REVISTA CUBANA
DE TRANSFORMACIÓN DIGITAL

RNPS 2487
Vol. 1 • Nro. 1 • enero-abril 2020 • págs. 132-148

ARTÍCULO ORIGINAL

Análisis de ventanas de tiempo en Sistemas de Detección de Tendencias

Time Windows Analysis in Trend Detection Systems

Rocío Cruz-Linares
rociocl@matcom.uh.cu • https://orcid.org/0000-0002-0069-8950
Alejandro Piad-Morffis
apiad@matcom.uh.cu • https://orcid.org/0000-0001-9522-3239
Yudivián Almeida-Cruz
yudy@matcom.uh.cu • https://orcid.org/0000-0002-2345-1387

FACULTAD DE MATEMÁTICA Y COMPUTACIÓN, UNIVERSIDAD DE LA HABANA, CUBA

Recibido: 2019-12-12 • Aceptado: 2020-03-15

RESUMEN

Los servicios de *microblogging* son potenciales fuentes de datos actualizados. Su uso constante los ha convertido en un campo de acción idóneo para detectar tendencias. La mayor parte de los esfuerzos empleados en desarrollar Sistemas de Detección de Tendencias (SDT) apuntan a la definición de los modelos. Sin embargo, estas investigaciones han obviado el análisis de elementos tan importantes como las ventanas de tiempo, las cuales, mal configuradas, pueden ocasionar un erróneo funcionamiento del sistema. En esta investigación se analiza la influencia del uso de ventanas de tiempo estáticas en Sistemas de detección de tendencias. Se definen, además, dos metodologías para generar configuraciones de ventanas. Una de ellas, determina el tamaño de ventana óptimo para ventanas de tiempo estáticas. Mientras que la otra, modelada como un problema de optimización, construye configuraciones de ventanas capaces de adaptarse al flujo de los datos. Una vez experimentadas, ambas son comparadas mediante un proceso de evaluación de SDT propuesto en el documento. En la comparación, se tienen en cuenta elementos como la estructura de las ventanas, la convergencia de la optimización y las evaluaciones alcanzadas. Como resultado, se refleja la superioridad de las ventanas no estáticas respecto a las estáticas y queda enfatizado el papel determinante que juegan las ventanas de tiempo en los Sistemas de detección de tendencias.

Palabras clave: metaheurísticas; sistemas de detección de tendencias; ventanas de tiempo; Twitter.

ABSTRACT

Microblogging services are potential sources of updated data. Their constant use has turned them into a suitable field of action to detect trends. Most of the efforts employed in developing Trend Detection Systems (TDS) point to the definition of the models. However, these investigations have overlooked the analysis of important elements such as time windows, which, misconfigured, can cause the system to malfunction. In this research, the influence of the use of static time windows in TDS is analyzed. In addition, two methodologies are defined to generate window configurations. One of them determines the optimal window size for static time windows. While the other, modeled as an optimization problem, builds window configurations capable of adapting to the flow of data. Once experimented, both methodologies are compared through a SDT evaluation process proposed in the document. Elements such as window structure, the convergence of optimization and the evaluations achieved are taken into account in the comparison. As a result, the superiority of non-static windows over static ones is shown and the decisive role played by the time windows in the TDS is emphasized.

Keywords: *metaheuristcs; trend detection systems; time windows; Twitter.*

INTRODUCCIÓN

Hoy en día, alrededor de 330 millones de personas son usuarios activos de *Twitter*[1] (Wolfe, 2019) y generan un promedio 6 000 mensajes por segundo (*Twitter Usage Statistics,* 2019). Esta red social se ha convertido en una plataforma esencial para el seguimiento, difusión y coordinación de eventos de diversa naturaleza y relevancia como pueden ser una campaña presidencial, una situación de desastre, un conflicto bélico o la repercusión de una noticia (Bollen, Mao & Pepe, 2011). Gracias a la variedad e inmediatez de los mensajes que se comparten segundo a segundo, *Twitter* es considerado una fuente de información valiosa para realizar múltiples tareas de Extracción de información y Procesamiento de lenguaje natural, entre ellas, la Detección de tendencias.

La Detección de tendencias (TD, por sus siglas en inglés), responde a preguntas como: ¿cuáles son los temas más populares que representan los intereses de la Sociedad en un período de

[1] https://twitter.com/

tiempo?, ¿existe algún cambio significativo en los temas más tratados?, ¿cuáles son los temas emergentes o que desaparecen?, o ¿qué temas tienen un comportamiento significativamente contrario a la tendencia general? (Montes, *et al.,* 2001). El estudio de las tendencias obtenidas a partir de Sistemas de Detección de Tendencias (SDT), es una forma de comprender los intereses de la Sociedad al establecer una correlación entre estos elementos y los temas emergentes. Además, su análisis ayuda a identificar la presencia de tendencias o noticias falsas, que pueden ser el resultado de usuarios corruptos intentando modificar la realidad (Figueira, *et al.,* 2018). Dado su importancia, se impone la necesidad de analizar y tratar de perfeccionar este tipo de sistemas.

La detección de tendencias en redes sociales es una problemática actual estudiada desde muchas aristas. Desde el punto de vista computacional, se puede entender como identificar en un instante de tiempo, si un tópico, personalidad, evento, noticia, o cualquier otra entidad, está suscitando de repente un interés desproporcionado. Las aplicaciones son obvias, desde la evaluación de campañas publicitarias, hasta la recomendación de productos y servicios. Por estos motivos, la investigación sobre técnicas computacionales que permitan acercarse a una solución a este problema es de gran actualidad, como lo demuestra la amplia bibliografía existente al respecto. Muchas de las propuestas actuales se basan en dividir un conjunto de interacciones (mensajes, *clicks, likes)* en ventanas de tiempo y computar ciertas métricas de agregación sobre estos subconjuntos.

Un elemento importante dentro de todo SDT es la configuración de los intervalos de tiempo, estos delimitan el conjunto de datos a analizar en cada momento. En la gran mayoría de los sistemas se utilizan intervalos fijos, o sea, del mismo tamaño. Este tipo de ventanas ocasionan varios problemas (Abdelhaq, *et al.,* 2013), por ejemplo: peticiones innecesarias a la red, errores en la detección de pequeñas tendencias y, además, obliga a que los algoritmos implementados definan estrategias para almacenar información de los intervalos anteriores.

Aun así, no existe una propuesta de intervalos de tiempo que sean definidos en dependencia del flujo de datos, ni un estudio que analice cuán determinante pueden ser los intervalos de tiempo en los SDT. Por lo que el objetivo de este trabajo es estudiar la influencia de los intervalos de tiempo en los SDT, y así poder crear configuraciones de ventanas de tiempo que reporten mejores resultados. Particularmente, la utilización de ventanas no fijas, que respondan a las características del flujo de datos.

Sistemas de Detección de Tendencias

En el marco del evento *Text REtrieval Conference* (TREC) en 1997, surge el programa de Detección y Seguimiento de Tópicos (TDT, por sus siglas en inglés) (Allan, 2002). El TDT buscaba el desarrollo de técnicas básicas para analizar informaciones extraídas de los medios tradicionales de comunicación, con el objetivo de mantener a los usuarios actualizados sobre las noticias y su desarrollo (Allan, *et al.,* 1998). Este programa se enfocaba en fuentes de publicación de bajo volumen y noticias bien redactadas.

En el caso particular de las tareas de TDT se encuentran los SDT, considerados una especialización de los Sistemas de detección de tópicos (Fiscus & Doddington, 2002). Los SDT

constituyen una de las líneas más activas dentro del programa TDT. Esto se debe, en parte, a la actual curiosidad de empresas, grupos políticos, personalidades famosas, cadenas televisivas, etc. por conocer qué quieren, piensan o comentan las personas.

La literatura recoge múltiples investigaciones cuyo objetivo es la construcción de SDT sobre flujos de datos de *microblogging.* Generalmente, hacen uso de *Twitter* debido a sus características y popularidad como red social (Lee & Sumiya, 2010); (Lau, *et al.,* 2012); (Osborne, *et al.,* 2012); (Guzmán & Poblete, 2013); (Mederos, *et al.,* 2013); (Tejeda & Cruz, 2016); (Hasan, *et al.,* 2018). Los resultados obtenidos para detectar tendencias son muy diversos en flujos de datos de *microblogging,* pero de manera general se pueden dividir en dos grupos: modelos basados en documentos (*document-pivot)* y modelos basados en características (*feature-pivot).*

Modelos basados en documentos

En los modelos basados en documentos, los mensajes son representados de la manera tradicional, como vectores de términos o bolsa de palabras. Como aspecto importante, en estos modelos el orden temporal de las palabras y los rasgos semánticos y sintácticos son descartados debido a que el modelo no puede capturar la similaridad entre tendencias relacionadas. Además, este tipo de modelos son generalmente aplicados a sistemas para detectar eventos (Atefeh & Khreich, 2013).

Modelos basados en características

Las técnicas basadas en características modelan las tendencias como un estallido de actividad. Analizan las distribuciones de las palabras y determinan las tendencias agrupando las palabras que presentan un crecimiento brusco en su frecuencia (conocidas como *keyburst* en inglés) (Guzmán & Poblete, 2013); (Mathioudakis & Koudas, 2010); (Xun, Feida & Li, 2011). Luego, una tendencia es convencionalmente representada por un número de palabras claves. Las implementaciones de estos modelos son generalmente aproximaciones estadísticas.

Guzmán propone un algoritmo para detectar *keybursts* en flujos de datos de *microblogs* escalable y eficiente. Como desventaja, necesita ser entrenado para la detección de idiomas y se ha de encontrar un intervalo de tiempo apropiado para que brinde resultados satisfactorios (Guzmán & Poblete 2013).

Lau (Lau, *et al.,* 2012) expone una estrategia para detectar tendencias y eventos en flujos de datos de Twitter basado en la modelación de tópicos. La propuesta de Lau se define inicialmente para trabajar de manera *offline,* pero al final del escrito, propone una variante online que hace uso de ventanas deslizantes. Los autores de esta investigación recomiendan descartar los tweets con menciones a usuarios

Mathioudakis describe la principal herramienta para detectar *trending topics* sobre flujos de datos de *Twitter: TwitterMonitor.* Debido a su implementación, existe una fuerte dependencia entre la herramienta y su conexión a Internet, así como un exceso de pedidos al API de *Twitter,* aun cuando no sea necesario, ya sea porque no se han acumulado suficientes *tweets* o porque la mayoría de los que se van a recolectar son ruido (Mathioudakis & Koudas, 2010).

Hendrickson (Hendrickson, *et al.*, 2015) realiza un breve análisis de tres enfoques analíticos para detectar tendencias en flujos de datos de *Twitter: Point-by-point Poisson Model, Cycle-corrected Poisson Model* y *A Data-driven Method.* De manera general, estos modelos definen un modelo origen mediante el cual se representa la hipótesis nula, también identificada como el caso de no tendencia. En esta investigación se hace referencia a las dificultades que pueden ocasionar la escala temporal de los cambios, la longitud de las ventanas y el tamaño de los datos a analizar. Esto se debe a que no existe una medida justa o exacta para delimitar estos elementos, la variación de estas magnitudes puede marcar la diferencia entre el correcto funcionamiento y la inutilidad de estas técnicas estadísticas

Por último, la tesis de licenciatura de Tejeda también responde a un modelo basado en características. El autor presenta una metodología para detectar tendencias en *Twitter* (Tejeda & Cruz 2016).

Como pudo ser observado, existen diversas estrategias para solucionar el problema de la Detección de tendencias. Donde además es notable cómo la configuración de las ventanas de tiempo puede influir en el resultado de la mayoría de estas investigaciones. Son muchos los aspectos que se deberían tener en cuenta, por ejemplo:

- Si las ventanas son muy grandes y las tendencias son muy rápidas, entonces puede que no se lleguen a detectar. Lo mismo sería en el caso contrario, donde las ventanas son muy pequeñas y las tendencias abarcan un espacio temporal mucho mayor.
- Con ventanas pequeñas se producen accesos reiterados a la red, que en reiteradas ocasiones no aportan ninguna información valiosa para el sistema.
- Si las ventanas de tiempo fraccionan la frecuencia de mención de un tópico, entonces los algoritmos de detección de tendencias tendrán que proponer estrategias para almacenar los datos anteriores y poderlos reutilizar.
- Si se establecen tamaños de ventanas fijos, sin importar las características de los tópicos a analizar, entonces puede que se generen accesos a la red sin necesidad.

Resulta evidente que la configuración de las ventanas de tiempo es un aspecto importante. Son muchos los SDT desarrollados y, aun así, no ha sido estudiado este fenómeno más allá que para la evaluación de unas pocas configuraciones y la selección de la mejor de ellas. Por ello, más que el estudio de los SDT, es interesante el análisis de la configuración de las ventanas de tiempo.

El contenido de este artículo se encuentra dividido en dos secciones. En la Sección Análisis de las Ventanas de Tiempo Estáticas, se realiza un estudio sobre la influencia de las ventanas de tiempo en los SDT y se propone una metodología para determinar el mejor tamaño de ventana para sistemas que trabajan con ventanas de tamaño fijo. En la Sección Distribución de Ventanas No Estáticas, se propone una metodología para generar ventanas de tiempo no estáticas. Para comprobar la efectividad de la solución, se presentan también los resultados de los experimentos realizados. Al final del documento se ofrecen las Conclusiones y Recomendaciones.

METODOLOGÍA

Bajo la premisa de estudiar los SDT, la presente investigación se propone analizar en un primer momento los SDT basados en características con ventanas de tiempo estáticas. Una vez descrito el SDT seleccionado, el conjunto de datos y la medida de evaluación a utilizar, se procede a investigar la influencia del tamaño de ventana, donde se comprueba que al modificar esta variable se obtienen diferentes resultados. Luego, se estudia la posible existencia de un tamaño de ventana óptimo.

En un segundo momento se profundiza en los SDT basados en características con ventanas de tiempo no estáticas. Con el objetivo de encontrar distribuciones de ventanas que alcancen una mejor evaluación que las distribuciones de ventanas estáticas, se modela el problema como un problema de optimización. Este modelo debe maximizar la función de evaluación de las distribuciones de ventana. Una propuesta basada en la metaheurística Escalador de colinas es descrita como solución computacional, y se realizan un grupo de experimentos para evaluarla. La convergencia de la optimización se examina mediante la diferencia relativa de las evaluaciones entre iteraciones consecutivas. Para conocer la estructura de las distribuciones de ventanas generadas se analiza la distribución de la variable aleatoria tamaño de ventana. Mientras que por último, se realiza una prueba *z-test* para el estudio de los resultados de la metaheurística respecto a las evaluaciones de las distribuciones de ventanas estáticas. Finalmente se arriban a conclusiones y se proponen futuras líneas de desarrollo.

ANÁLISIS DE LAS VENTANAS DE TIEMPO ESTÁTICAS

Los SDT sobre flujos de datos de *microblogs,* en su mayoría, trabajan en tiempo real. Para ello definen ventanas deslizantes (o intervalos de tiempo) que les permiten discretizar la llegada continua de mensajes. Las ventanas poseen dos parámetros: el tamaño (*l)* y el desplazamiento (*r).* Si $r<l$, las ventanas se consideran ventanas solapadas y si $r \geq l$ las ventanas se consideran ventanas no solapadas. El caso en que $r>l$ nunca es tratado puesto que se quedarían documentos sin analizar. A su vez, las ventanas pueden ser estáticas o dinámicas (Laguna, Olaya & Borrajo, 2011):

- Ventanas estáticas: se mantienen los valores de l y r fijos. Los intervalos de documentos analizados son de la forma: $\{\ldots,[t-l,t],[t-l+r,t+r],[t-l+2r,t+2r],\ldots\}$. Donde t puede ser cualquier instante de tiempo.
- Ventanas dinámicas: no poseen valores fijos para l o para r.

En la literatura no existe ningún proceso estándar mediante el cual se definan las configuraciones de las ventanas de tiempo. Cada sistema las define de acuerdo a sus intereses y funcionamiento (Lau, *et al.,* 2012); (Guzmán & Poblete, 2013); (Tejeda & Cruz, 2016). Sin embargo, es común para todos los SDT que sus resultados dependan de las configuraciones de las ventanas utilizadas. En esta sección se realizará un análisis de la influencia de las ventanas de tiempo estáticas en un SDT para estudiar cómo varían los resultados ante el cambio de configuración de las ventanas. Como son casi inexistentes los casos que tratan ventanas solapadas, en esta investigación solo se analizarán las ventanas de tiempo estáticas no solapadas, por lo que la única configuración variable será el tamaño de la ventana.

Para llevar a cabo un análisis de la influencia de las ventanas de tiempo en los SDT basados en características se hace necesario uno, o más SDT. Desafortunadamente, la mayoría de los SDT consultados no son de código abierto. Sus implementaciones son privadas o los artículos donde se describen no es posible replicar el sistema (Sankaranarayanan, *et al.,* 2009); (Mathioudakis & Koudas, 2010); (Xun, Feida & Li, 2011). Lo mismo ocurre con los conjuntos de datos utilizados para detectar tendencias y realizar las evaluaciones. Después de una extensa búsqueda, se pudo acceder a la implementación de un SDT basado en características que hace uso de series de tiempo (Hendrickson, *et al.,* 2015). El autor propone un SDT sobre el flujo de datos de redes sociales basado en características. Su algoritmo, detecta cambios en las series de tiempo de los tópicos simulando el proceso de la generación de tendencias como un modelo *Poisson.* Sin embargo, el repositorio del proyecto no contiene un conjunto de datos verdadero, solo brinda un ejemplo de cómo deberían ser los archivos de entrada. Por lo que se optó por utilizar un conjunto de datos anotados que fue utilizado para el entrenamiento de un algoritmo de clasificación de tipos de tendencias (Zubiaga, *et al.,* 2014).

Otro elemento necesario para comparar la efectividad de los SDT son las medidas de evaluación. En la literatura existen diversas estrategias para realizar las evaluaciones (Weiler, *et al.,* 2015) por ejemplo: comprobar los resultados manualmente, ya sea con las noticias de periódicos locales o con los *trending topics* que brinda *Twitter* (Tejeda & Cruz, 2016), comparar los resultados con los resultados de otros SDT (Xun & Li, 2011), respaldar los resultados con información extraída de *Wikipedia* (Osborne, *et al.,* 2012) y, en muy pocas ocasiones, análisis estadísticos (Lau, *et al.,* 2012); (Ozdikis, *et al.,* 2012).

Son múltiples las vías para evaluar las configuraciones de ventanas, la mayoría de ellas muy subjetivas, y esto se debe a la carencia de conjuntos de datos anotados que se ajusten a los sistemas desarrollados. Con la presencia de un *corpus* anotado se hace posible el uso de medidas como la precisión, el recobrado y las *F*-medidas. Por esto, se decidió utilizar como medida la función *S* definida en la Tesis de Licenciatura de R. C. Linares (Linares, Morffis & Cruz, 2018), $S: D \rightarrow \mathbb{R}$ donde *D* es el conjunto que contiene todas las posibles distribuciones de ventanas:

$$S(d) = \alpha * F_2(d) - \beta * RR_T(d)$$

Ecuación 1.
Medida de evaluación de las distribuciones de ventanas de tiempo.

En la ecuación (1), $F_2(d)$ es la medida F_2 de la distribución de ventana *d*, $RR_T(d)$ es la cantidad de multidetecciones[2], y α, β son factores de peso. Estos factores de peso deben ser fijados de manera tal que la función *S(d)* penalice aquellas distribuciones de ventanas que tengan más multidetecciones.

Con el objetivo de estudiar la influencia de la configuración de las ventanas de tiempo estáticas en los SDT, se realizaron varios experimentos.

[2] Detección simultánea no necesaria de una tendencia en una distribución de ventanas.

Influencia del tamaño de ventana

Bajo la hipótesis de que el tamaño de ventana es un factor determinante en los resultados de los SDT, se analizó cómo al cambiar el tamaño de ventana se obtienen diferentes resultados. Para ello se realizaron ejecuciones del algoritmo con ventanas de tamaño de un minuto, una hora y un día. Los resultados demostraron que, efectivamente, varían la cantidad de momentos en que se detectó tendencia para cada configuración. De ahí la importancia de ajustar correctamente el tamaño de ventana para obtener resultados satisfactorios en los SDT. En la figura 1 se pueden observar los resultados obtenidos para el día primero de marzo del 2011. Cada gráfico contiene tres representaciones. En la parte superior se encuentran las series de tiempo de frecuencia de cada tópico, en la parte intermedia, la aproximación de cada serie de tiempo modelada con los valores η del modelo *Poisson* y en la parte inferior se muestran, marcados con una línea vertical y un punto, los momentos en que fueron detectadas las tendencias según el algoritmo.

Tamaño de ventana óptimo

Una incógnita latente en el ámbito de los SDT es el desconocimiento sobre la existencia o no de un valor óptimo de tamaño de ventana. Este no es un tema que se aborde en la literatura, puesto que, como se analizó, cada sistema define las configuraciones de las ventanas de tiempo acuerdo a sus intereses y funcionamiento, sin ningún procedimiento predefinido.

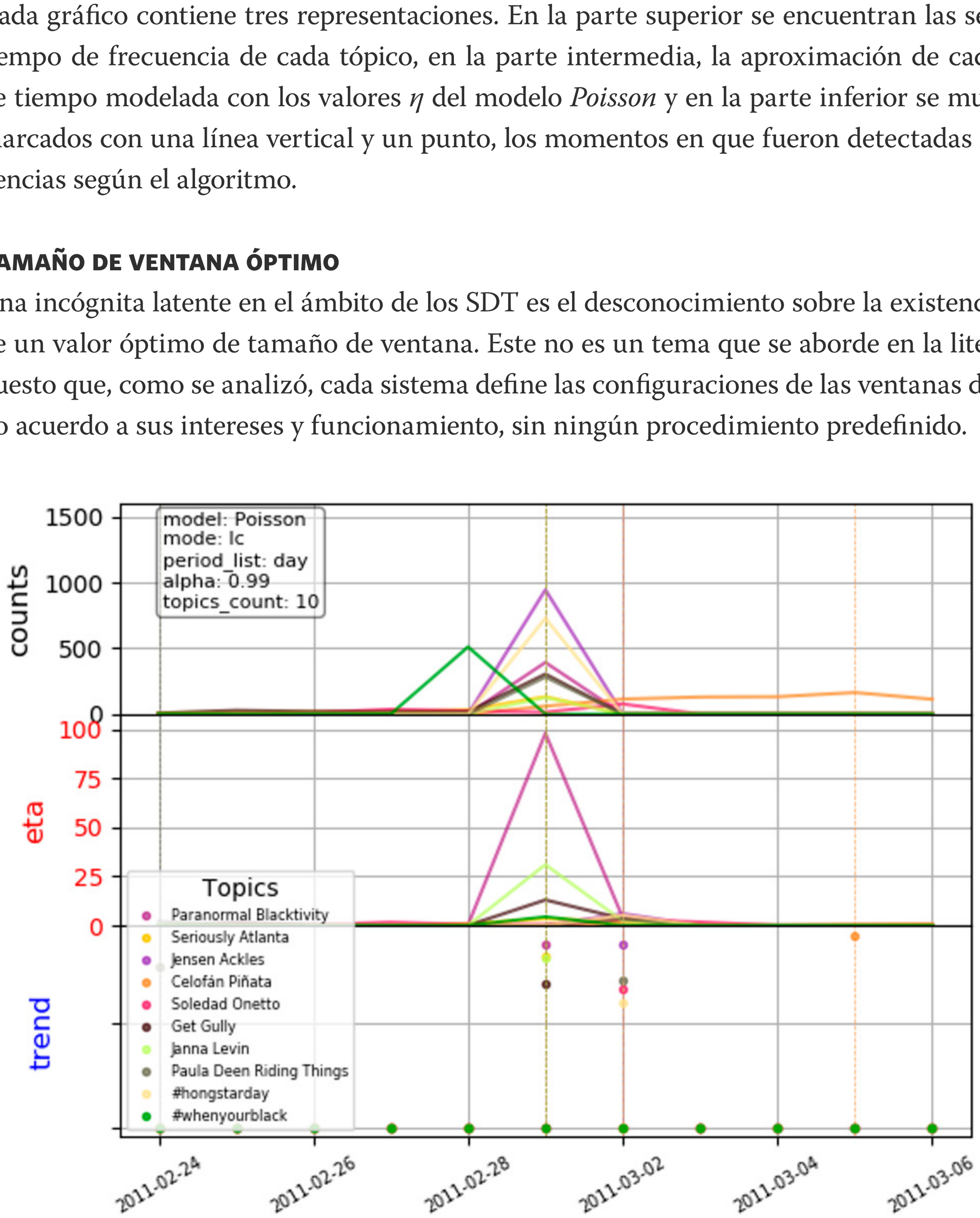

Figura 1a. *Trending Topics* del día 1ro de marzo 2011 (ventanas de un día).

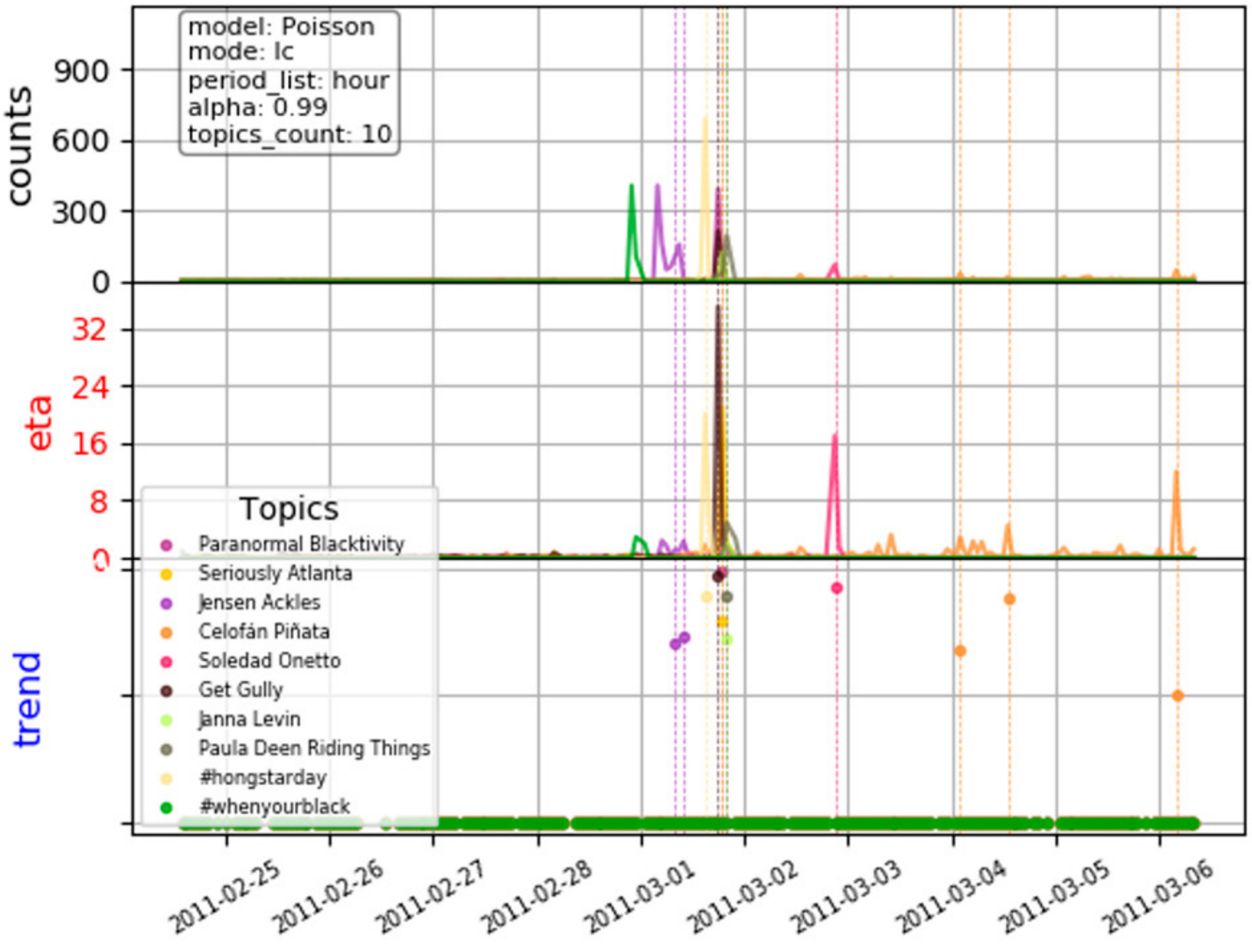

Figura 1b. *Trending Topics* del día 1ro de marzo 2011 (ventanas de una hora).

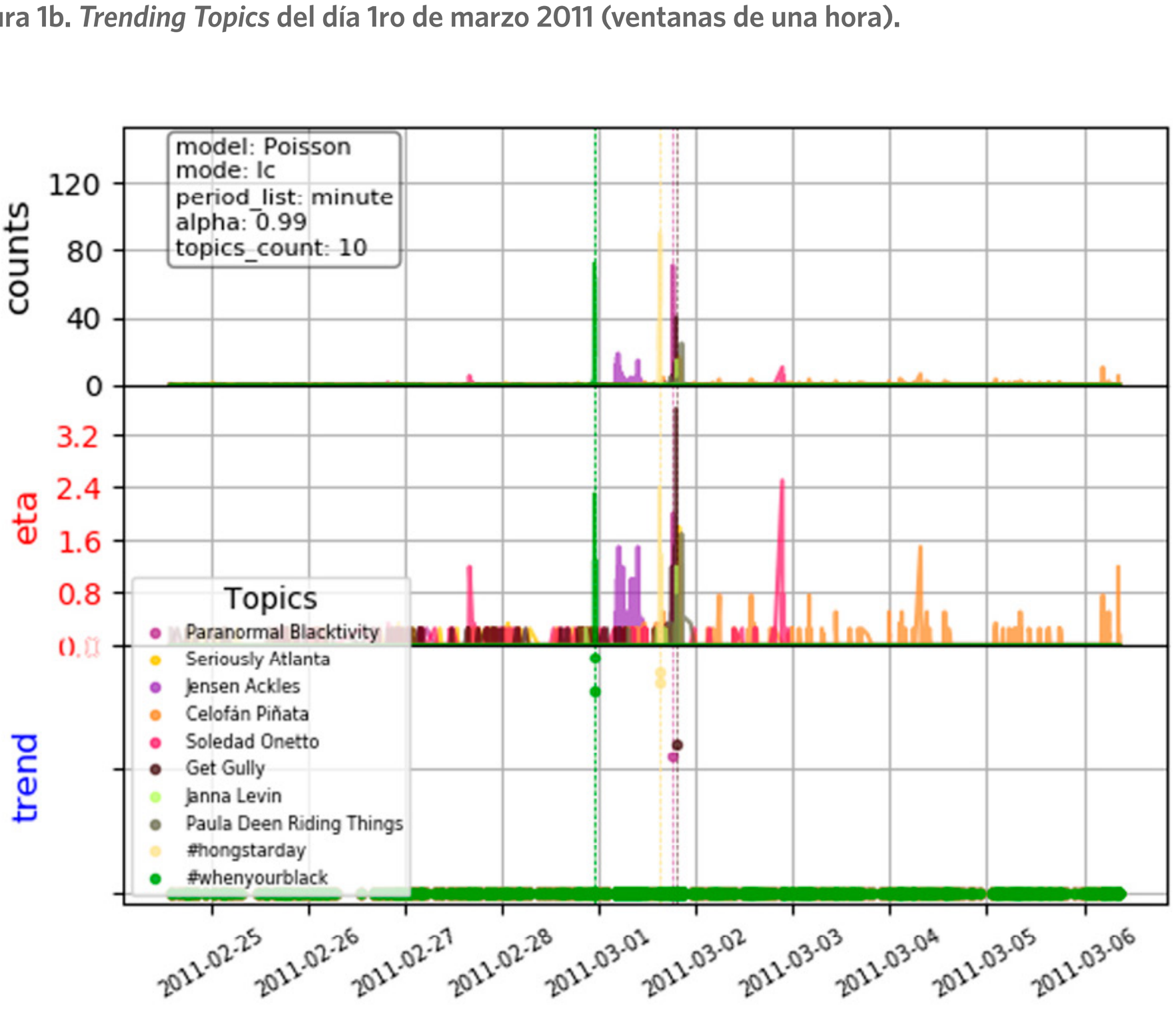

Figura 1c. *Trending Topics* del día 1ro de marzo 2011 (ventanas de un minuto).

Si se deseara conocer el tamaño de ventana óptimo para un SDT dado un conjunto de datos, se debería iterar por todos los posibles tamaños de ventanas, realizar la evaluación de las distribuciones de ventanas generadas y finalmente, tomar como óptimo el tamaño que mayor evaluación reporte. Sin embargo, como el tamaño de la ventana es una unidad de medida continua, no es practicable explorar todo el espacio de posibles valores.

Una posible manera de recorrer este dominio es determinar el valor *k* para el cual se obtienen los mejores resultados, siendo *k* la cantidad de ventanas en que se divide el espacio temporal en que están comprendidos los *tweets* del *corpus*. De esta manera, para mayores valores de *k* se generan ventanas de corta duración y para menores valores, se obtienen ventanas más grandes. Una vez que se tienen las evaluaciones de cada uno de los conjuntos de ventanas generados por los valores de *k*, se debe seleccionar el valor *k** (K ópt) que reporte mejores resultados.

Son muchos los factores que influyen en la selección de *k**, por ejemplo, el algoritmo utilizado como SDT, el *corpus* con que se evalúa y la metodología de evaluación. Por lo que una vez obtenido *k** para una muestra, cabe la duda de cuán estable es este valor. Se ejecutó el algoritmo con ventanas divididas desde cuatro hasta 1500 partes, de dos días hasta diez minutos respectivamente para determinar el tamaño de ventana óptimo para el conjunto de datos estudiado. El resultado obtenido fue que *k** toma valor 20, lo cual corresponde a ventanas de trece horas aproximadamente.

Para continuar con el análisis de la variable *k*, fue aplicado el mismo procedimiento a dos grupos de subconjuntos del *corpus:* el primero agrupados por días y el segundo agrupados por categorías de tendencia. Las tablas 1 y 2 muestran los valores *k** y los tamaños de ventanas correspondientes en segundos (T segs) para cada conjunto de *tweets.* Los resultados de la tabla 1 reportan que *k** sufre pocas variaciones, alcanzando una media de 1469 segundos y una desviación estándar de 1062 segundos para los valores obtenidos según la función *S*, sin tomar en cuenta el valor aislado del primer día. Por lo que se puede resumir que este caso, dividir el *corpus* por períodos de tiempo, no representa un factor determinante al calcular el valor *k**. Sin embargo, en el caso de la tabla 2, los cambios son más notables. Las ventanas toman tamaños entre dos y quince horas para la función de evaluación *S*. Estos valores presentan una media de 35 209 segundos y una desviación estándar de 17 407 segundos. Por ello se puede concluir que el tipo de tendencias que contenga el *corpus* es un factor determinante para la selección de *k**.

Como resultado, se puede afirmar que escoger el tamaño de ventana es un proceso complejo, nada estable y extremadamente subjetivo. Este problema lo han tenido que afrontar todos los SDT, sin embargo, no ha sido solucionado. En su mayoría han optado por experimentar con un pequeño grupo de valores y elegir el que mejor resultado reporte. En cambio, las ventanas dinámicas siguen siendo un campo sin explorar. A continuación, se propone una metodología con la cual obtener mejores resultados en los SDT mediante el uso de ventanas que se ajusten al flujo de datos.

Tabla 1. Tamaños de ventanas óptimos reportados para el corpus dividido por días.

Fecha	S		F_2	
	K ópt	**T segs**	**K ópt**	**T segs**
20110301	5	140 853	5	140 853
20110302	694	432	404	742
20110303	985	610	880	683
20110304	133	2 700	440	820
20110305	878	706	579	1 070
20110306	641	1 228	476	1 653
20110307	138	3 141	305	1 427

Tabla 2. Tamaños de ventanas óptimos reportados para el corpus dividido por categoría de tendencia.

Categoría	S		F_2	
	K ópt	**T segs**	**K ópt**	**T segs**
meme	113	8 365	920	1 035
ongoing event	21	41 665	730	1 253
news	27	34 404	62	15 290
commemorative	14	56 402	159	5 287

DISTRIBUCIÓN DE VENTANAS NO ESTÁTICAS

Se busca encontrar distribuciones de ventanas que alcancen una mejor evaluación que las distribuciones de ventanas estáticas. Para ello se podría modelar el problema como un problema de optimización donde se va a maximizar la función *S* definida en (1), tal que una distribución de ventanas contenga todos los elementos del *corpus* y no contenga ventanas vacías:
Son características de este problema:

$$max\ S(d)\ s.a.: \qquad (i)\ \sum_{v \in d} |v| = T \qquad (ii)\ v \neq \emptyset,\ \forall\ v \in d \qquad (2)$$

- Las instancias de entrada de este problema, distribuciones de ventanas, son de gran tamaño. En el capítulo anterior se pudo ver cómo una distribución de ventanas puede llegar a tener 1500 ventanas o más, dependiendo del espacio temporal que comprende el *corpus*.
- El espacio de búsqueda es extremadamente grande. En el caso del *corpus* utilizado posee 50 355 *tweets,* por lo que el espacio de búsqueda estaría compuesto por todas las posibles distribuciones de ventanas que no contengan ventanas vacías.
- No es necesario llegar a un óptimo global. Las soluciones aproximadas son aceptables, ya que, el valor óptimo no es conocido.

Por lo que teniendo en cuenta la amplitud del espacio de soluciones, la calidad de las evaluaciones esperadas y la complejidad y dificultad del problema, se propone un modelo metaheurístico como método de solución.

Solución Computacional

Se seleccionó el algoritmo *Escalador de colinas* (*Hill Climbing)* como modelo metaheurístico para proponer una solución al problema de optimización definido en (2) (Linares, *et al.* 2018). Una ejecución de la metaheurística recibe una solución inicial y a partir de esta, genera un grupo de distribuciones de ventanas vecinas. Cada solución vecina es evaluada y de ellas se selecciona la que mejor evaluación de la función objetivo obtiene. Después de cierto número de iteraciones realizando estas operaciones, se retorna una solución final (Talbi, 2018).

El operador de vecindad está compuesto por tres acciones: clonar, unir y dividir. Las cuales son aplicadas con cierta probabilidad a cada ventana de la distribución de ventanas. De esta manera se obtienen distribuciones de ventanas vecinas, con gran probabilidad de ser diferentes.

En la figura 2 se puede observar la distribución de ventanas inicial con que se realizó una ejecución de *Hill Climbing,* una de las soluciones intermedias del algoritmo y la distribución de ventanas final, con evaluaciones 0.271, 0.327 y 0.351 respectivamente. La imagen muestra (fig. 2-c) una distribución de ventanas no estática, totalmente distinta a la inicial (fig. 2-a), con una menor cantidad de ventanas, un menor número de momentos de detección de tendencias y una mejor evaluación. Estos elementos pudieran ser un primer indicador del adecuado comportamiento de la optimización.

Para un estudio más detallado del funcionamiento de la optimización se ejecutó la metaheurística *Hill Climbing* 30 veces sobre el *corpus* general de los 70 tópicos. Cada ejecución con 20 vecinos, 300 iteraciones, y como solución inicial, la distribución de ventanas estática obtenida a partir de $k^* = 20$. Un conjunto de experimentos fueron definidos para evaluar y comparar el comportamiento de esta metodología con respecto a la metodología para obtener ventanas estáticas.

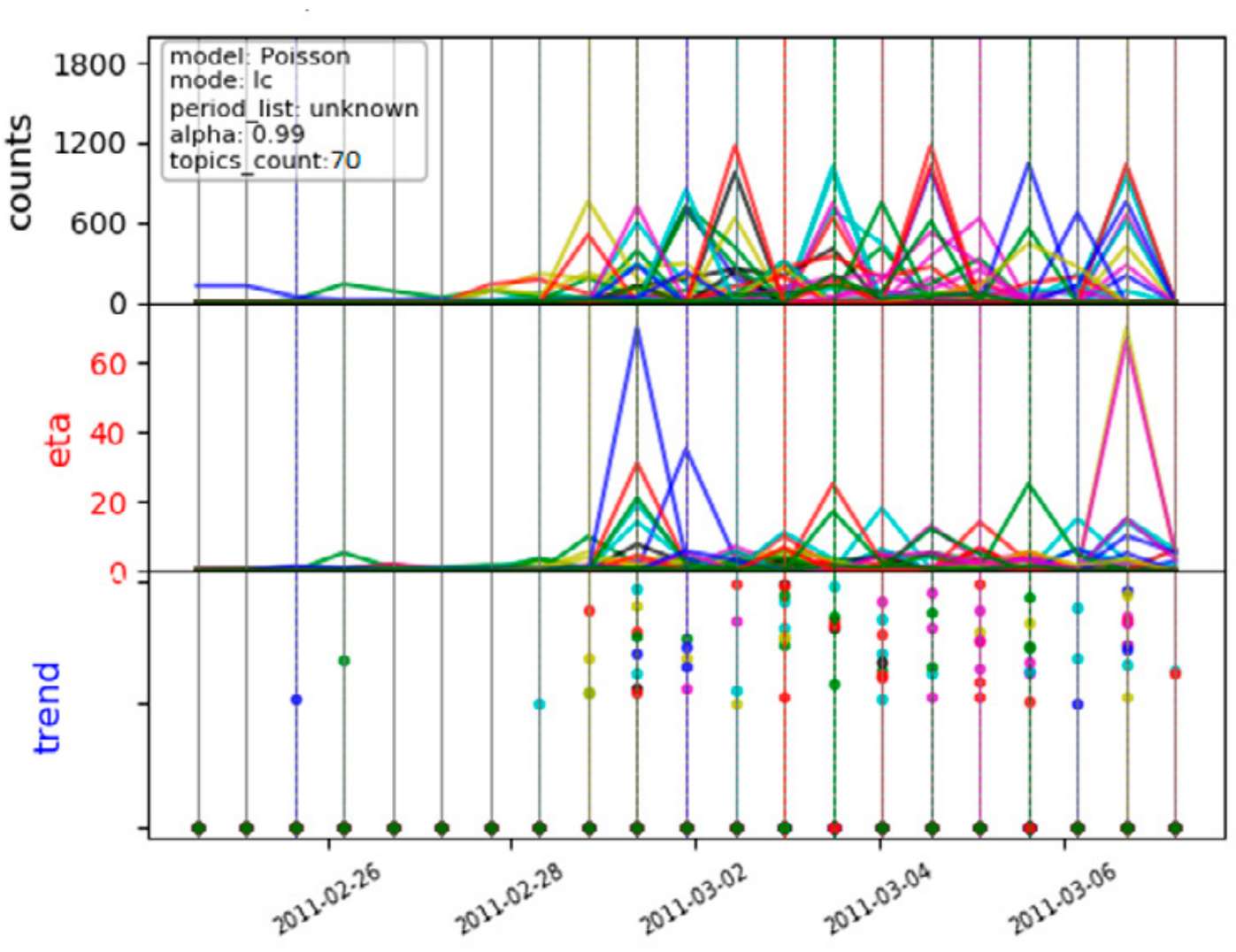

Figura 2a: Distribuciones de ventanas obtenidas durante el proceso de optimización de una ejecución de *Hill Climbing*. Distribución de ventanas inicial.

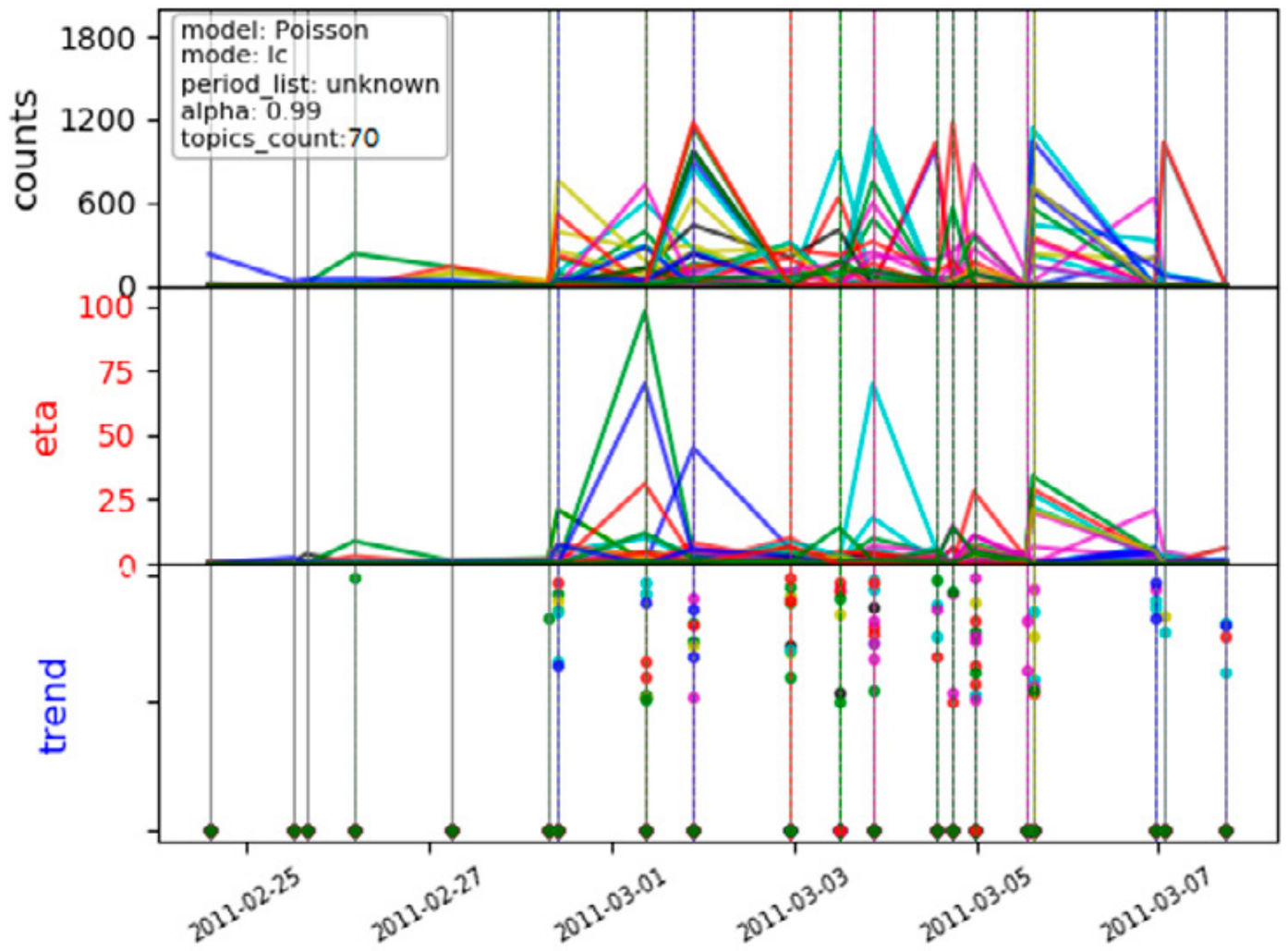

Figura 2b: Distribuciones de ventanas obtenidas durante el proceso de optimización de una ejecución de *Hill Climbing*. Distribución de ventanas intermedia.

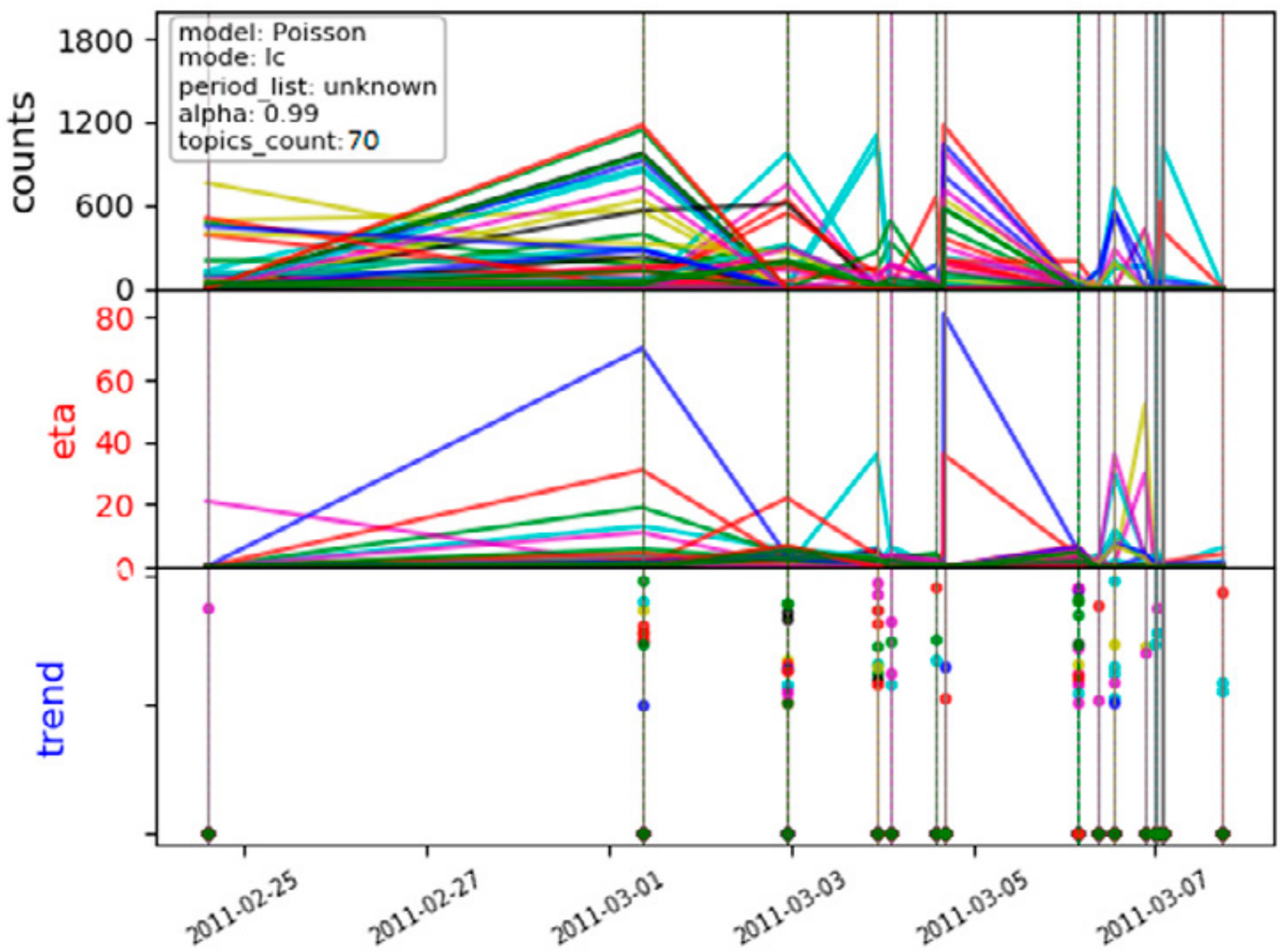

Figura 2c: Distribuciones de ventanas obtenidas durante el proceso de optimización de una ejecución de *Hill Climbing*. Distribución de ventanas final.

Convergencia de la optimización

Como primer paso es importante analizar si la metaheurística converge. Para ello se realizó un estudio de la iteración donde se alcanza el valor óptimo. Además, se observó el comportamiento de la diferencia relativa de las evaluaciones entre iteraciones consecutivas. Efectivamente, se comprobó que, como promedio, a partir de la iteración 241 de 300 la diferencia relativa entre soluciones consecutivas no suele ser mayor que 0.005. Este pequeño valor no es más que el 10% de la diferencia máxima, lo cual ratifica el carácter decreciente de la función. Además, se determinó que la optimización converge, puesto que se obtiene la mejor solución encontrada, aproxima-damente, en la iteración 194, con una desviación estándar de 80 iteraciones.

Estructura de las distribuciones de ventanas generadas

Las nuevas distribuciones de ventanas no son obligatoriamente estáticas. Por lo que se analizó la distribución de la variable aleatoria tamaño de ventana para descubrir la estructura de las nuevas distribuciones de ventanas. Los tamaños de ventanas en segundos de las 30 distribuciones de ventanas resultantes, en su mayoría, tienen un tamaño entre cero y 17 000 segundos aproximadamente, pero existen otros grupos de ventanas de mayor tamaño, lo que hace que esta variable aleatoria no se comporte de forma constante. Además, se calculó la media y la desviación estándar de la muestra, y los valores alcanzados fueron 52 783 y 77 038 segundos respectivamente. Puesto que la desviación estándar es de 21 horas aproximadamente, un amplio rango de variabilidad de los tamaños de ventanas, se puede apreciar cómo estos valores afirman la no uniformidad de los tamaños de ventanas generados por la metaheurística.

Evaluación de las distribuciones de ventanas

Finalmente se debe determinar si los resultados de la metaheurística son significativamente mejores. Para ello se estudió el comportamiento de los resultados de la metaheurística respecto a las evaluaciones de las distribuciones de ventanas estáticas. La prueba realizada fue un *z-test,* la cual analiza si una muestra de datos *A* presenta una distribución normal tomando como hipótesis nula que un valor *v* es la media de la distribución. Como lo que se deseaba era comprobar con qué probabilidad las evaluaciones finales de la metaheurística son peores o iguales que la evaluación del k^*, se definió la muestra *A* como el conjunto de las evaluaciones de las 30 distribuciones de ventanas resultantes, y *v* es la evaluación de la distribución de ventana generada a partir del k^*=20. A raíz de esto se estableció como hipótesis alternativa que la diferencia en promedio es mayor que el valor dado. El resultado obtenido fue un *p_valor* = 5.829832343873788e - 96, valor extremadamente bajo, que hizo rechazar la hipótesis nula y aceptar la alternativa, es decir, que la media de *A* es mayor que *v*. Por lo que es muy baja la probabilidad de obtener, con la metaheurística, soluciones peores o iguales que la evaluación de la distribución estática.

De manera general, se obtuvieron resultados satisfactorios. Las hipótesis asumidas fueron respaldadas a partir de los experimentos. La metodología propuesta para generar ventanas no estáticas reporta resultados más precisos que la metodología de las ventanas estáticas. En consecuencia, se obtuvo un proceso de optimización que converge y las ventanas generadas presentaron una estructura no uniforme.

CONCLUSIONES

En la presente investigación se estudió la influencia de las ventanas de tiempo sobre SDT basados en características. Se realizaron un grupo de experimentos analizando el comportamiento del sistema ante el cambio del tamaño de las ventanas. Al evaluar los resultados, se comprobó que la configuración de las ventanas es un factor determinante para la efectividad de este tipo de sistemas. Como ya era sospechado, determinar el tamaño de las ventanas es un proceso extremadamente complejo.

Ante la necesidad de dar una posible solución a este problema, se definió una metodología para generar configuraciones de ventanas que tuvieran en cuenta el flujo de los datos. El problema fue modelado como un problema de optimización y para la solución computacional se propuso la implementación de la metaheurística *Hill Climbing*.

Los resultados de esta metodología fueron satisfactorios, las distribuciones de ventanas dependientes del flujo presentaron una estructura no uniforme y las evaluaciones de las soluciones resultaron ser más efectivas que las distribuciones estáticas.

Esta investigación posee muchas aristas en las que seguir profundizando. Sin embargo, es un avance en el perfeccionamiento de los SDT, sistemas que van tomando gran relevancia en la digitalización de la sociedad cubana. La detección de catástrofes o accidentes, la recomendación de productos y el análisis de impacto de las nuevas políticas, son algunas de las áreas donde se puede reconocer la aplicación de este tipo de sistemas. Se recomienda profundizar en el ajuste de parámetros, definir nuevos operadores de búsqueda que tengan en cuenta la estructura de los datos, y experimentar con conjuntos de datos de mayor volumen.

REFERENCIAS

Abdelhaq, H., Sengstock, C. & Gertz, M. (2013). Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment,* vol. 6, no. 12, pp. 1326–1329.

Allan, J. (2002). Introduction to topic detection and tracking. *Topic detection and tracking,* pp. 1–16.

Allan, J., Carbonell, J.G., Doddington, G., Yamron, J. & Yang, Y. (1998). Topic detection and tracking pilot study final report. *Carnegie Mellon University.*

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence,* 31(1), 132-164.

Bollen, J., Mao, H. & Pepe, A. (2011). *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Icwsm,* vol. 11, pp. 450–453.

Figueira, Á., Guimarães, N. & Torgo, L. (2018). *Current State of the Art to Detect Fake News in Social Media: Global Trendings and Next Challenges.* In WEBIST, pp. 332-339.

Fiscus, J. G. & Doddington, G. R. (2002). *Topic detection and tracking evaluation overview.* In J. G. Fiscus, & G. R. Doddington, Topic detection and tracking. Springer, pp. 17-31.

Guzmán, J. & Poblete, B. (2013). On-line Relevant Anomaly Detection in the Twitter Stream: An Efficient Bursty Keyword Detection Model. En: event-place: Chicago, Illinois, *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description* [en línea]. New York, NY, USA: ACM, pp. 31–39. ISBN 978-1-4503-2335-2. DOI 10.1145/2500853.2500860. Disponible en: http://doi.acm.org/10.1145/2500853.2500860.

Hasan, M., Orgun, M. A. & Schwitter, R. (2018). *A survey on real-time event detection from the twitter data stream.* Journal of Information Science, 44(4), pp. 443-463.

Hendrickson, S., Kolb, J., Lehman, B. y Montague, J. (2015). Trend detection in social data. *Twitter Blog.*

Laguna, J.O., Olaya, A.G. & Borrajo, D. (2011). A dynamic sliding window approach for activity recognition. *International Conference on User Modeling, Adaptation, and Personalization.* S.l.: Springer, pp. 219–230.

Lau, J.H., Collier, N. & Baldwin, T. (2012). On-line Trend Analysis with Topic Models:$\backslash backslash$# twitter Trends Detection Topic Model Online. COLING. S.l.: s.n., pp. 1519–1534.

Lee, R., & Sumiya, K. (2010). *Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection.* In Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks. New York, NY, USA: ACM, pp. 1-10.

Linares, R.C., Morffis, A.P. & Cruz, Y.A. (2018). *Distribución de Ventanas en Sistemas de Detección de Tendencias: Análisis y Propuesta para un Modelo de Configuraciones.* Licenciatura. S.l.: Universidad de La Habana.

Mathioudakis, M. & Koudas, N. (2010). Twitter Monitor: Trend Detection over the Twitter Stream. En: event-place: Indianapolis, Indiana, USA, *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data* [en línea]. New York, NY, USA: ACM, pp. 1155–1158. ISBN 978-1-4503-0032-2. DOI 10.1145/1807167.1807306. Disponible en: http://doi.acm.org/10.1145/1807167.1807306.

Mederos, O., Hernández, A. & Almeida-Cruz, Y. (2013). *Detección de tópicos en Twitter.* Tesis de Licenciatura, Facultad de Matemática y Computación, Universidad de La Habana.

Montes, M., Gelbukh, A. & López, A.L. (2001). Mining the news: trends, associations, and deviations. *Computación y Sistemas,* vol. 5, no. 1, pp. 14–24.

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C. & Ounis, I. (2012). *Bieber no more: First story detection using Twitter and Wikipedia.* In SIGIR 2012 Workshop on Time-aware Information Access.

Ozdikis, O., Senkul, P., & Oguztuzun, H. (2012). Semantic expansion of tweet contents for enhanced event detection in twitter. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 20-24). IEEE.

Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D. & Sperling, J. (2009). Twitterstand: news in tweets. *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems.* S.l.: ACM, pp. 42–51.

Talbi, E.-G., 2009. *Metaheuristics. From Desing to Implementation.* New Jersey: Wiley. John Wiley & Sons, Inc., Publication.

Tejeda, R.G. & Cruz, Y.A. (2016). *Detección de tendencias en Twitter.* Licenciatura. S.l.: Universidad de La Habana.

Twitter Usage Statistics. *Internet Live Stats* [en línea], (2019). [Consulta: 10 diciembre 2019]. Disponible en: https://www.internetlivestats.com/twitter-statistics/#source.

Weiler, A., Grossniklaus, M., & Scholl, M. H. (2015). *Evaluation measures for event detection techniques on twitter data streams.* In British International Conference on Databases. Springer, Cham, pp. 108-119.

Wolfe, L. (2019). Twitter User Statistics for 2019. *The Balance Careers* [en línea]. [Consulta: 10 diciembre 2019]. Disponible en: https://www.thebalancecareers.com/twitter-statistics-2008-2009-2010-2011-3515899

Xun Wang, J.J., Feida Zhu & LI, S. (2011). Real Time Event Detection in Twitter. *Springer-Verlag,* vol. 21.

Zubiaga, A., Spina, D., Martinez, R., & Fresno, V. (2014). Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology,* vol. 66, no. 3, pp. 462–473.