Reflexión sobre Datos masivos y su impacto

Bit&Byte conversó con el prestigioso Doctor José Ángel Olivas de la Universidad Castilla La Mancha de España, acerca de cómo afectan los datos masivos en la sociedad y de sus consecuencias políticas y económicas.

A lo largo de la conversación, también remarcó que en el mercado informático hay una gran demanda de ingenieros de datos y que la formación de los profesionales de informática debe atender esta necesidad

¿Cómo podría explicar qué son los datos masivos?

Hay que diferenciar claramente entre lo que son datos, lo que es información y lo que es conocimiento. Para mí los datos son los elementos en crudo y los registros, lo que proveen los sensores. En cambio, el conocimiento es una abstracción de mucho más alto nivel, y en el medio está la información que es lo que muchas veces se entiende por visualización de los datos.

Los datos masivos vienen de muchas fuentes, hoy en día podríamos distinguir los datos estructurados de los no estructurados. Los estructurados son aquellos de las bases de datos relacionales de toda la vida, que están perfectamente divididos entre campos y registros; y los no estructurados son de fuentes que no tienen esa organización, por ejemplo, el texto plano. Suelen venir de las redes sociales, comentarios, ficheros de texto, de imagen de videos en streaming.
El dato masivo hoy en día es "internet of things" (IoT), que son los datos que vienen de sensores, o las "smartcity" basadas en la "internet of things", son datos que vienen, por ejemplo, de los sensores de nuestros móviles.

¿Por qué creé que se habla de "la revolución del Big Data"?

Yo creo que se habla por error, el Big data no es revolucionario, sino que es una continuidad. Es decir, el término Big data es relativo a nuestras capacidades. Yo trabajaba en la década del 90' con un ordenador con 20 megas de disco duro y 4 megas de memoria RAM y ya analizaba datos y utilizaba casi los mismos algoritmos que se utilizan ahora.

Lo que ocurre es una cuestión de tamaño nada más. Es decir, hoy en día si tengo un equipo con 4 teras de memoria puedo procesar una cantidad ingente de datos.
Sin embargo, pienso que no hay una separación clara entre lo que es el análisis de datos de toda la vida -previo a la computación incluso- de lo que es el big data.
Es una cuestión de tamaño que es relativo, si lo tengo que hacer en mi portátil no puedo, si lo tengo que hacer en una estación con 4 teras de RAM, pues tengo más capacidad.

¿Piensa que esta revolución del Big Data tiene consecuencias sociales y económicas?

Absolutamente, hoy en día hay un montón de información y de datos, entonces dependemos mucho de su análisis. Por ejemplo, las empresas dependen de los comentarios de los clientes, un mal comentario puede hacer que caigan las ventas de un producto.

Hay casos en los que una cadena de opinión hace fracasar a un producto, entonces como hay tantos datos, la sociedad depende de ellos. Por ejemplo, un hotel o un restaurante se pueden hundir por algunos comentarios negativos en Tripadvisor. Hay que estar atentos y tener cuidado con los datos, los tweets, todo esto que llamamos revolución digital. Las revoluciones ya no se hacen en la calle, se hacen en redes sociales. Los datos son imprescindibles, no podemos vivir ignorando la información digital

¿Qué significan los tres factores "volumen", "variedad" y "velocidad" a la hora de clasificar Big Data?

Bueno también se habla de 4 o incluso de 8. Pero refiriéndome a estos 3 en particular, el volumen significa que el tamaño es inmenso; la variedad

que son de muy distinto formato, el concepto de data lake (lago de datos) que se habla en las empresas, pues hay datos duplicados, datos de video, estructurados, no estructurados, etc. Y la velocidad es tanto la que generan como la que transmiten o puede transmitirse, un teléfono puede estar mandando la medida de 20 sensores por segundo, por ejemplo.

Sin embargo, hay más factores, la veracidad, por ejemplo, cuan fiables son los datos. La volatilidad, hay datos que sólo duran un momento.

¿Por qué se relaciona la Inteligencia Artificial y los Sistemas Inteligentes en general con Big Data?

Porque la única forma de aprovechar el conocimiento que nos pueden suministrar los datos es mediante el uso de técnicas sofisticadas y no de informática convencional, que son las técnicas que provee la inteligencia artificial, y en particular una disciplina de la inteligencia artificial que es el Machine Learning (aprendizaje automático).

Hay dos grandes fuentes de herramientas o técnicas para el análisis de datos con la intención de extraer conocimiento. Por un lado, aquellas que provienen del mundo más clásico de la probabilidad y de la estadística: extrapolación de series temporales, técnicas de regresión, técnicas de modelos gráficos probabilistas. Por otra parte, las que son propias del mundo de la inteligencia artificial como las que vienen del paradigma conexionista, es decir las redes neuronales, deep learning que está tan de moda. También están las que derivan del paradigma evolutivo, los algoritmos genéticos, la lógica borrosa, etc.

¿Concuerda con que la economía mundial experimenta grandes cambios a partir de la revolución de los datos masivos?

Totalmente, es dependiente de ellos. No sólo la economía, la política también. Hoy en día el criterio por el que la mayoría de la gente vota a un partido u otro son comentarios en las redes sociales, las influencias, con lo cual la revolución de los datos masivos es la que mueve el mundo.

Una empresa puede quebrar por una cadena de comentarios malos sobre un producto, conozco casos reales sobre esto, entonces el mundo depende de los datos hoy en día. Tanto para tomar decisiones, que es lo que suelen hacer las empresas, como para manejar situaciones.

Repito el concepto de smartcity: si tengo muchos datos de los sensores de los coches, es posible que pueda optimizar los semáforos y mi ciudad sea más ecológica. Incluso en la dimensión humana: si analizo los mensajes, los tweets, una red social como Instagram, puedo estudiar regularidades en el comportamiento humano que me permitan decidir si mi partido político está trabajando bien o mal, si la política de mi empresa no es la adecuada, etc.

Por lo tanto,creo que eso sí es una revolución, el big data es una evolución. Desde que se empezaron a tratar los datos, que no es cosa ni del siglo pasado, ya en el siglo XVIII Y XIX había datos y se trataban.

El big data es una evolución: la sociedad digital. Hoy en día se generan tantos datos o se pueden generar, que está habiendo un problema porque no se están tratando y no hay mucha posibilidad de tratarlos. Vamos detrás, es decir, se generan más datos de la capacidad que tenemos de procesarlos o analizarlos, por eso tenemos que trabajar mucho en estos ámbitos para mejorar las técnicas.

¿Qué temas son importantes en la formación de profesionales en Informática para manejar problemas de Big Data?

Esto es muy importante. Hay dos figuras fundamentales, tanto en el mundo de la empresa como en el de la investigación: el data scientist y el data engineers.

El data engineers es alguien que sabe almacenar, generar una arquitectura, transmitir, guardar grandes volúmenes de datos, pero no tiene por qué saber dotarlos de inteligencia, de análisis. Un ingeniero de datos es simplemente alguien que se dedica a manejar los datos como si fuera una mercancía para que el data scientist sea capaz de analizarlos y sacarles partido, es decir, extraer conocimiento de ellos.

Esta es la figura que me interesa para la formación particularmente porque es muy complicada y hay muy pocos. Justamente porque como es un tema complejo tiene que ser una persona muy bien formada. Tiene que saber mucho de inteligencia artificial, sus fundamentos, conocer los algoritmos, cuáles son las limitaciones, para qué sirve, en qué casos es mejor utilizarlo.

Eso es una parte: el machine learning. Pero también tiene que tener conocimientos, por ejemplo, de sociología, de antropología social, de psicología, de por qué en una red social los comentarios se dirigen de determinada manera o qué intención pueden tener.

Es un saber muy renacentista, de fundamentos matemáticos, entonces, ocurre que cada vez escasea más la figura de un buen data scientist y hoy en día mucha gente sabe las herramientas, los algoritmos, conoce las librerías que se suelen utilizar en big data pero no con la suficiente profundidad como para seleccionar el más adecuado y afrontar un problema concreto desde los datos hasta el conocimiento.

No es sólo aplicar el algoritmo, hay un pre proceso, una selección de qué datos voy a utilizar, quitarles el ruido, lo que se llama data cleaning. Es todo un proceso desde que llego de los datos al conocimiento: tengo que tener muy claro como científico de datos, el conocimiento que puedo esperar. Hay un error muy frecuente y es que mucha gente que se dedica a ser científico de datos aplica a un conjunto de datos algoritmos a ciegas para ver qué sale. Es un error fatal, en cualquier proceso de análisis de datos hay que tener muy en claro qué tengo, lo que puedo necesitar y puedo encontrar en otras bases de datos, y dónde quiero llegar. Esto puede ser predicción, pronóstico, prospección, estimación, prescripción.

No es lo mismo hacer un sistema de pronóstico que uno de predicción porque ésta tiene que ver con la extrapolación de un comportamiento y el pronóstico tiene que ver con anticipar un hecho puntual en base a pocas alternativas.

Por lo tanto, un data scientist no sólo tiene que conocer los algoritmos sino tiene que saber muy bien las posibles salidas y cómo manejarlas porque eso determinará qué algoritmos o herramientas utilizar y cómo preprocesar eso datos.

Por lo tanto, la formación de un data scientist es compleja, es una persona que tiene que estar muy formada. Por desgracia es una figura escasa en el mercado, que se paga muy bien y que en la mayoría de las demandas de puestos ofertados están relacionados al análisis de datos. Debemos fomentar en las escuelas de informática la formación seria y compleja para formar profesionales que puedan ser en un futuro un buen científico de datos •