GABINETE TÉCNICO.
CENTRO DE ANÁLISIS Y PROSPECTIVA.

BOLETÍN DE ANÁLISIS Y SEGURIDAD INTERNACIONAL

2/2017

GABINETE TÉCNICO DE LA GUARDIA CIVIL
CENTRO DE ANÁLISIS Y PROSPECTIVA

Datos no tan grandes
La calidad frente al tamaño

Si quieres entender cómo viven los animales no vayas al zoológico, vete a la jungla.
Martin Lindstrom

Introducción

Los datos masivos son un tema de actualidad. Por ello, en artículos anteriores hemos comentado las ventajas que proporciona al análisis de inteligencia la combinación de la potencia computacional con la abundancia de datos de la era digital. Pero también hemos analizado algunos de los problemas que plantean tanto la confianza excesiva en los datos como la dependencia acrítica en los algoritmos que los procesan. Acabaremos esta línea argumental con algunas palabras para hacer un pequeño apunte que ponga en valor la importancia en la investigación de los pequeños datos, los indicios significativos o los llamados "datos de calidad".

Sabemos que nuestras interacciones diarias en Internet generan grandes cantidades de información que muestran nuestras preferencias de compra, nuestras actitudes y nuestras tendencias políticas. Esto ha supuesto un nuevo campo en el que los emprendedores buscan métodos nuevos de hacer dinero organizando y analizando todos esos datos. La comunidad de inteligencia busca a su vez extraer información relevante que permita perseguir e incluso anticipar todo tipo de delitos.

Hasta ahora, la discusión acerca de estos datos masivos ha oscilado de una forma un tanto acrítica, desde el entusiasmo al fervor religioso, llegándose a afirmar que el análisis de datos será el método dominante para evaluar casi todo. A partir de ahora, no necesitaremos nada más; sólo analizar de modo significativo los datos masivos[1].

Sin embargo hay un concepto paralelo. Los llamados "pequeños datos" están adquiriendo un carácter significativo en la actualidad. No tanto de forma contrapuesta sino como complemento al análisis de datos masivos. La importancia del detalle y de la observación cualitativa relevante tienen una trascendencia obvia para la seguridad, las ciencias forenses, criminalística, peritajes, etc..

Los datos pequeños o cualitativos son aquellos que se obtienen a partir de observaciones de informaciones relevantes, realizadas frecuentemente de forma personal, y que permiten captar detalles de gran importancia para la identificación y análisis de sucesos o tendencias[2].

Por ejemplo, si quisiéramos saber qué y porqué eligen los consumidores, lo mejor que podríamos hacer es ir a sus casas y reunir los detalles relevantes de su

[1] *The end of theory: the data deluge makes the scientific method obsolete*. Chris Anderson. Wired. 23/06/2008.
[2] *El milagro del 'Small Data': análisis de pequeñas cantidades de información que ayuda a grandes empresas*. Antena 3 Noticias. 26/08/2016.

comportamiento para estudiarlos y extraer el significado de los "pequeños datos" que nos muestran su vida cotidiana.

Pequeños datos

Los motores de búsqueda de las empresas de la tecnología de la información pueden detectar correlaciones significativas −algo que no implica causación− pero realmente pueden decir pocas cosas a nivel individual. Algunos buscadores, conscientes de esto, han acudido a consultores e investigadores para examinar lo que hay detrás de los datos masivos: los datos significativos o *small data*.

No hemos sabido que teníamos datos pequeños o que había datos de alta calidad hasta hace poco tiempo. Nos dimos cuenta de ello cuando, hace un par de décadas, nos inundaron los datos masivos y supimos que los mismos podrían ser tratados en bruto por su cantidad. **Pero lo que caracteriza a los datos pequeños** es precisamente que su volumen no es grande y por tanto son **accesibles, informativos y operativos**.

El hecho es que en la vida cotidiana todos dejamos cada día una serie de pequeñas cantidades de información que muestran cómo somos realmente −hábitos, rituales, gestos, preferencias…− y que no tienen por qué ser procesados por las empresas dedicadas a recolectar datos masivos en Internet. Aunque parezca difícil de concebir hoy en día y sea casi contraintuitivo, no todo está en la Telaraña Mundial y existen grandes límites en la capacidad analítica y predictiva de los datos masivos.

Los grandes datos tienen problemas… y puede que los pequeños datos puedan ayudar a superarlos. Las grandes bases de datos están definidas normalmente bajo criterios muy restringidos, muchas veces demasiado limitados para crear conocimiento. Los algoritmos también son, con mucha frecuencia, demasiado ajustados para lograr por simple aprendizaje, grandes avances.

Y los datos son datos. No debemos esperar que las redes neuronales extraigan cualidades emocionales de las bases de datos. La belleza, la amistad, el amor… y otros conceptos, pueden estar fuera del alcance de algoritmos devoradores de datos que anteponen el análisis ante la emoción. En algunos campos, sencillamente los algoritmos y los datos masivos son extremadamente incompetentes.

No se debe confiar en el procesamiento de datos masivos como una panacea o concebirlo como una especie de fe en el solucionismo tecnológico. A los datos masivos hay que añadir los datos relevantes −*rich data*− y los datos profundos −*deep data*− que, muchas veces, pueden venir a nosotros en forma de pequeños datos.

Cuando el análisis y la investigación se realizan desde la óptica de estos "pequeños datos" pueden abrirse perspectivas que habíamos dejado de lado en el contexto de la burbuja de la información. Evidentemente, un pequeño dato o una información anecdótica no es base suficiente para extraer un significado relevante o plantear una hipótesis. Pero sí puede combinarse con otras observaciones y, de esta forma, proporcionar la base de un conocimiento futuro.

No queremos decir en absoluto que los grandes datos no tengan importancia. Lo que es relevante aquí es que la integración de las grandes bases de datos con los datos

significativos, pequeños pero cualificados, es un ingrediente fundamental del análisis en los tiempos modernos.

Datos informativos

Juntar pequeños fragmentos de datos significativos, considerando de entre ellos cuales son los que tienen importancia, es algo parecido a un arte. Reunir pistas no es un proceso lineal, no es una minería automática guiada por algoritmos. Algunas pistas no llevarán a ninguna parte, otras serán significativas, otras potencialmente interesantes, la mayor parte… irrelevantes. Sin embargo, un dato específico puede ser lo suficientemente importante como para ser la fundación de un concepto, un avance o una nueva explicación. Puede ser la clave para predecir el futuro que se oculta tras la niebla de la incertidumbre.

Esta clase de información se consigue, no mediante el proceso analítico interminable de bases de datos enormes sino mediante la subversión de las reglas de las propias bases de datos. Si alguna vez vamos a un país extranjero, no debemos mirar estadísticas sino las señales inconscientes que nos ofrecen las gentes que viven en él. Tal vez es mejor simplemente mirar con un ojo perspicaz.

Maskenfreiheit

Analizar grandes cantidades de datos es como estudiar a una gran multitud que, en carnaval, se oculta tras una máscara. Esto es especialmente cierto en Internet. Aunque muchos comportamientos son compartidos y la individualidad se funde en el grupo, ¿podemos pretender que hemos desentrañado la personalidad, el comportamiento o los deseos de cada individuo? ¿Quién puede decir que sabe realmente lo que hay detrás de cada máscara?

En alemán “maskenfreiheit” es la libertad que se esconde tras la máscara y en nuestro contexto se refiere a la cantidad y, sobre todo, la calidad de los datos que se esconden detrás del Big Data. Lo que hay detrás de la máscara digital es la persona real, que lo es cuando está desconectada… Gracias a la tecnología podemos ser dos entidades: una etérea basada en bits y electrones y otra real, basada en carne, ladrillos y cemento, que frecuentemente se solapan.

Pero no siempre. Sin cara y sin identidad, navegamos en la red como una versión de nosotros, sin ser nosotros.

Kulturbrille

Las “lentes culturales” fueron conceptualizadas y descritas por Franz Boas para definir los filtros con los que nos contemplamos a nosotros mismos. Por un lado son útiles para dar sentido a nuestras observaciones y por otro son una forma de ceguera, peligrosa en tanto no somos conscientes de ella.

Culturalmente es posible que la cocina sea un área “exclusiva” para las mujeres −ni siquiera es necesario poner ejemplos, pero por poner uno, citemos a Japón−. Con

esto no vamos a justificar, ni siquiera a considerar, esta pintoresca costumbre. Pero si somos parte del equipo de ventas de una multinacional que comercializa electrodomésticos, debemos tener en cuenta este pequeño dato cultural.

La observación individual es la parte fundamental de los "datos relevantes" o pequeños datos. De otro modo ¿cómo explicar que, pese a que 3.424.971.237 humanos aproximadamente[3] estamos conectados diariamente a la red y utilizamos el buscador de Alphabet Inc. (conocido como Google), esta empresa sepa a fin de cuentas tan poco sobre nosotros[4]?

En muchas ocasiones lo realmente revelador no es la solución que se origina en los grandes datos, o el descubrimiento proporcionado por el aprendizaje automático. Ni siquiera la que proviene de un estudio internacional, o regional o local, ni tampoco la que se ha obtenido mediante el concurso de una gran cantidad de equipos de consultores. En ocasiones puede ser simplemente la idiosincrasia de un lugar o ponerse en el caso o en el lugar del objeto de nuestro estudio.

Además el suministro continuo de información en tiempo real confunde y puede ser desproporcionadamente alarmista. Es como ver un telediario o estar expuesto en Internet a un flujo continuo de malas noticias, sin ninguna perspectiva. Se trata de algo análogo a seguir los mercados financieros en tiempo real. Se pierde la perspectiva y la cantidad de información deja de tener verdadero valor informativo.

Sabemos que, en la actualidad, estamos expuestos y además parecemos desear cantidades masivas de datos que no estamos preparados para procesar. Como una especie de "comida basura", la información nos sacia momentáneamente pero acto seguido volvemos a estar hambrientos y regresamos a por más.

Agua por todas partes, pero ni una gota para beber

Los datos masivos son el presente. Los algoritmos, el aprendizaje automático y otras técnicas de inteligencia artificial forman parte del momento actual. Y además esto es algo bueno: el problema no es la tecnología en sí, ni los datos, por masivos que sean. El problema es el desequilibrio y la asimetría.

Nuestra percepción del mundo es casi siempre local y se centra en nuestro entorno, en nuestros vecinos, nuestras tradiciones o creencias. Cada cultura tiene su propia idiosincrasia, sus temas de conversación, sus narrativas y costumbres… Pero en ciertos aspectos el mundo ya no es local −los souvenirs han dejado prácticamente de existir porque es raro el objeto que no se puede conseguir desde cualquier sitio− aunque quedan sectores que todavía pueden sorprendernos por su carácter local… Pero hoy en día nos comparamos y nos guiamos no por nuestros amigos o vecinos, sino por millones de personas de todo el planeta.

Existen también diferencias individuales en nuestra percepción del mundo. Todos los seres humanos tendemos a ver la realidad de forma diferente, aun cuando

[3] *Internet live stats*. www.internetlivestats.com/internet-users/ Consultado el 26/01/2017.
[4] *Small Data. The Tiny Clues That Uncover Huge Trends*. Martin Lindstrom. St. Martin's Press. 2016, p. 21.

todavía sea más semejante de lo que podemos imaginar. Pero los datos masivos no nos tratan como individuos sino en el marco de procesos de segmentación y grupos demográficos.

El proceso de grandes datos es una solución valiosa pero incompleta. El problema es que puede ocultar percepciones importantes e ideas de alta calidad. Internet sigue proporcionando datos acerca de quienes somos y cómo nos comportamos, pero en su entorno, esto es; en un ambiente altamente idealizado y protegido.

Los datos masivos pueden proporcionar información exacta conectando y correlacionando millones de datos pero tienen problemas cuando los individuos humanos se comportan como… humanos. Parece lógico pensar que, al tiempo que los grandes datos modificarán nuestras vidas, las personas evolucionarán simultáneamente para adaptarse a estos cambios que la tecnología nos proporciona. Al final los datos masivos, los pequeños datos y los datos de calidad necesitarán encontrar un equilibrio.

Es en los pequeños detalles donde se encuentra la evidencia fundamental de quienes somos y qué deseamos. Los datos masivos son masivos, sin estructura intrínseca y totalmente heterogéneos… y por eso son incomprensibles a escala humana. Si no podemos hacer que la tecnología nos ayude significativamente, acabaremos ahogándonos en un mar de datos sobre nosotros mismos.

Y no sólo es el dato sino la carencia del mismo. Sabemos que para determinados análisis es más importante advertir la ausencia de algún detalle que la presencia del mismo. Tanto más cuando estamos tan inundados de datos que no podemos separar el silencio del ruido, ni de la información.

Con los datos masivos estamos creando pirámides. Pero solo en el ápice de la pirámide se puede tener una visión de lo que está debajo y ahí es donde trabajan los grandes datos. Pero estos datos son útiles sólo en función de las interpretaciones que hagamos de ellos y de las conclusiones que extraigamos de los mismos.

El problema fundamental de los datos masivos es que, cuanto más abundantes sean, mayor es la probabilidad de que los análisis de cualquier tipo, encuentren fundamento en ellos para soluciones que no existen en realidad. Fundamento incluso para resolver problemas que ni siquiera se han planteado. Cuantos más datos, mayor es la posibilidad de encontrar correlaciones aleatorias y sin sentido entre ellos.

Puede ser un problema que las grandes expectativas confiadas al procesamiento masivo de grandes datos acaben en decepción. Esto sería algo que acabaría con el valor real que puede aportarnos el análisis de la información masiva de la era digital. Las grandes cantidades de datos no garantizan directamente grandes cantidades de información accesible[5]. La entropía limita la cantidad de información disponible en un sistema a partir de un punto desde el que se hace extremadamente difícil extraer una señal informativamente importante del ruido de fondo. La información debe encontrarse en un tipo especial de "zona de habitabilidad" en la que no sea ni demasiado escasa ni demasiado abundante[6].

[5] *La señal y el ruido. Cómo navegar por la maraña de datos que nos inunda, localizar los que son relevantes y utilizarlos para elaborar predicciones infalibles*. Nate Silver. Eds. Península. 2014.
[6] *Goldilocks communication: Just the right amount of information*. Cam Barber. The Vivid Method. 18/05/2011.

Sabemos que cuando analizamos grandes cantidades de datos y buscamos correlaciones significativas entre un número pequeño de variables, los datos son sensibles a muchas otras. Cuantas más variables controlas, la muestra se hace demasiado pequeña como para extraer conclusiones significativas.

Tener grandes cantidades de datos no supone automáticamente que los mismos tengan valor en sí, ni que de ellos se pueda extraer este valor o que de ellos se seguirán necesariamente grandes ideas. Puedes tener un gran camión lleno de arena pero los castillos que pueden hacerse son innumerables. Los datos masivos, sin personal experto que extraiga conclusiones y sentido de ellos, son nada más que eso: contenedores llenos de arena.

Conclusión

Para una inteligencia artificial alimentada por datos masivos es relativamente fácil describir a una persona. Es mucho más difícil para una persona mirar en un espejo y describirse a sí misma.

Pese al gran ruido mediático −en cierto modo justificado− en torno al aprendizaje profundo, los algoritmos y el adiestramiento automático, hasta el momento nada puede compararse a la habilidad del ser humano para establecer hipótesis acerca de conjuntos de datos y establecer relaciones entre ellos mediante el "sentido común". Correlación no implica causación; más que datos por datos, al final es interesante que los datos sean verificados por seres humanos

En realidad no debe exagerarse con el término "small data". Los pequeños datos, los datos significativos y de calidad siempre han existido. Siempre han coexistido con la información y con el ruido. Es lo que anteriormente se conocía simplemente como *datos*.

Sin embargo, el valor del concepto de *pequeños datos* es que nos permite recordar la importancia de los datos clásicos y, combinados con la tecnología y con la inteligencia del analista, su relevancia para encontrar nuevas soluciones y direcciones correctas y significativas a la hora de realizar estudios y considerar eventualidades.

Pero no es suficiente; si hacemos demasiado hincapié en los datos pequeños o los datos de calidad podemos incurrir en el error de Kant: alcanzar una conclusión demasiado general a partir de una experiencia insuficiente. Para el filósofo germano no se podía crear nueva materia, ni la materia existente podía dejar de ser, y esto era tan obvio que constituyó su ejemplo de razonamiento a priori. Pero hoy sabemos que la materia y la energía están relacionadas y la una puede transformarse en la otra. Después de Kant, vino Einstein.

Un pequeño dato o una información de calidad no suele ser suficiente para crear una hipótesis válida o una base estratégica viable. Es más bien su visión global y su presentación lo que determinarán su valor.

Al final, en los análisis de seguridad y en los estudios prospectivos debemos recordar constantemente las palabras de Robert Ingersoll cuando afirmaba que: *la razón, la observación y la experiencia son la santísima trinidad de la ciencia*. Un pequeño dato; muy válido en el análisis de inteligencia.

Lecturas adicionales

Small Data. The Tiny Clues That Uncover Huge Trends. Martin Lindstrom. St. Martin's Press. 2016.

La locura del solucionismo tecnológico. Evgeny Morozov. Katz. 2015.

No Comercial (Non commercial): La explotación de la obra queda limitada a usos no comerciales.

Compartir Igual (Share alike): La explotación autorizada incluye la creación de obras derivadas siempre que mantengan la misma licencia al ser divulgadas.