Detección de patrones de movilidad en datos de localización aplicando modelos de tópicos

Manuel Arturo Suárez-Améndola[1], Salvador Ruiz-Correa[2], Rogelio Hasimoto-Beltrán[3]

[1] Instituto Tecnológico Superior de Escárcega, México

[2] Center for Mobile Life Studies, México

[3] Centro de Investigación en Matemáticas A.C., México

masuareb@hotmail.com, src@cmls.pw, hasimoto@cimat.mx

Resumen. De manera natural las personas vamos formando rutinas para adaptarnos al entorno social en el cuál nos desempeñamos. Estas rutinas se ven reflejadas en las secuencias de ubicaciones que visitamos para desplazarnos hacia los lugares donde realizamos nuestras actividades diarias, sin embargo, ¿cómo podemos determinar la presencia de ellas? El presente trabajo consiste en un análisis exploratorio de un conjunto de datos de movilidad de usuarios utilizando modelos de tópicos con el fin de determinar si existe una estructura subyacente en la movilidad de las personas de acuerdo a las rutinas que éstas manifiestan en sus vidas. Se presenta la introducción al problema explicando el modelo a utilizar así como su desarrollo teórico, validación experimental de su funcionamiento a través de conjuntos de datos sintéticos y su aplicación a un conjunto de datos real con presentación de los resultados obtenidos y conclusiones.

Palabras clave: reconocimiento de patrones, análisis exploratorio, datos móviles, modelos de tópicos.

1. Introducción

Es interesante observar cómo las personas van formando hábitos en sus vidas generando rutinas para la mayoría de sus actividades a través de la continua repetición de las mismas. Si bien, el comportamiento humano es un aspecto muy difícil de analizar, intuitivamente sabemos que las personas definen regiones de permanencia y secuencias de ubicaciónes que visitan basados en los roles que como persona desempeñan dentro de una sociedad. Aunque el ser humano tiene el potencial de exhibir patrones de comportamiento aleatorio, existen ciertas rutinas fácilmente identificables en el contexto de su quehacer diario; por ejemplo, consideremos los pasos típicos de una secuencia de ubicaciones para una persona en un día laboral:

1. Levantarse a una hora definida.
2. Salir de su casa y tomar transporte (ya sea privado ó público).

3. Permanecer en su lugar de trabajo un período de tiempo definido.
4. Salir de la oficina y tomar transporte.
5. Regresar a casa.

Esta serie de ubicaciones en ese orden específico sugieren la presencia de patrones de comportamiento en los seres humanos en las ubicaciones que visita con base en sus rutinas de comportamiento en su vida diaria.

1.1. Modelos de tópicos

Los modelos de tópicos se desarrollaron inicialmente para el análisis de grandes colecciones de documentos de texto [6,12]; de manera simultánea e independiente, modelos similares se aplicaron en el área de biología, específicamente, en genética, donde los modelos de tópicos se han aplicado para la identificación y clasificación de poblaciones de genotipos [16]. Actualmente se han aplicado satisfactoriamente en otros contextos tales como ubicación [8] y proximidad física [2,7]. La capacidad de los modelos para extraer la información semántica y contextual de un documento los convierten en una herramienta fundamental para el análisis de texto, tomemos por ejemplo, la Enciclopedia Libre *Wikipedia* [4] [1], la cuál es un compendio de más de 4 millones de artículos en Inglés (al 10 de Abril de 2015 según la página principal) clasificados en varios temas principales (Artes, Historia, Sociedad, Matemáticas, etc.). De manera natural, existen artículos que pertenecen a más de una clasificación por su contenido temático, por lo cuál podemos preguntarnos, ¿existen clasificaciones que no se hayan considerado entre las actuales?, ¿hay clasificaciones a las que un documento pertenezca que no se hayan detectado?. Las anteriores y otras preguntas surgen al imaginar la extensión del conocimiento plasmado en un número tan grande de artículos y que continuamente va creciendo.

Fig. 1. Buscador por Tópicos de Wikipedia

[4] www.wikipedia.org

Wikipedia Topics [5] [4] es un ejemplo de la aplicación de un modelo de tópicos a un conjunto de artículos de la *Wikipedia*. En la Figura 1 podemos ver la organización de los tópicos estimados a partir de la aplicación del modelo en el conjunto de documentos y en la Figura 2 podemos ver el conjunto de documentos relacionados a un tópico. Podemos ver que los Modelos de Tópicos nos permiten inferir de manera automática un contenido semántico implícito en un conjunto de documentos de acuerdo al contenido temático de cada uno de ellos.

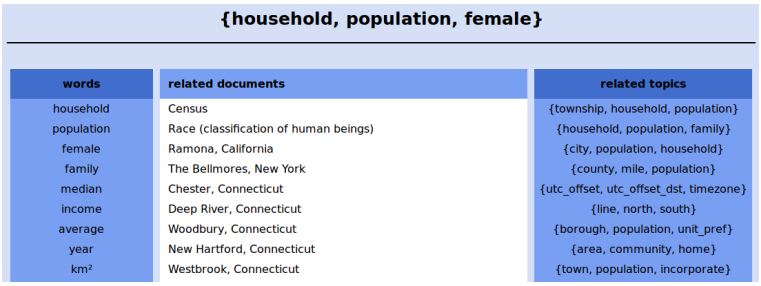

Fig. 2. Artículos relacionados a un tópico

Por su capacidad de análisis sobre conjuntos de datos discretos, los modelos de tópicos se vuelven una herramienta muy útil para el reconocimiento de patrones. Por lo tanto, el interés principal de éste trabajo es la identificación de patrones estructurales en los conjuntos de datos de ubicaciones geográficas obtenidas con dispositivos móviles aplicando modelos de tópicos.

1.2. Estructura del trabajo realizado

En este artículo se analiza un modelo de tópicos aplicado a un conjunto de datos que representan secuencias de ubicaciones geográficas con el objetivo de identificar una estructura presente en los datos que representen patrones de comportamiento en las personas. Se analizan aspectos teóricos del modelo así como una verificación experimental del funcionamiento del mismo utilizando datos sintéticos generados específicamente para probar aspectos del funcionamiento del modelo. Finalmente se aplica el modelo a un conjunto de datos de movilidad de usuarios observando los resultados obtenidos así como las conclusiones y trabajo a futuro.

2. Distant N-Gram Topic Model (DNTM)

Dos de los aspectos mas relevantes para el análisis de secuencias de datos discretos son la capacidad del modelo para expresar la pertenencia de un elemento

[5] http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-presence.html

a más de una categoría de clasificación y la capacidad para incluir información de dependencia entre los elementos en las secuencias de datos. Estas ventajas las presenta el modelo DNTM, el cual se describe brevemente en la sección 2.3 y a detalle en la sección 3.

2.1. Modelos de pertenencia mixta

$$p(v) = \sum_{h=1}^{H} p(v \mid h)p(h) \tag{1}$$

Un modelo mixto es uno en el cuál un conjunto de modelos es combinado para producir un modelo más expresivo [3]. En un modelo mixto, la variable v es visible y la variable h la indexa conformando los grupos a los cuáles puede pertenecer la variable v. Específicamente, podemos observar que una palabra puede pertenecer a más de un tópico lo cuál se indica con la medida de probabilidad asignada a ella en cada uno de los tópicos. Sin embargo, en un modelo mixto tradicional, un elemento sólo pertenece a un único grupo, por ejemplo en el algoritmo K-Means [14], un algoritmo que geométricamente descubre K grupos de elementos dada una medida de distancia. En K-Means cada elemento del conjunto de datos se asocia a un grupo lo cuál se conoce como *hard-clustering*. A diferencia, el *soft-clustering* considera que un elemento puede pertenecer a más de un grupo, de acuerdo a cierta medida de similitud. Los modelos de tópicos como LDA [6] y PLSI [12] consideran que un elemento de datos puede pertenecer a más de una clasificación (tópicos) de acuerdo a cierta medida de probabilidad.

2.2. Modelado de las dependencias

De manera natural, las personas inician sus rutinas diarias en un lugar en específico a partir del cuál se van moviendo a lo largo de sus rutinas definidas. Por lo tanto, es deseable que el modelo probabilístico a aplicar tenga ésta consideración en cuenta agregando dichas dependencias para poder capturar de manera correcta el significado semántico de su comportamiento. Sin embargo, se requiere evitar la explosión paramétrica que ocurriría si se modelan de manera explícita las dependencias entre cada uno de los elementos de la secuencia. Por tanto, en [10] se sugiere el modelo expresado en la Figura 3, donde se puede apreciar que, en primer lugar, se define el modelo con base en secuencias de ubicaciones $q = (w_1, w_2, ..., w_N)$ en vez de ubicaciones individuales, y en segundo lugar, que se define una dependencia a la primera ubicación visitada en la secuencia.

2.3. Descripción del modelo

El modelo DNTM [10] es un modelo derivado de LDA [5] el cuál es un modelo generativo bajo el cuál cada documento en una colección de texto (corpus) es representado con una distribución multinomial de tópicos los cuáles capturan el

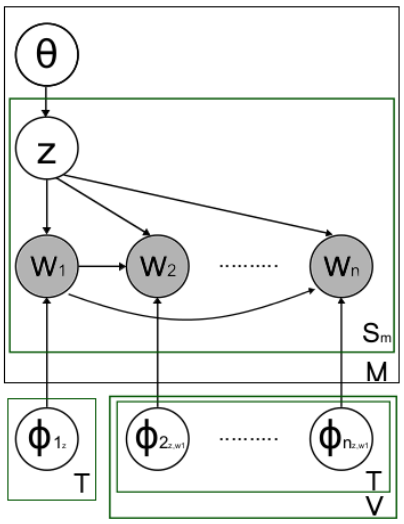

Fig. 3. Modelo Gráfico del modelo DNTM [10]

significado semántico de cada documento. A su vez, cada tópico es representado con una distribución multinomial sobre las palabras de un diccionario las cuáles capturan el significado semántico del tópico. En la modelación de patrones estructurales de movilidad de las personas, se considera una ubicación geográfica como una analogía a una palabra dentro de un documento y un intervalo de tiempo en el cuál se han dado las ubicaciones como una analogía a un documento. Es decir, se tiene un conjunto de elementos (ubicaciones geográficas) que pertenecen a diferentes clasificaciones (rutinas) de acuerdo a la sucesión de ellas. De ésta manera, el modelo de tópicos intenta descubrir o capturar el significado semántico de las secuencias de ubicaciones que un usuario visita con base en sus rutinas de movilidad.

3. Desarrollo del modelo

Como se observa en la Figura 3, θ y ϕ son las realizaciones Dirichlet de las distribuciones de los tópicos sobre los documentos (distribuciones multinomiales) las cuáles influyen directamente en la generación de la variable z (los tópicos latentes) y las variables w_n las cuáles dependen del tópico generado z, las variables ϕ_n y, para $j > 1$ del primer elemento w_1. Los parámetros del modelo se muestran en el cuadro 1.

3.1. Estimación de parámetros

La estimación exacta de parámetros para el modelo DNTM como en el modelo LDA es intratable, por lo que se han desarrollado métodos de apro-

N	Longitud de la secuencia.
q	Una secuencia de N etiquetas consecutivas $(w_1, ..., w_N)$.
m	Una instancia de un documento.
S_m	El número total de secuencias q en el documento m.
M	El número total de documentos en el corpus.
T	El número de tópicos latentes.
z	Una instancia de un tópico.
V	El tamaño del vocabulario.
Θ	La distribución de los tópicos dados los documentos.
Φ	La distribución de las secuencias dados los tópicos, donde $\Phi = \{\Phi_{1_z}, \Phi_{2_{z,w_1}}, ..., \Phi_{n_{z,w_1}}\}$
Φ_{1_z}	La distribución de las etiquetas w_1 dados los tópicos.
$\Phi_{j_{z,w_1}}$	La distribución de las etiquetas w_j dada la primera etiqueta w_1 y los tópicos.

Tabla 1. Descripción de símbolos de los componentes del modelo DNTM

ximación basados en inferencia variacional [6] y simulaciones de Monte Carlo (MCMC) [11]. En [9] y [10] se desarrolla un método de inferencia por Monte Carlo utilizando un esquema de Gibbs Sampling colapsado para estimar los parámetros del modelo θ y ϕ a partir de la simulación de la distribución condicional $p(z \mid q)$, por lo que se requiere de la expresión $p(z_i \mid z_{-i}, q)$ donde z_{-i} denota la probabilidad condicional para la simulación de la distribución de un sólo tópico z_i condicionado a las observaciones q y los tópicos restantes z_{-i}.

3.2. Desarrollo del modelo (colapsado de variables)

A partir de la Figura 3 y bajo la asunción de *i.i.d.* presente en los datos, determinamos la expresión de la verosimilitud para desarrollar el colapsado de variables (marginalización de las variables θ y ϕ para determinar la probabilidad $p(z, w_{1:n})$. Adoptamos la siguiente notación para escribir la expresión de la verosimilitud:

1. Z el conjunto de tópicos (variables latentes) z_i.
2. W el conjunto de observaciones (secuencias) $w_{1:n}$.
3. Θ el conjunto de realizaciones Dirichlet de las distribuciones de tópicos sobre documentos θ.
4. Φ el conjunto de realizaciones Dirichlet de las distribuciones de palabras sobre tópicos ϕ.

$$P(Z, W, \Theta, \Phi|\alpha, \beta_{1:n}) = \prod_{m=1}^{M}\left[p(\theta_m|\alpha)\prod_{s=1}^{S}\left[p(z_{ms}|\theta_m)p(w_{1ms}|z_{ms}, \varphi_1)\prod_{j=1}^{n}p(w_{jms}|w_{1ms}, z_{ms}, \varphi_{jw_1})\right]\right] \prod_{t=1}^{T}\left[p(\varphi_{1t}|\beta_1)\prod_{j=2}^{n}\prod_{v=1}^{V}p(\varphi_{jtv}|\beta_j)\right] \quad (2)$$

A partir de la expresión de la verosimilitud (2), se procede a integrar sobre los parámetros θ y ϕ para poder obtener la distribución requerida para el procedimiento de inferencia $p(z|q)$. Definiendo variables para la composición de los

productos en sumatorias sobre los exponentes de las distribuciones de probabilidad ν_{mt}, τ_{tv} y $\xi_{jw_1w_2}$, procedemos a la normalización de las distribuciones Dirichlet con el fin de realizar su integración la cuál es igual a 1. (La integración de una distribución de probabilidad sobre todo su dominio es igual a 1), lo que resulta en la distribución de probabilidad conjunta de Z y W.

$$\begin{aligned} &P(Z,W|\alpha,\beta_{1:n}) = \\ &\prod_{m=1}^{M} \frac{\Gamma(\alpha T)}{\Gamma(\alpha)^T} \frac{\prod_{t=1}^{T} \Gamma(\nu_{mt}+\alpha)}{\Gamma(\sum_{t=1}^{T} \nu_{mt}+\alpha)} \\ &\prod_{t=1}^{T} \frac{\Gamma(\beta_1 V)}{\Gamma(\beta_1)^V} \frac{\prod_{v=1}^{V} \Gamma(\tau_{tv}+\beta_1)}{\Gamma(\sum_{v=1}^{V} \tau_{tv}+\beta_1)} \\ &\prod_{t=1}^{T} \prod_{j=2}^{n} \frac{\Gamma(\beta_j)^V}{\Gamma(\beta_j)^{V^2}} \frac{\prod_{w_1=1}^{V} \prod_{w_2=1}^{V} \Gamma(\xi_{jtw_1w_2}+\beta_j)}{\Gamma(\sum_{w_1=1}^{V} \sum_{w_2=1}^{V} \xi_{jtw_1w_2}+\beta_j)} \end{aligned} \tag{3}$$

Recordemos que la inferencia sobre ésta distribución es intratable [11] por lo que recurrimos a métodos aproximados y para utilizar Gibbs Sampling, requerimos de la probabilidad condicional $p(z_i \mid z_{-i}, w)$, es decir, la expresión que corresponda a la probabilidad condicional de una tópico z dados los tópicos restantes z_{-i} y las observaciones w. Por tanto, observando que para una observación z_i en el modelo su valor es igual a 1, aplicamos las propiedades de la función Gamma para simplificar la expresión de la verosimilitud obteniendo las expresiones para las probabilidades condicionales requeridas para el muestreador Gibbs (4).

$$\begin{aligned} &p(z_{ms}=k|z^{\neg ms}, w_{1\,n}, \alpha, \beta_{1\,n}) \propto \\ &(\nu_{mt}^{\neg ms}+\alpha) \frac{\tau_{tv}^{\neg ms}+\beta_1}{\sum_{v=1}^{V}(\tau_{tv}^{\neg sm}+\beta_1} \prod_{j=2}^{n} \frac{\xi_{jtw_1w_2}^{\neg sm}+\beta_j}{\sum_{w_1=1}^{V}\sum_{w_2=1}^{V} \xi_{jtw_1w_2}^{\neg sm}+\beta_j} \end{aligned} \tag{4}$$

4. Verificación del modelo

4.1. Estructura de los datos sintéticos

Conjunto de Datos Sintéticos No. 1. El primer conjunto de datos se realiza con un vocabulario de 10 etiquetas distribuidas en 5 tópicos cada uno representado como una secuencia de 6 ubicaciones. Se genera un documento con 2,000 secuencias aleatorias asumiendo la misma probabilidad para todos los tópicos dentro de un mismo documento.

En la Figura (4) se muestran las distribuciones de probabilidad correspondientes a cada uno de los 5 tópicos, donde cada tópico contiene una secuencia de 6 posibles ubicaciones (eje x); la primera ubicación sobre el eje x corresponde a la etiqueta que se genera en primera posición de las secuencias y las ubicacions posteriores sobre el mismo eje corresponden a las probabilidades de las etiquetas siguientes. El eje y corresponde a las posibles etiquetas del vocabulario ($V = 10$).

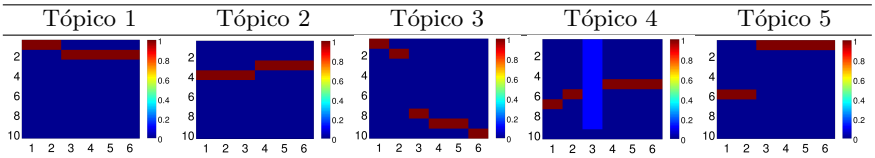

Fig. 4. Distribuciones de probabilidad de los tópicos del primer conjunto de datos sintéticos.

Obsérvsese que en el tópico 4 generamos en la tercera posición probabilidades similares para los elementos del vocabulario excepto el último.

Conjunto de Datos Sintéticos No. 2. La característica más relevante del conjunto de datos sintéticos No. 2 es la asignación de probabilidades para la generación de más de una etiqueta en la primera posición en las secuencias del documento, lo cuál se muestra en la figura 5.

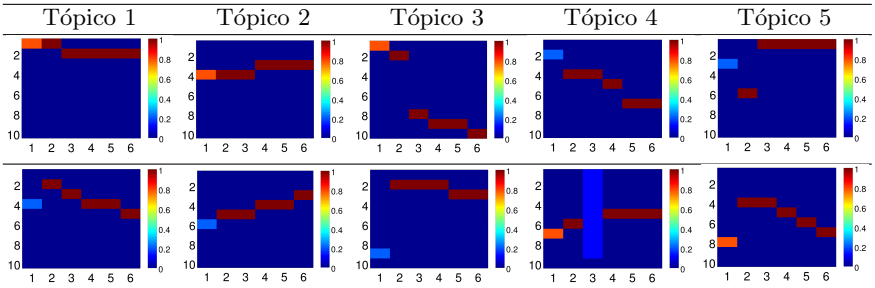

Fig. 5. Distribuciones de probabilidad de los tópicos del segundo conjunto de datos sintéticos.

Conjunto de Datos Sintéticos No. 3. Como podemos observar en la Figura 6, para el tercer conjunto de datos se generan 5 tópicos con distribuciones similares, donde sólo en el último elemento de las secuencias se observa una diferencia entre las etiquetas generadas. De ésta manera queremos determinar si el modelo es capaz de inferir diferencias entre estructuras de movilidades similares.

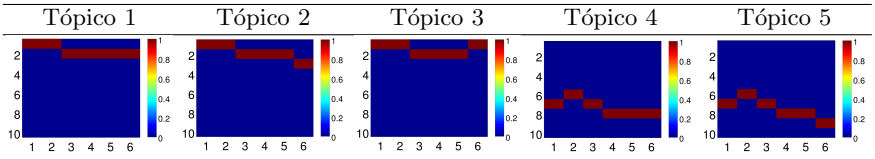

Fig. 6. Distribuciones de probabilidad de los tópicos del tercer conjunto de datos sintéticos.

Conjunto de Datos Sintéticos No. 4. En el último conjunto de datos deseamos verificar que el modelo sea capaz de identificar secuencias de longitudes diferentes aun a pesar de que él mismo no integre dentro de su estructura un mecanismo para determinar la longitud de cada secuencia. Se diseña entonces, un conjunto de 6 tópicos consistentes en secuencias de diferente longitud $N = 6$ y $N = 9$. En la Figura 7 se muestran las distribuciones de probabilidad para los tópicos, donde los tópicos 4 y 5 son los tópicos de longitud 9 y los restantes de longitud 6.

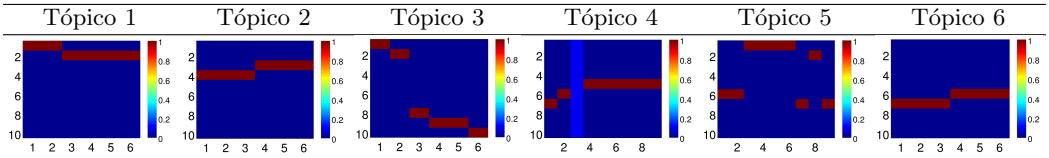

Fig. 7. Distribuciones de probabilidad de los tópicos del cuarto conjunto de datos sintéticos.

4.2. Resultados del modelo con datos sintéticos

Resultados del conjunto de Datos Sintéticos No. 1. Aplicando el modelo con los valores de $T = 5$ y $N = 6$, obtenemos en la Figura 8 que los resultados son bastantes precisos para el conjunto de datos, observando que el tópico 3 mezcló secuencias de tópicos diferentes y los tópicos 3 y 4 del modelo original se combinaron en el tópico 4.

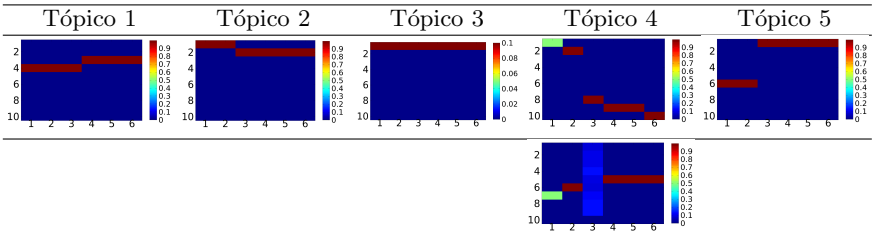

Fig. 8. Distribuciones de probabilidad obtenidas con el modelo con $N = 6$ y $T = 5$.

Resultados del conjunto de Datos Sintéticos No. 2. Podemos observar en la Figura 9 que los tópicos obtenidos son más precisos, sin embargo, uno de ellos no fue recuperado, lo cuál puede indicar que es mejor considerar un número mayor de tópicos para permitir que el modelo pueda explorar el espacio de búsqueda de distribuciones de probabilidad que puedan corresponder a los parámetros que el modelo va a inferir sobre los datos.

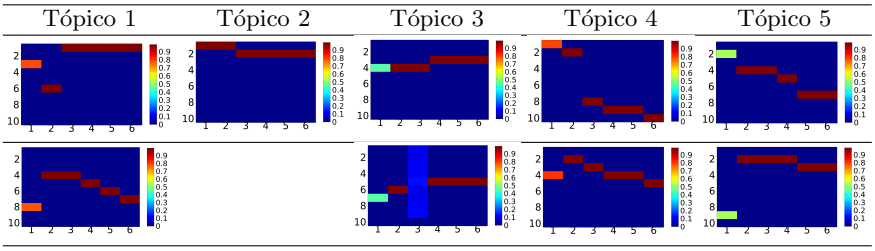

Fig. 9. Distribuciones de probabilidad obtenidas con el modelo con $N = 6$ y $T = 5$.

Resultados del conjunto de Datos Sintéticos No. 3. Podemos observar que el modelo no distingue de manera tan precisa entre los tópicos de estructura similar, como podemos apreciar en la Figura 10, los tópicos 1, 2, 4 y 5 tienen la misma estructura de secuencias difiriendo en la última pero combinando los tres tópicos de las distribuciones originales en los mismos tópicos. En el tópico 3 combinó los dos tópicos similares. Esto indica que el modelo es muy sensible a la presencia de estructuras de movilidad muy similares entre sí, no pudiendo distinguir pequeñas diferencias entre ellas.

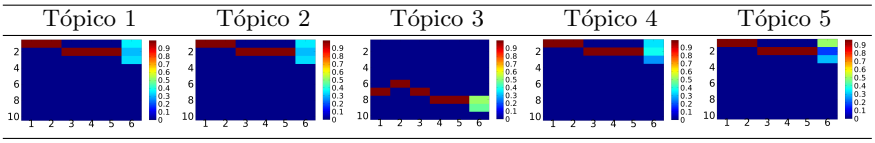

Fig. 10. Distribuciones de probabilidad obtenidas con el modelo con $N = 6$ y $T = 5$.

Resultados del conjunto de Datos Sintéticos No. 4. En las pruebas realizadas para éste conjunto de datos, se utilizaron valores de longitudes de secuencias $N = 6$, $N = 9$ y $N = 12$, observando que para $N = 6$ el modelo no alcanza a reconocer las estructuras de longitudes mayores diviendolo entre los tópicos detectados y para $N = 9$ el modelo reconoce de manera correcta longitudes menores pero las mayores las divide igualmente entre los tópicos. Mostramos entonces los resultados para $N = 12$ en la Figura 11 donde se muestran las distribuciones de máxima probabilidad para las secuencias detectadas.

5. Reconocimiento de patrones de movilidad en datos de ubicación geográfica

5.1. Datos Nokia: Lausanne data collection campaign [13]

El conjunto de datos móviles analizados consiste en el seguimiento de datos reales en la vida de 25 usuarios llevando un teléfono móvil Nokia modelo N95 en un periodo de nueve meses (del 01-10-2009 hasta el 01-07-2010).

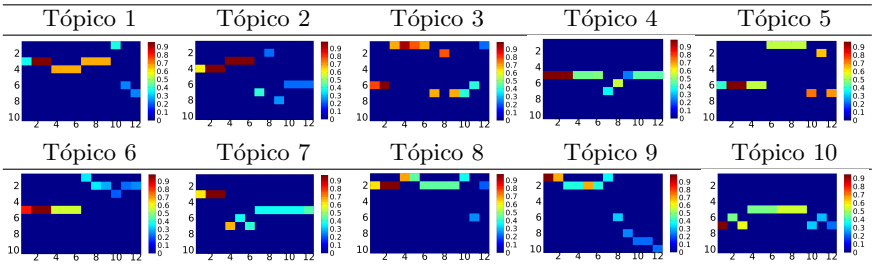

Fig. 11. Distribuciones de máxima probabilidad obtenidas con el modelo con $N = 12$ y $T = 10$.

En [10], obtienen los datos crudos de los teléfonos móviles y utilizan un algoritmo de extracción de datos [15] para el agrupamiento de las ubicaciones geográficas detectadas en un intervalo de tiempo basado en técnicas de agrupamientos de datos. Por tanto, se dispone de un conjunto de datos discreto multidimensional de tamaño 25x335x140 (es decir, 25 usuarios, 335 días para cada uno de ellos y 140 mediciones de intervalos de 10 minutos para cada día).

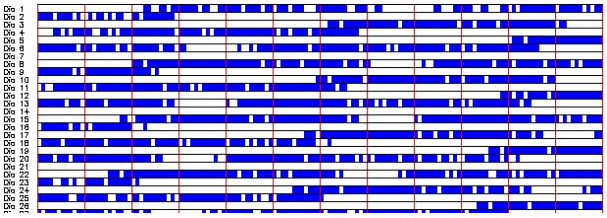

Fig. 12. Detección de ubicaciones geográficas correspondientes a un usuario

En la Figura 12 podemos observar una gráfica correspondiente a las detecciones de ubicaciones geográficas del usuario No. 1 del Conjunto de datos Nokia, en el eje x tenemos los 140 intervalos de tiempo correspondientes a 10 minutos cada uno de ellos, en el eje y tenemos los días durante los cuáles se hizo la detección y el color de cada celda corresponde a distintas ubicaciones pero donde el blanco se utiliza para indicar que no hubo detección correcta de ella.

5.2. Tópicos más probables

En la Figura 13 observamos un gráfico señalando los tópicos más probables para cada usuario, sobre el eje x tenemos cada uno de los 10 tópicos $T = 10$ y sobre el eje y tenemos a los usuarios, el color de cada celda indica la probabilidad de cada tópico para cada usuario. Observamos que en general, algunos tópicos definen la mayoría de la probabilidad de ocurrencia para cada usuario, teniendo algunos casos donde dicha probabilidad se distribuye de manera más uniforme.

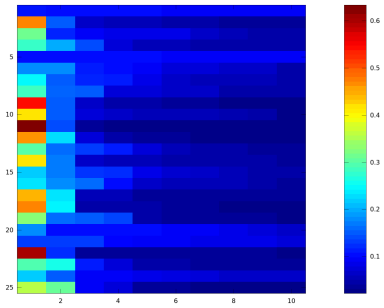

Fig. 13. Tópicos más probables obtenidos por el modelo DNTM sobre los 25 usuarios con 10 tópicos $T = 10$ y longitud de secuencia $N = 6$

5.3. Días más probables

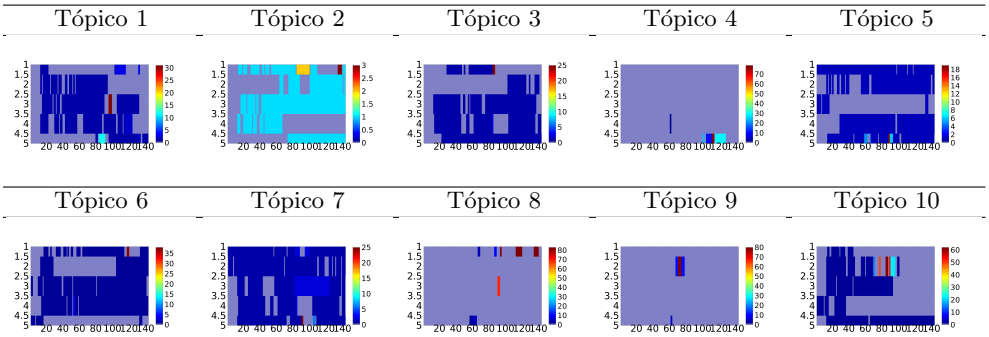

Fig. 14. Distribución de los 5 días más probables dados los tópicos.

En la Figura 14 podemos visualizar los 5 días mas probables para el usuario No. 1 con $T = 10$ y longitud de secuencia $N = 6$. Sobre el eje x se encuentran los intervalos de tiempo de 10 minutos definidos en el día (140) y sobre el eje y se encuentran los días más probables para cada tópico. El color de cada celda indica las diferentes etiquetas de las regiones de permanencia definidas para el usuario y las celdas en color gris indican que no hubo detección de ubicación en ese intervalo de tiempo.

5.4. Secuencias más probables

En la Figura 15 podemos visualizar algunos tópicos descubiertos por el modelo para el usuario No. 2, observamos la predominancia de las primeras etiquetas en los tópicos lo cuál proviene del hecho de que las etiquetas se ordenaron por

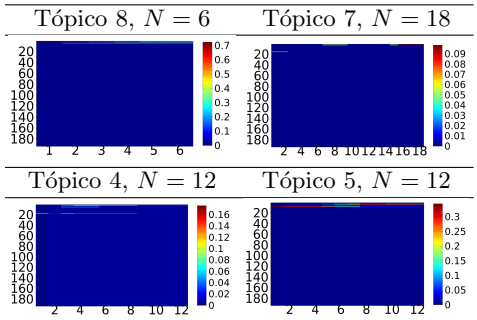

Fig. 15. Usuario 2, Distribución de Secuencias

frecuencia de aparición, aun así, observamos que la probabilidad se distribuye entre ellas. En los tópicos número 4 y 5 ($N = 12$) observamos también que las secuencias descubiertas por el modelo integran información de etiquetas diversas tal como en los datos sintéticos.

6. Conclusiones y trabajo a futuro

Los modelos de tópicos se han venido desarrollando desde sus primeras aplicaciones para modelado de texto y aplicaciones en genética, hasta dar el salto hacia su aplicación en el contexto de "reality mining", lo cuál es una extensión natural debido a sus características inherentes al análisis de datos discretos y su potencial aplicación para discernir estructuras subyacentes en un conjunto de datos. Con base en los resultados obtenidos con los datos sintéticos y el conjunto de datos Nokia de movilidad se puede concluir que un modelo de tópicos es una herramienta adecuada para el reconocimiento de patrones en el conjunto discretizado de ubicaciones geográficas y que permiten la detección de estructuras de movilidad en las personas.

Durante el desarrollo del modelo se observaron varias áreas de oportunidad para continuar con el trabajo y mejorar los resultados, a continuación se presentan algunas sugerencias del trabajo que se pueden desarrollar:

- **Identificación de las longitudes de secuencia**.- Aplicar métodos no paramétricos para identificar de manera automática las longitudes de secuencia subyacentes en los datos.
- **Modelos alternativos de análisis**.- Aplicar extensiones al modelo probabilístico para identificar una mayor descomposición de la información, como el modelo de autor para identificar patrones comunes de movilidad en los usuarios (movilidad en grupos).
- **Mayor exploración del espacio de resultados**.- Implementar "Parallel Tempering" en el modelo para permitir la mayor exploración del espacio de resultados con el muestreador MCMC.

Referencias

1. Wikipedia la enciclopedia libre. http://www.wikipedia.org/ (2015), [Fecha de consulta: 10-Abril-2015]
2. Bao, T., Cao, H., Chen, E., Tian, J., Xiong, H.: An unsupervised approach to modeling personalized contexts of mobile users. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. pp. 38–47 (Dec 2010)
3. Barber, D.: Bayesian Reasoning and Machine Learning. Cambridge University Press (2012)
4. Barlow, A.: Wikipedia topics. http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-presence.html/ (2015), [Fecha de consulta: 10-Abril-2015]
5. Blei, D.M.: Probabilistic topic models. Commun. ACM 55(4), 77–84 (Apr 2012), http://doi.acm.org/10.1145/2133806.2133826
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003), http://dl.acm.org/citation.cfm?id=944919.944937
7. Do, T.M.T., Gatica-Perez, D.: Groupus: Smartphone proximity data and human interaction type mining. In: Proceedings of the 2011 15th Annual International Symposium on Wearable Computers. pp. 21–28. ISWC '11, IEEE Computer Society, Washington, DC, USA (2011), http://dx.doi.org/10.1109/ISWC.2011.28
8. Farrahi, K., Gatica-Perez, D.: Probabilistic mining of socio-geographic routines from mobile phone data. Selected Topics in Signal Processing, IEEE Journal of 4(4), 746–755 (Aug 2010)
9. Farrahi, K.: A Probabilistic Approach to Socio-Geographic Reality Mining. Ph.D. thesis, École Polytechnique Fédérale de Lausanne (2011)
10. Farrahi, K., Gatica-Perez, D.: A probabilistic approach to mining mobile phone data sequences. Personal and Ubiquitous Computing 18(1), 223–238 (2014), http://dx.doi.org/10.1007/s00779-013-0640-8
11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101(suppl 1), 5228–5235 (2004), http://www.pnas.org/content/101/suppl_1/5228.abstract
12. Hofmann, T.: Probabilistic latent semantic analysis. In: In Proc. of Uncertainty in Artificial Intelligence, UAI'99. pp. 289–296 (1999)
13. Kiukkonen, N., J., B., Dousse, O., Gatica-Perez, D., J., L.: Towards rich mobile phone datasets: Lausanne data collection campaign. In: Proc. ACM Int. Conf. on Pervasive Services (ICPS), Berlin. (7 2010)
14. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. University of California Press (1967)
15. Montoliu, R., Gatica-Perez, D.: Discovering human places of interest from multimodal mobile phone data. In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia. pp. 12:1–12:10. MUM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1899475.1899487
16. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W.: Genetic structure of human populations. Science 298(5602), 2381–2385 (2002), http://www.sciencemag.org/content/298/5602/2381.abstract