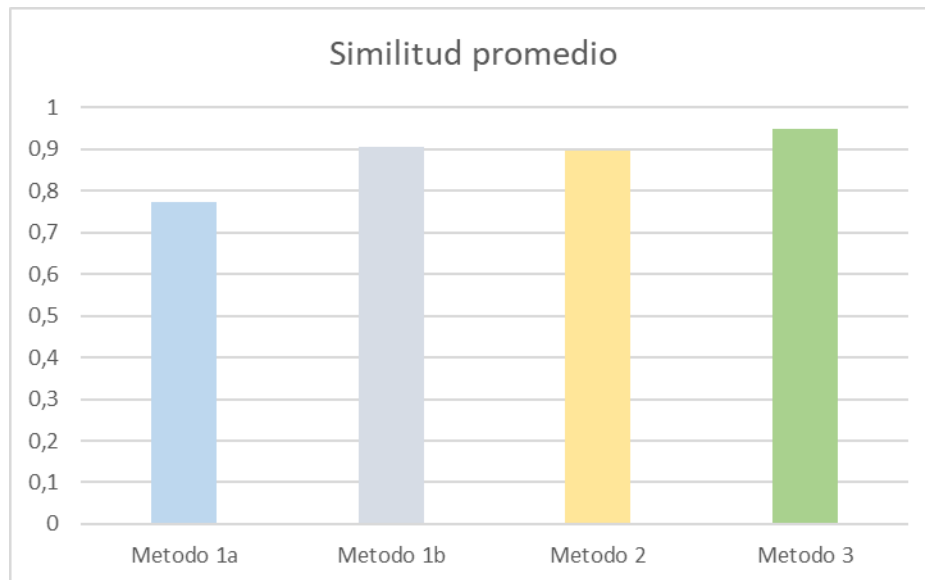


Figura 39. Similitud promedio para cada uno de los métodos

La Figura 39 muestra la similitud promedio de cada uno de los métodos, para los 50 pares de trayectorias. El método con que mayor similitud se obtiene es el 3 puesto que al permitir “huecos” suele obtener una secuencia común bastante larga en la primera y única iteración, además que a diferencia de los demás métodos no se “penaliza” por cada iteración. Con los métodos 1b y 2 se tienen resultados muy similares y mayores a los

resultados del método 1a, puesto que con los dos primeros se pueden formar nuevas subsecuencias por naturaleza de estos, lo cual no ocurre en el método 1a que es el más rígido. Los resultados también son acordes a lo esperado según el nivel de dureza de los métodos que se explica en la Tabla 11.

El código y los datos utilizados para los experimentos puede ser accedidos a través del link: <https://www.dropbox.com/sh/q586exkc2cpdtgg/AADXSMJJvZ56ITCkl4IEKJjKa?dl=0>

## **4. Conclusiones y recomendaciones**

### **4.1 Conclusiones**

Las trayectorias representan gustos y preferencias de los usuarios, y su análisis es de gran interés en áreas como marketing, turismo y redes sociales. Diferentes métodos para encontrar similitud entre trayectorias crudas han sido propuestos; sin embargo, pocos trabajos han considerado la similitud entre trayectorias semánticas, las cuales poseen información de más alto nivel que permite análisis más valiosos.

En esta tesis se ha propuesto una medida de similitud para trayectorias semánticas de objetos móviles que considera tanto sitios visitados como actividades ejecutadas. Esta propuesta incluye dos métodos para calcular la similitud y es flexible porque permite al analista definir sus propios arboles de categorías para la clasificación de sitios y actividades. Además de esto, en el método 2 es posible asignar pesos a los nodos de los arboles con el fin de establecer su importancia en el cálculo de la similitud.

En el aspecto temporal se tiene un gran conjunto de posibles criterios a considerar, añadiendo un nivel de complejidad importante al cálculo de la similitud entre dos trayectorias, más aún cuando se considera tanto sitios como actividades. En esta tesis se selecciona el aspecto de cronología de los episodios, es decir, el orden en que suceden los eventos.

Se definen cuatro métodos diferentes para calcular la similitud basada en la cronología y se propone una fórmula básica de ponderación de subsecuencias. Esta propuesta permite seleccionar al analista el método que más se ajuste a sus intereses, ofreciendo flexibilidad y adaptación al dominio de aplicación.

## 4.2 Cumplimiento de objetivos

A continuación, se relaciona el objetivo general y los objetivos específicos de la propuesta, y la forma como se cumplieron en esta tesis:

**Objetivo general:** *Proponer un método para el cálculo de la similitud entre dos trayectorias semánticas basado en los sitios visitados, las actividades ejecutadas y la cronología de los episodios (stops).*

En el capítulo 2 se hace una propuesta de un método para encontrar la similitud de trayectorias considerando los sitios visitados por el usuario y las actividades ejecutadas a través de un árbol de categorías para la clasificación de sitios (CTCS) y actividades (CTCA). En el capítulo 3 se incorporan elementos relacionados con la cronología de los episodios, permitiendo considerar el orden en que los usuarios visitan los sitios y ejecutadas las actividades.

### **Objetivos específicos:**

*1. Identificar los métodos para el cálculo de similitud en trayectorias semánticas y determinar las ventajas y desventajas de cada uno.*

En el capítulo 1 se analizan diferentes trabajos recientes relacionados con la similitud de trayectorias, se analiza cada uno y se identifican las falencias halladas. Además se analizan con mayor profundidad cuatro de los trabajos más relevantes del dominio.

*2. Desarrollar un método para el cálculo de la similitud entre dos trayectorias semánticas basado en los sitios visitados y las actividades ejecutadas.*

En el capítulo 2 se propone un método para calcular la similitud de trayectorias basada en los sitios visitados, las actividades ejecutadas, o ambos criterios. Dos variantes del método son propuestas para ser usadas según criterio del analista.

*3. Extender el método desarrollado en el objetivo anterior considerando además la cronología de los episodios.*

En el capítulo 3 se incorporan criterios temporales relacionados con el orden de ocurrencia de los eventos, los cuales pueden ser analizados en conjunto con los sitios visitados y las actividades ejecutadas

4. *Desarrollar un prototipo y evaluar el método desarrollado frente a otro método.*

En los capítulos 2 y 3 se realizan diversos experimentos a partir de datos recolectados anónimamente de Foursquare y se analizan los resultados frente a otros métodos.

### 4.3 Recomendaciones

La medida de similitud temporal propuesta en esta tesis es un punto de partida inicial y considera los aspectos básicos, sin embargo, se hace necesario más investigación en este campo con el objetivo obtener un comportamiento más similar al esperado. También es primordial considerar otros criterios temporales, ya que, si bien la cronología es un criterio importante en el cálculo de la similitud entre usuarios, no es el único ni el más determinante, por lo que debe ser combinado con otros criterios como la duración, frecuencia, momentos de tiempo, entre otros. Los diferentes parámetros que el analista puede ajustar le dan flexibilidad y mayor aplicación a los métodos propuestos, pero también incrementan el grado de complejidad de los mismos, por lo cual se hace importante encontrar unos valores óptimos para los mismos según diferentes intereses o dominios. De manera adicional, la eficiencia de los métodos propuestos puede ser mejorada considerablemente mediante optimización y paralelismo, ya que cuando se cuenta con una gran cantidad de datos generados por los usuarios, la eficiencia se reduce de manera significativa debido a la cantidad de caminos y combinaciones posibles, en especial cuando aplica como función para *clustering*.

### 4.4 Difusión

A la fecha de entrega de esta tesis, se cuenta con los siguientes elementos de difusión:

- Congreso internacional

Título: An algorithm for trajectory semantic similarity

Autores: Francisco Javier Moreno, Santiago Román Fernández, Vania Borgorny

Congreso: Mathematical modelling in engineering & human behavior

Año: 2015

Ubicación: Valencia, España

- Capítulo de libro

Título: Trajectories similarity: A proposal and some problems

Autores: Francisco Javier Moreno, Santiago Román Fernández, Vania Borgorny

Libro: Modelling Human Behavior: Individuals and Organizations

Año: 2016

ISBN: 978-1-53610-216-1

- Artículo publicado en revista internacional

Título: Towards a Semantic Trajectory Similarity Measuring

Autores: Francisco Javier Moreno, Santiago Román Fernández, Vania Borgorny

Revista: Indian journal of science and technology

Volumen 10, Issue 18, mayo 2017

Categoría A2 Colciencias

- Artículo publicado en revista internacional

Título: Semantic Trajectory Similarity based on Chronological Aspects

Autores: Francisco Javier Moreno, Santiago Román Fernández, Jaime Alberto Guzmán

Revista: Indian journal of science and technology

Volumen 11, Issue 32, agosto 2018

Categoría A2 Colciencias

- Artículo sometido en revista nacional

Título: Una aproximación a la similitud de trayectorias

Revista: Revista Ingenierías Universidad de Medellín

## A. Anexo: Demostración de que $C_{ns,T_i,T_j,nmsw} \in [0,1]$

**Demostración:** 
$$C_{ns,T_i,T_j,nmsw} = \frac{|POI_{ns,T_i} \cap POI_{ns,T_j}| + nmsw * nnms}{|POI_{ns,T_i} \cup POI_{ns,T_j}| - nmsw * nnms} \in [0,1]$$

Sea  $n = |POI_{ns,T_i}|$  y  $m = |POI_{ns,T_j}|$

- Si  $POI_{ns,T_i} \cap POI_{ns,T_j} = \emptyset$  entonces  $C_{ns,T_i,T_j,nmsw} = \frac{nmsw * nnms}{n+m - nmsw * nnms}$  y  $nnms$  puede ser  $n$  o  $m$  dependiendo de cuál conjunto ( $POI_{ns,T_i}$  or  $POI_{ns,T_j}$ ) es más grande.

Suponga que  $|POI_{ns,T_j}| > |POI_{ns,T_i}|$  entonces  $nnms = n$  y  $C_{ns,T_i,T_j,nmsw} = \frac{nmsw * n}{n(1-nmsw)+m}$ ,  $0 \leq nmsw \leq 1$ .

Si  $nmsw = 0$  entonces  $C_{ns,T_i,T_j,nmsw} = 0$  y si  $nmsw = 1$  entonces  $C_{ns,T_i,T_j,nmsw} = \frac{n}{m} < 1$  ya que  $n < m$ .

Similarmente cuando  $|POI_{ns,T_j}| < |POI_{ns,T_i}|$

- Si  $POI_{ns,T_i} \cap POI_{ns,T_j} = POI_{ns,T_i}$  entonces  $nnms = 0$  ya que  $POI_{ns,T_i} - POI_{ns,T_j} = \emptyset$   
 $C_{ns,T_i,T_j,nmsw} = \frac{n}{m} < 1$  ya que  $n < m$ .

Similarmente cuando  $POI_{ns,T_i} \cap POI_{ns,T_j} = POI_{ns,T_j}$

- Si  $POI_{ns,T_i} \cap POI_{ns,T_j} = X$  cuando  $X$  es un subconjunto de  $POI_{ns,T_i}$  and  $POI_{ns,T_j}$  y  $x = |X|$

Cuando  $nmsw = 0$ ,  $C_{ns,T_i,T_j,nmsw}$  es la definición del índice de Jaccard, el cual  $\in [0, 1]$ .

Cuando  $nmsw = 1$ ,  $C_{ns,T_i,T_j,nmsw} = \frac{x+nnms}{n+m-x-nnms}$  y  $nnms = n - x$  o  $nnms = m - x$  dependiendo de cuál conjunto ( $POI_{ns,T_i}$  or  $POI_{ns,T_j}$ ) es más grande.

Suponga que  $|POI_{ns,T_j}| > |POI_{ns,T_i}|$  entonces  $nnms = n - x$  y  $C_{ns,T_i,T_j,nmsw} = \frac{n}{m} < 1$  ya que  $n < m$ .

Similarmente cuando  $|POI_{ns,T_j}| < |POI_{ns,T_i}|$ .

## Bibliografía

- Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., & Vaisman, A. (2007). A Model for Enriching Trajectories with Semantic Geographical Information. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems* (p. 22:1--22:8). New York, NY, USA: ACM. <https://doi.org/10.1145/1341012.1341041>
- Bergroth, L., Hakonen, H., & Raita, T. (2000). A Survey of Longest Common Subsequence Algorithms. In *Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00)* (p. 39). Washington, DC, USA: IEEE Computer Society.
- Bogorny, V., Renso, C., de Aquino, A. R., de Lucca Siqueira, F., & Alvares, L. O. (2014). CONSTAnT – A Conceptual Data Model for Semantic Trajectories of Moving Objects. *Transactions in GIS*, 18(1), 66–88. <https://doi.org/10.1111/tgis.12011>
- Cao, H., Wolfson, O., & Trajcevski, G. (2006). Spatio-temporal Data Reduction with Deterministic Error Bounds. *The VLDB Journal*, 15(3), 211–228. <https://doi.org/10.1007/s00778-005-0163-7>
- Chang, J.-W., Bista, R., Kim, Y.-C., & Kim, Y.-K. (2007). Spatio-temporal Similarity Measure Algorithm for Moving Objects on Spatial Networks. In O. Gervasi & M. L. Gavrilova (Eds.), *Computational Science and Its Applications -- ICCSA 2007* (pp. 1165–1178). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Chen, L., Özsu, M. T., & Oria, V. (2005). Robust and Fast Similarity Search for Moving Object Trajectories. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (pp. 491–502). New York, NY, USA: ACM. <https://doi.org/10.1145/1066157.1066213>
- Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. *Icwsn, 2010(Cholera)*, 81–88.
- Furtado, A. S., Kopanaki, D., Alvares, L. O., & Bogorny, V. (2015). Multidimensional Similarity Measuring for Semantic Trajectories. *Transactions in GIS*.

- Keogh, E., & Ratanamahatana, A. C. (2004). Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3), 358–386. <https://doi.org/10.1007/s10115-004-0154-9>
- Kondaveeti, A. (2012). *Spatio-Temporal Data Mining to Detect Changes and Clusters in Trajectories*.
- Kreveld, M. Van, & Luo, J. (2007). Trajectory Similarity of Moving Objects. *Young Researcher Forum*, 229–232.
- Kruskal, J. B. (1983). An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules. *SIAM Review*, 25(2), 201–237. Retrieved from <http://www.jstor.org/stable/2030214>
- Lee, J.-G., Han, J., & Whang, K.-Y. (2007). Trajectory Clustering: A Partition-and-group Framework. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data* (pp. 593–604). New York, NY, USA: ACM. <https://doi.org/10.1145/1247480.1247546>
- Lee, M.-J., & Chung, C.-W. (2011). Database Systems for Advanced Applications: 16th International Conference, DASFAA 2011, Hong Kong, China, April 22-25, 2011, Proceedings, Part I. In J. X. Yu, M. H. Kim, & R. Unland (Eds.) (pp. 38–52). Berlin, Heidelberg: Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-20149-3\\_5](https://doi.org/10.1007/978-3-642-20149-3_5)
- Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W.-Y. (2008). Mining User Similarity Based on Location History. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 34:1--34:10). New York, NY, USA: ACM. <https://doi.org/10.1145/1463434.1463477>
- Li, Z., Kays, R., & Nye, P. (2010). Mining Periodic Behaviors for Moving Objects. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1099–1108). Washington, DC, USA.
- Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News: The Newsletter of the R Project*, 2(3), 18–22.
- Liu, H., & Schneider, M. (2012). Similarity Measurement of Moving Object Trajectories. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming* (pp. 19–22). New York, NY, USA: ACM. <https://doi.org/10.1145/2442968.2442971>
- Moreno, F. J., Bogorny, V., & Román, S. (2017). Towards a Semantic Trajectory Similarity Measuring. *Indian Journal of Science & Technology*, 10(18), 1–14.

- Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., ... Yan, Z. (2013). Semantic Trajectories Modeling and Analysis. *ACM Computing Surveys (CSUR)*, 45(4), 42:1--42:32. <https://doi.org/10.1145/2501654.2501656>
- Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). Chapter 1 - Challenges of Sentiment Analysis in Social Networks: An Overview. In F. A. Pozzi, E. Fersini, E. Messina, & B. Liu (Eds.), *Sentiment Analysis in Social Networks* (pp. 1–11). Boston: Morgan Kaufmann. <https://doi.org/https://doi.org/10.1016/B978-0-12-804412-4.00001-2>
- Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., & Vangenot, C. (2008). A Conceptual View on Trajectories. *Data & Knowledge Engineering*, 65(1), 126–146. <https://doi.org/10.1016/j.datak.2007.10.008>
- Tiakas, E., Papadopoulos, A. N., Nanopoulos, A., Manolopoulos, Y., Stojanovic, D., & Djordjevic-Kajan, S. (2009). Searching for similar trajectories in spatial networks. *Journal of Systems and Software*, 82(5), 772–788. <https://doi.org/http://dx.doi.org/10.1016/j.jss.2008.11.832>
- Trajcevski, G., Ding, H., Scheuermann, P., Tamassia, R., & Vaccaro, D. (2007). Dynamics-aware Similarity of Moving Objects Trajectories. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems* (p. 11:1--11:8). New York, NY, USA: ACM. <https://doi.org/10.1145/1341012.1341027>
- Vlachos, M., Kollios, G., & Gunopulos, D. (2002). Discovering similar multidimensional trajectories. In *Data Engineering, 2002. Proceedings. 18th International Conference on* (pp. 673–684). <https://doi.org/10.1109/ICDE.2002.994784>
- Xiao, X., Zheng, Y., Luo, Q., & Xie, X. (2012). Inferring social ties between users with human location history. *Journal of Ambient Intelligence and Humanized Computing*, 5(1), 3–19. <https://doi.org/10.1007/s12652-012-0117-z>
- Yanagisawa, Y., Akahani, J., & Satoh, T. (2003). Shape-Based Similarity Query for Trajectory of Mobile Objects. In *Proceedings of the 4th International Conference on Mobile Data Management* (pp. 63–77). London, UK, UK: Springer-Verlag. Retrieved from <http://dl.acm.org/citation.cfm?id=648060.747275>
- Ying, J. J.-C., Lu, E. H.-C., Lee, W.-C., Weng, T.-C., & Tseng, V. S. (2010). Mining User Similarity from Semantic Trajectories. In *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks* (pp. 19–26). New York,

NY, USA: ACM. <https://doi.org/10.1145/1867699.1867703>

Zhao, H., Han, Q., Pan, H., & Yin, G. (2009). Spatio-temporal Similarity Measure for Trajectories on Road Networks. In *Internet Computing for Science and Engineering (ICICSE), 2009 Fourth International Conference on* (pp. 189–193).

<https://doi.org/10.1109/ICICSE.2009.18>

Zhao, X. (2011). Progressive refinement for clustering spatio-temporal semantic trajectories. In *Proceedings of 2011 International Conference on Computer Science and Network Technology, ICCSNT 2011* (Vol. 4, pp. 2695–2699). Harbin, China:

IEEE. <https://doi.org/10.1109/ICCSNT.2011.6182522>