Un universo de datos: situación y perspectivas

A pesar de las innovaciones que han tenido lugar en los últimos 20 años, existen una serie de hándicaps que deben ser superados para llegar a alcanzar al menos las 3 Vs básicas del Big-Data: Volumen – Velocidad – Variedad. Es precisamente en la última de estas «V» donde se deben dedicar esfuerzos en los próximos años.

Roberto Knop @rkm4457 | *Director asociado del área Analytics de Afi*
Francisco Jesús Rodriguez | *Consultor del área de Analytics de Afi*

Con total seguridad habrás escuchado en diversos medios máximas como Data is the new oil o que vivimos en la «Sociedad del Dato» o que las empresas deben ser Data Driven.

Sin duda, si algo ha marcado el inicio de este nuevo milenio, ha sido hechos claves como:

- la **aparición de numerosas fuentes de información** gracias al desarrollo de Internet,
- el **incremento de la capacidad de cómputo y de almacenaje de los procesadores** y, más recientemente,
- las facilidades que nos ofrece **la computación en la nube**.

Todos estos elementos ofrecen a amplios espectros de empresas la posibilidad de almacenar y gestionar grandes cantidades de datos, lo cual hace apenas 20 años estaba prácticamente relegado a grandes corporaciones que eran las que podían hacer grandes inversiones en equipos con suficiente potencia.

Vivimos en una nueva era, no hay duda, pero como sucede en cada era de la humanidad, aparecen soluciones a los problemas de la anterior, pero también surgen otros nuevos que seguramente se resolverán en la próxima. Si nos centramos en el momento actual, **gran parte de las tecnologías del tipo open-source se están integrado como parte del día a día de las grandes corporaciones, de las empresas de tamaño medio y forman parte del ADN de las numerosas start-ups** que surgen en la actualidad. Y es que el procesamiento de grandes

volúmenes de información está al alcance de cualquier institución con el simple gesto de solicitar un procesador potente a través de alguno de los proveedores de software de la nube conocidos. Hoy en día, a golpe de clic cualquier técnico o incluso cualquier persona sin tener formación informática, puede «encender» una máquina con la potencia que desee, utilizarla durante un tiempo para un determinado cálculo, apagarla y ceder dicho uso a otro usuario que lo necesite.

La extensión del uso de las tecnologías conocidas como Big-Data está en proceso, pero existen retos que deben superarse. Una visión simplista del Big-Data, que es la metodología de tratamiento de datos que cumplen las 3 condiciones básicas conocidas como las 3 «V», actualmente numerosas compañías pueden procesar y modelizar con datos muy pesados, como por ejemplo los procedentes de imágenes de satélite, de cámara de vídeo e incluso aplicar estos avances a través de la Inteligencia Artificial para la conducción autónoma o la gestión del tráfico. Sin embargo, **el problema se plantea cuando se quiere dotar de variedad a la información**. Aunque existirían problemas éticos y legales como los que se plantean en el Reglamente General de Protección de Datos, numerosas situaciones en las que cruzar una diversidad de fuentes de información son totalmente lícitas.

Normalización

Si nos fijamos en los numerosos datos abiertos que se ofrecen a nivel de estadística pública que comprenden mapas, imágenes, tablas, documentos, todos ofrecidos de modo gratuito y cumpliendo las especificaciones éticas y legales pertinentes, ¿cómo de fácil o de difícil sería asociar, por ejemplo, a un automóvil autónomo del futuro información relevante de una determinada región donde el conductor desee ir? Un problema como este es el que afrontan diariamente las empresas, y si bien **es cierto que existen enormes cantidades de datos disponibles y procesables, el problema está en cómo unirlos para que sean diversos**.

Un ejemplo sencillo es intentar relacionar información a un nivel tan agregado como es el de los municipios españoles. Este ejercicio debería ser sencillo a día de hoy, porque se supone que se está ante un nivel de dato en teoría tratable. Sin embargo, choca con algunos problemas como el de la normalización. Así, por ejemplo, aunque existe un código municipal que suele utilizar el INE, distintas administraciones tanto públicas como privadas en general no lo utilizan y publican datos usando sólo el nombre del municipio. Esto genera problemas al querer unificar información, ya que el nombre de un municipio como por ejemplo «El Álamo», es habitual encontrarlo con distintas denominaciones como «Álamo (El)» o «Álamo, El». Ello, sin contar las particularidades idiomáticas de las distintas regiones, donde en ocasiones se escribe en las dos lenguas y otras veces en una de ellas. Por tanto, de un hecho formado sólo por unas 8100 entidades municipales, se requiere un trabajo intenso de unificación y de mantenimiento para la actualización necesaria conforme se producen cambios a lo largo del tiempo por apariciones, desapariciones y fusiones de distintos municipios.

Indicadores de alta frecuencia

Las empresas y la sociedad en general quieren anticipar problemas, comportamientos de los distintos agentes de la sociedad, la evolución de la economía. Por ello, **la construcción de indicadores de alta frecuencia capaces de anticipar problemas con alertas diarias, horarias o personalizadas es sin duda otro de los retos que requieren la integración de distintas fuentes**, tales como las publicaciones de determinados usuarios en Twitter, la aglutinación de las noticias que aparecen en distintos medios de prensa, o la renovación constante de determinadas bases de datos de las instituciones públicas. Todo resulta procesable, todo resultar ser un gran volumen de información, pero falta el eslabón de la unificación.

Finalmente, **nuestros sistemas de Machine Learning e Inteligencia Artificial podrían ser realmente potentes si fuesen capaces de integrar toda esta información**. Posibilitar el uso de aquello que se publica libremente, pero con cierto grado de unificación y centralización es necesario si se desea que esas fuentes de información que han eclosionado en la última década y que son servidas y utilizables por el público en general, resulten realmente útiles y potentes para la sociedad en su conjunto ::