Google *Flu Trends* y los datos masivos: ¿extrapolable al ébola?

Pilar José López, UCM, España

Resumen: *Millones de personas navegan en Internet a través del buscador de Google. Esta compañía aprovechando la información de sus usuarios desarrolló Google Flu Trends en 2008. Esta herramienta nace con el objetivo de recabar datos para obtener la incidencia de la gripe en un país determinado con una gran precisión. Esta aplicación registra las consultas que hacen los cibernautas a través de su buscador Google y con esos datos obtiene sus propias conclusiones, como si de un estudio de epidemiología se tratara. Tres años más tarde del desarrollo de esta herramienta, en 2011, la información que ofrecían los datos no se asemejaban a la realidad. ¿Qué había pasado? El Periodismo de datos estaba fallando. Muchos usuarios que no padecían la gripe buscaban información en Internet y Google Trends Flu los contaba cómo enfermos. Con este paper se pretende analizar esta herramienta y comparar su evolución y resultados con la enfermedad del ébola.*

Palabras clave: *periodismo de datos, Google Trends Flu, España, ébola, investigación*

Abstract: *Millions of people surf the Internet through the Google search engine. This company leveraging in- training your users Google Flu Trends developed in 2008. This tool was created with the aim of collecting data for the incidence of influenza in a country with high precision. This application records queries that netizens through its search engine Google and the data obtained their own conclusions, as if from a study of epidemiology is involved. Three years later the development of this tool, in 2011, the information you offer data did not resemble reality. What had happened? The Data Journalism was failing. Many users who did not have the flu seeking information on the Internet and Google Flu Trends counted them how sick. With this paper is to analyze this tool and compare their progress and results with Ebola disease.*

Keywords: *Data Journalism, Google Trends Flu, Spain, Ébola, Research*

Introducción

Cada minuto se realizan en el mundo dos millones de búsquedas en Internet. Los usuarios navegan por el espacio cibernético en busca de información utilizando diferentes buscadores, el más común y utilizado en esta era digital es *Google Chrome*. Este navegador mueve cada día una media de 25 petabytes de datos, es decir, un millón de gigabytes. Las nuevas tecnologías están cambiando la manera de vivir de la sociedad. Según Youtube (2015), su plataforma cuenta con más de mil millones de usuarios que suben cada minuto a Internet 300 horas de vídeo. La red ha revolucionado la rutina diaria de miles de personas sin que ellas sean conscientes, ya que no sólo los usuarios son los que consiguen información a través de un navegador, sino que también obtienen datos las compañías que operan con los buscadores.

En esta nueva forma de entender el mundo, Google resulta ser uno de los grandes beneficiados con su buscador. Según la última encuesta de la AIMC (Asociación para la Investigación de Medios de Comunicación) seis de cada diez españoles usan el navegador *Google Chrome*. Lo que no saben algunos consumidores de Internet es que cuando realizan una búsqueda a través de Chrome, Google recopila esa información en una gran base de datos. Esos conjuntos de datos forman parte de lo que se denomina "Big Data", definido según la empresa tecnológica americana *Gartner* como

> activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones. (Pries, 2013; 62)

Revista Internacional de Tecnología, Ciencia y Sociedad
Volumen 5, Número 1, <http://tecnociencia-sociedad.com>, ISSN 2530-4895

Republicado de *Revista Internacional de Tecnología, Conocimiento y Sociedad* 5(1), 2016 (pp. 1-8)

Estos grandes volúmenes de datos pueden ser de cualquier índole y procedencia, pueden estar estructurados o no, es decir, que pueden encontrarse organizados en bases de datos para su posterior análisis o pueden hallarse sin clasificar dificultando su tratamiento. Pero, ¿cuáles son las fuentes de información de big data más habituales para una empresa? Según el último estudio del Cisco (2014), las compañías utilizan las fuentes de datos que generan ellas mismas.

Una de las fuentes de información de Google son las que el investigador *Meyer* denomina '*desechos de datos'*, que son aquellos términos o palabras clave que un usuario introduce en el buscador de Google para realizar una consulta. A través de estos *'desechos de datos'*, Google puede llegar a conocer nuestra vida, nuestros gustos y la información que requerimos en cada momento. Google ha sabido aprovechar esa información que ofrecen los usuarios de forma gratuita reutilizándola para su procesamiento y análisis. De estos datos masivos, Google creó *Flu Trends*.

Flu trends

Google conocía el potencial del que disponía con la información que recuperaba de las búsquedas hechas con *Google Chrome*, y en 2008 realizó un mapa que mostraba en tiempo real la propagación del virus de la gripe (H1N1) en Estados Unidos llamado *Flu Trends*. Esta aplicación perseguiría predecir con exactitud la expansión del virus en el país y nacería en respuesta a la existente y tardía identificación del brote y a una larga espera de tiempo en contener esta enfermedad en el país. Esta herramienta, actualmente en funcionamiento continúa ofreciendo datos actualizados de la expansión del virus, frente a los datos que ofrece el *Centro para el Control y la Prevención de Enfermedades de EEUU* (CDC) que tarda entre 7 y 15 días en obtener los datos de la gripe. *Flu Trends* se convertiría en lo que muchos investigadores entienden como infodemiología

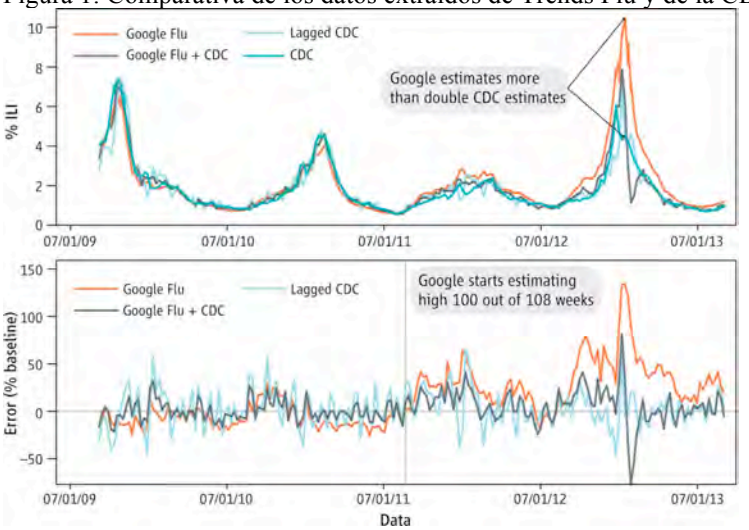
> Ciencia que evalúa la distribución y los factores determinantes de la información por internet o en una población, con el fin de recolectar información que permita evaluar la salud pública y las políticas estatales en el área. Supera a su ciencia madre, la epidemiología, encargada tradicional de evaluar la distribución y los factores determinantes de la enfermedad en las poblaciones, al ofrecer un método novedoso de consecución de la información, la "infovigilancia", proceso que permite la observación de los eventos en el instante mismo en que van ocurriendo, o incluso antes, mediante el análisis continuo y estructurado de la información disponible en Internet. (Roncancio, 2009)

La herramienta era sencilla de realizar, ya que varios estudios corroboraron la relación que se establecía entre la cantidad de informaciones que se buscaban en el motor de búsquedas de Google sobre la gripe en Estados Unidos y el aumento de esta enfermedad en una determinada zona del país. Este navegador Web se llegaba a convertir en el denominado "*Dr. Google"*, muchos pacientes tendían a consultar información en Internet cuando creían tener síntomas del virus, antes que acudir a una consulta médica. Google conociendo esa información fue recabando las búsquedas realizadas por palabras clave como 'síntomas gripe', 'virus gripe' y 'enfermedad gripe' y la localización geográfica de los usuarios a través de la IP del usuario y así mediante datos *no estructurados* se construyó *Flu Trends*. Cuando se notaba un incremento en el número de consultas en una determinada zona, Google hacía saltar la alarma y predecir una epidemia del virus.

El éxito de *Flu Trends* no se hizo esperar. En 2009 se adelantó al CDC en dos semanas, por las informaciones recogidas de sus usuarios en la red y ayudó a frenar una pandemia del virus H1N1 en Estados Unidos. Esta plataforma se fue expandiendo a otros países, como España. Pero lo que al principio parecía una herramienta muy útil y lo que se mostraba como una gran relación entre la ciencia y el big data, resultaría ser un fracaso.

La falta de rigor científico y la escasa actualización de los algoritmos de selección de las palabras hicieron que en la temporada 2012-2013 *Flu Trends* pronosticara el doble de visitas a los centros hospitalarios por el virus H1N1 que la CDC. Un estudio realizado por los investigadores estadounidenses David Lazer, Ryan Kennedy, Gary King y Alessandro Vespignani corroboró que de 108 semanas de análisis *Google Flu Trends* se equivocó 100 semanas.

Figura 1: Comparativa de los datos extraídos de Trends Flu y de la CDC



Fuente: Lazer, 2014.

También el estadístico Kaiser Fung analizó los datos de David Lazer, Ryan Kennedy, Gary King y Alessandro Vespignani mostrados en la *figura 1* y comprobó que se necesitaban cambiar los algoritmos, ya que las tendencias de los usuarios iban cambiando y la aplicación acertaba pocas veces. También discrepó en la forma en la que Google ocultaba el método a seguir para extraer los datos y pronosticar el brote del virus. Google no explicaba en que se basaba para ofrecer la información, ya que los datos que obtenían no estaban limpios, no eran estables y muchos menos se encontraban estructurados:

> La cantidad de datos todavía tiende a dominar la discusión de valor de grandes datos. Pero más datos no conducen a un mejor análisis, tal como se desprende de la evolución de la gripe. Los grandes conjuntos de datos no garantizan conjuntos de datos válidos. (Fung, 2014)

Fung también argumenta que aunque Google abarque el 80% del mercado de búsquedas no se garantiza que toda la información esté completa. A pesar de su fracaso en EEUU entre 2012 y 2013 Google no ha parado en su idea de controlar el virus H1N1 en el mundo mediante big data. Actualmente *Trends Flu* se encuentra en 29 países. En España se puso en funcionamiento en 2009 después de poder contrastar la información con el Centro Europeo para la Prevención y el Control de las Enfermedades (*CEPCE*) y corroborar que había una cierta correlación con las búsquedas realizadas por los españoles en el buscador y la incidencia del virus en el país.

Figura 2: Flu Trends

Fuente: Flu Trends, 2015.

La epidemia del ébola

El ébola ha sido el último brote epidémico que ha surgido a nivel mundial. El ébola tuvo su origen en 2013 en Guinea y se fue extendiendo a muchos países saliendo del continente africano, entre ellos España. El ébola aún no controlado en algunos países de África ya ha dejado según la Organización Mundial de la Salud, más de 26.500 afectados y más de 11.000 muertes en todo el mundo desde su inicio en diciembre de 2013.

En el caso concreto de España, entre 2013 y 2014 se activó el protocolo contra el ébola en 48 ocasiones por casos sospechosos. Los españoles empezaron a conocer qué es el ébola y cómo actúa este virus a partir del 7 de agosto de 2014 cuando el Gobierno español repatrió desde Liberia, al religioso Miguel Pajares, el primer español infectado por el virus, que falleció el 12 de agosto convirtiéndose en la primera víctima de Europa. El 22 de septiembre, España repatrió de Sierra Leona a otro religioso infectado por el virus, que tres días más tarde falleció en el hospital. El 6 de octubre una de las enfermeras que trató a ambos afectados por el ébola, Teresa Romero, ingresó en el hospital por haber contraído la enfermedad. 21 personas que se encontraban en contacto con la enfermera son también puestas en cuarentena. Ninguno de ellos presentó tener el ébola y abandonaron el hospital después de estar en observación tres semanas. Finalmente Teresa Romero superó el virus y el 5 de noviembre salió del hospital totalmente recuperada.

En España las autoridades políticas y sanitarias y los medios de comunicación fomentaron un alarmismo entre la ciudadanía.

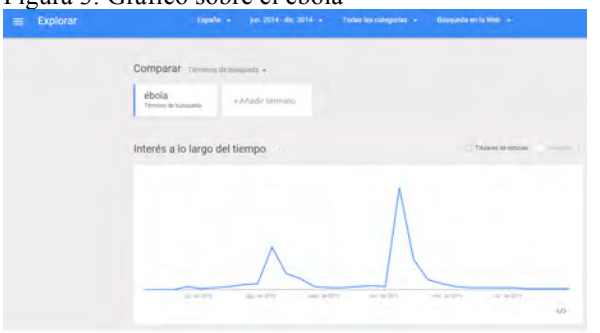
> El "efecto Guadiana" de la información sanitaria: primero, crisis; luego, alarma sanitaria; después, desinformación. La noticia sanitaria es un producto fugaz, efímero. Interesa cuando adquiere la categoría de alarma, pero después el ciudadano lector debe rebuscar en el interior del diario para encontrar algún reportaje sanitario, una reducida información sobre salud o una columna en página par. Explota como alarma y desaparece cuando se cronifica. (Jurado, 2015)

Enrique Jurado profesor investigador en ESADE explica como la falta de información por parte de las autoridades y la ausencia del rigor informativo en algunos medios de comunicación convirtieron el problema del ébola en un espectáculo televisivo.

¿Google *Flu Trends* podría ser extrapolable al ébola?

El objetivo de este paper es realizar un estudio comparativo del virus del ébola y del virus H1N1 analizando los datos masivos que Google recogió sobre el ébola, a través de *Google Trends*, correspondientes a los meses de junio a diciembre de 2014, en los que se desarrolló esta cepa en España y demostrar que estos volúmenes de datos no son tan fidedignos como aparentan. El término introducido en *Google Trends* ha sido ébola y hemos puesto como ámbito geográfico toda España. Observamos los siguientes gráficos:

Figura 3: Gráfico sobre el ébola



Fuente: Google Trends, 2015.

Figura 4: Mapa sobre la influencia del ébola

Fuente: Google Trends, 2015.

Figura 5: Búsquedas relacionadas sobre el ébola

Fuente: Google Trends, 2015.

Después de analizar las diferentes búsquedas realizadas se extraen los siguientes aspectos:

1. No existe ninguna tendencia sobre consultas de información entorno al ébola en los meses previos al 3 de agosto en España.
2. Del 3 al 9 de agosto como se aprecia en la *figura 3*, coincidiendo con la repatriación del primer religioso, Miguel Pajares, se observa como hay una gran presencia de consultas sobre el ébola.
3. El pico más alto se registra a principios de octubre coincidiendo con el ingreso de Teresa Romero en el hospital, la primera víctima infectada por el ébola en Europa. Como vemos hay un gran aumento de consultas sobre el virus y sobre las últimas noticias respecto a la enfermedad.
4. No todas las comunidades autónomas registran el mismo número de consultas. Madrid, Extremadura, Canarias y Castilla La Mancha son las comunidades en las que los usuarios realizaron más búsquedas sobre el ébola. Las palabras clave que utilizaron los usuarios fueron: 'el ébola', 'ébola España', 'enfermedad ébola'… Lo podemos ver en *las figuras 4 y 5*.

Conclusión

Si comparamos la incidencia del virus H1N1 con la incidencia del ébola está claro que no todo el mundo que consultaba información en el motor de búsqueda *Google Chrome* estaba infectado por el virus del ébola. Sólo una persona había contraído el ébola en España y el número de consultas en Internet sobre este virus se había disparado. Según datos de Google (Zeitgeist 2014) la palabra "ébola" se posicionó en el cuarto puesto de los términos más buscados en 2014. Lo que nos muestra esta

comparativa es que el miedo de la población ante la falta de información y el pánico creado por los medios de comunicación, entre otros factores, hicieron que los usuarios navegaran por la red en busca de información, pero sin padecer esta enfermedad ninguno de ellos.

Este dato también lo corrobora Gustavo Eduardo en la publicación *Ayudas desde la red para el control de la epidemia*, en el que explica que *Google Trends* que se basa en la misma metodología que *Flu Trends* no sirve como ciencia exacta:

> Debe señalarse que esta novedosa forma de pesquisa de epidemias no ha sido diseñada para reemplazar a las redes de vigilancia tradicional, sobre todo teniendo en cuenta que la investigación online no permite conocer datos demográficas, y que en el caso de una pandemia se recolecta información tanto de la población enferma como de la sana, la cual, producto del pánico, aumenta el número de consultas y la sobreestimación de los datos. (Eduardo, 2009)

Como hemos visto anteriormente *Flu Trends* falló en Estados Unidos. El motivo fue una de las variables que consideraron para la aplicación; el número de búsquedas realizadas por los usuarios, algo que si lo comparamos con el caso del ébola no resulta ser una variable fiable. No todo el mundo que consultó "gripe" y "virus H1N1" tenían esta enfermedad como se ha visto anteriormente. Con esta comparación se quiere demostrar que aún el Big Data se encuentra en fase de desarrollo y queda mucho por hacer para una correcta gestión y utilización de estos macrodatos. Los grandes volúmenes de información tienen muchas interpretaciones y como hemos podido comprobar con la herramienta *Google Trends* aplicado al ébola, no todos los *desechos de datos*, nos sirven siempre para obtener información con un rigor científico y exacto, ni mucho menos para predecir en base a un número de consultas la incidencia de una enfermedad, porque muchas personas pueden buscar información sobre una enfermedad y no padecerla**.**

REFERENCIAS

Casacuberta, D. (2013). Innovación, Big Data y Epidemiología. *Revista Iberoamericana de Argumentación, 7.*

Cook, S., Conrad, C., Fowlkes, A.L., (2011). Assessing Google flu trends perfor-mance in the United States during the 2009 influenza virus A (H1N1) pandemic. Disponible en: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023610

Fung, K. (2014). *Google Flu Trends' Failure Shows Good Data > Big Data*. Harvard Business Review. Disponible el recurso online: https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data/#signin

Ginsberg, J.; Mohebbi, M.H. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(19). Disponible el recurso online: http://dx.doi.org/10.1038/nature07634

Jurado, E. (2015). Los errores de comunicación en la crisis del ébola. *Cuadernos de Periodistas, 29.*

Lazer, D., Kennedy, R., King, G. y Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science, 343*(6176), pp. 1203–1205.

Pries, K.H. & Dunnigan, R. (2015). *Big Data Analytics: A practical guide for managers.* Boca Raton: CRC Press

Roncancio, G. E. (2009). Ayudas desde la red para el control de la epidemia. *Infectio, 13*(3), pp. 217-222. Retrieved March 23, 2015. Disponible el recurso online: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123939220090003000009&lng=en&tlng=es.

Tascón, M. (2013). Introducción: Big Data. Pasado, presente y futuro. *Revista Telos, 95.*

Valdivia Pérez, A., Benito M. A., Escortell Mayr, E. (2010) ¿Se puede predecir la epidemia de gripe mediante datos de búsquedas en Internet? Gac Sanit.

Viktor, MS. & Kenneth, C. (2013). *Big data: a revolution that will transform how we live, work, and think.* British: Hodderand Stoughton.

SOBRE LA AUTORA

Pilar José López: Licenciada en periodismo por la universidad de Murcia (2007) e ingeniera técnica en Informática de Sistemas por la universidad de Almería (2005). Ha realizado el máster en Estudios Avanzados en Comunicación por la universidad de Murcia y el máster de Periodismo de Televisión del Instituto de Televisión Española. Ha trabajado durante más de cinco años en medios audiovisuales como TVE, Telemadrid, 7 Región de Murcia o Thader Televisión. Actualmente se encuentra realizando su doctorado en periodismo de datos en la universidad Complutense, bajo la tutela del profesor Dr. D. Jesús Miguel Flores. Sus estudios de doctorado están enfocados al periodismo de precisión y a la visualización de los datos por ser ámbitos que conoce por su trayectoria como informática y periodista. Ha publicado varias comunicaciones relacionadas con las nuevas tecnologías, entre las que destacan el "Uso Político del Blog" en el VI Congreso Internacional de Periodismo en la Red de la UCM; "De viaje por Internet", X Congreso de Periodismo Digital de la universidad de Huesca y "Ciudades Digitales: ciudadanía en red", de la universidad de La Laguna.