



Boletín GripeNet Temporada 2, 4º entrega.

Big Data y Epidemiología: Google flu trends (2): ¿Cómo funciona?

En la anterior entrega de nuestro boletín antes de las vacaciones navideñas, les hablábamos de **Google flu trends**, la herramienta para la monitorización de la gripe desarrollada por el gigante informático Google en colaboración con el organismo estadounidense para el control y la prevención de enfermedades (CDC).

El funcionamiento de Google flu trends se basa en registrar el número de búsquedas relacionadas con la gripe que se producen en una determinada área geográfica en cada momento, para, a partir de ahí, obtener una predicción cuantitativa del nivel de incidencia de la gripe en ese momento y lugar. En este boletín, explicaremos los principales problemas que presenta este tipo de análisis y la manera en la que Google flu trends los afronta.

Una búsqueda en Google no indica necesariamente enfermedad.

El principal escollo de este tipo de enfoque de vigilancia epidemiológica lo resume perfectamente el título de esta sección: el número de búsquedas en Google relacionadas con la gripe no es un indicativo directo del número de casos de gripe, sino una variable correlacionada, en el mejor de los casos, con la incidencia gripal.

De hecho, el principal problema a resolver es qué tipo de búsquedas han de considerarse estrictamente relacionadas con los casos de gripe y cuáles han de ser descartadas. Para ello, podríamos restringir nuestras búsquedas a aquellas que contengan la palabra “gripe”, pero quizás el modelo se comportaría mejor si contemplásemos otras búsquedas relacionadas con términos más específicos como: “fiebre”, “dolor muscular”, “congestión”, “resfriado”, “síntomas gripales”, etcétera. La mejor manera de distinguir qué términos se deben considerar para obtener un funcionamiento óptimo de este método, es analizando en retrospectiva aquellos términos que estén más correlacionados con los datos oficiales de incidencia gripal registrados por los organismos de salud en un periodo de tiempo concreto en el que los datos de búsqueda y de incidencia gripal hayan quedado registrados.

Con este propósito, el equipo de Google flu trends analizó una lista de más de 50 millones de términos y frases breves para encontrar, entre todos ellos, un total de 45 que eran los que proporcionaban la mejor correlación con el nivel de incidencia gripal registrado en el pasado, en concreto, con las temporadas comprendidas entre los años 2003 y 2006. De este modo, el principal escollo ya estaba solventado: identificar los términos o frases que las personas enfermas de gripe tienden a introducir en el buscador con una frecuencia muy superior a la que cualquier otra persona, que no padece la enfermedad, haría por mera curiosidad o por cualquier otro motivo.

La escala del mapa: del número de búsquedas al número de casos

Una vez se han identificado estos términos de búsqueda específicos, es necesario establecer una escala adecuada para obtener, a partir del número de búsquedas claves en un área determinada, una estimación del número de casos de gripe en la población.

Para ello, es necesario calibrar de nuevo los resultados comparándolos con las tasas de incidencia registradas por un sistema de vigilancia epidemiológica tradicional en cada área. Es importante señalar que este calibrado, responsable de definir la relación entre número de búsquedas y número de casos, no tiene por qué ser necesariamente igual en todas las regiones analizadas, sino que puede cambiar y depender de variables como la densidad de población en las distintas áreas, el grado de difusión de Internet en la población y, más concretamente, del grado de popularidad del propio buscador de Google.

Los investigadores de Google y el CDC estadounidense, lograron identificar estos problemas y resolverlos hasta construir una herramienta fiable, capaz de predecir los niveles de incidencia gripal con notable precisión y antelación. Sin embargo, el sistema ha experimentado una serie de problemas desde su puesta en marcha que le han llevado a protagonizar al menos dos sonados tropiezos en las temporadas gripales del 2009 y 2013, y que han llegado a poner en duda el verdadero alcance de este tipo de método de vigilancia epidemiológica pasiva. Sobre estos problemas, sus causas y limitaciones, hablaremos en la próxima entrega del boletín GripeNet.

El Big Data plantea en la actualidad una serie de retos a los que la sociedad y la tecnología tienen que hacer frente. La enorme cantidad de información generada cada día en la red demanda un uso correcto de la misma y la implementación de protocolos de seguridad fiables. El tratamiento, la interpretación y el análisis adecuado de toda esta valiosa e ingente cantidad de información es fundamental para el desarrollo de modelos de predicción y propagación de enfermedades, rumores, tendencias, entre otros. Es por ello que vuestra participación en GripeNet y los datos facilitados por cada uno de los usuarios en tiempo real son indispensables para el trabajo investigativo de nuestro grupo COSNET en aras de implementar modelos epidemiológicos y de predicción más precisos y fiables.

Para saber más: “Los fallos de Google Flu trends evidencian los problemas del Big Data” Artículo en “The Conversation” (inglés)

<http://theconversation.com/googles-flu-fail-shows-the-problem-with-big-data-19363>

