BIG DATA: ¿LA MISMA CERVEZA PERO CON OTRO ENVASE?

JUAN VELÁSQUEZ

Profesor Asociado del Departamento de Ingeniería Industrial (DII), de la U. de Chile. Dr. en Ingeniería de la Información de la Universidad de Tokio, Japón.

jvelasqu@dii.uchile.cl

Hace unos 20 años, cuando comenzábamos a trabajar con algoritmos de procesamiento masivo de datos, específicamente utilizando Redes Neuronales Artificiales y Algoritmos Genéticos, nos maravillábamos con la posibilidad de analizar estos grandes volúmenes de información en sólo unas cuantas horas usando súper computadores. La idea central era encontrar patrones en los datos que nos permitieran crear un modelo predictivo del fenómeno en estudio, algo así como una bola de cristal que "adivinaba el futuro".

Datos, información y conocimiento. ¿Cuáles son las distinciones fundamentales? Los datos son sólo un registro de un evento, por ejemplo la temperatura de un objeto. La información nos permite tomar decisiones. Por ejemplo, si un indicador de gestión nos dice que estamos mal en las ventas de un producto, hay que tomar acciones al respecto. El conocimiento es algo un poco más elaborado, y del punto de vista tecnológico tiene que ver con "patrones y reglas de uso". Por ejemplo, si el paciente manifiesta X,Y y Z síntomas, entonces debemos aplicar el procedimiento P.

Pongamos las cosas en un contexto simple. Con un bit se puede almacenar un estado de "verdadero/falso". Un byte (8 bits) almacena un carácter. 1024 bytes corresponden a 1Kilobyte, lo cual permite almacenar una oración. 1 Mega byte, o 1024 Kilobytes, sirven para almacenar una novela corta, y en 1 Gigaybte, o 1024 Megabytes, podemos guardar una película. En un disco de 1TeraByte se pueden almacenar los documentos/libros de una biblioteca grande, como la del Congreso Nacional. Las próximas escalas ya comienzan a ser cifras tan enormes que con la tecnología actual se necesitarían edificios llenos de discos duros de 1Terabyte para poder almacenar todos esos datos.

El concepto Big Data nos propone trabajar con Terabytes de datos, y en algunos casos con PetaBytes o más (es decir, millones de Gigabytes), en una enorme gama de disciplinas con un sueño común: encontrar patrones que permitan descubrir un nuevo conocimiento desde las cordilleras de datos. Es importante notar que en este caso los datos no están estructurados utilizando tan solo las bases de datos relacionales tradicionales, sino que se va mucho más allá. Son las bases de datos no estructuradas, donde todos los tipos de datos de la historia están presentes, las que más concitan la atención de quienes desarrollamos algoritmos y herramientas Big Data.

PROBLEMAS Y DESAFÍOS

Son los de antaño, los de siempre, es decir, cómo capturar, almacenar, buscar, comparar, analizar y visualizar grandes volúmenes de datos. Pero con la sutileza de que ahora estamos muy lejos de contar con una solución tecnológica para la creación de discos que sean capaces de almacenar millones de TeraBytes, y con un acceso lo suficientemente rápido como para ser transmitidos casi sin retardos por una red de alta velocidad. Hasta el momento la solución más práctica ha consistido en colocar sendos arreglos de discos duros para almacenar los datos en forma distribuida, para luego hacer uso de estos a través de computación paralela. Pero todo tiene su costo, y el principal es la energía utilizada para mantener esta infraestructura.

Revisitemos nuevamente un viejo problema: ¿cómo obtener información valiosa a partir de los datos que nos permita tomar la mejor decisión táctica/estratégica para una situación en particular? Más aún, ¿qué nuevo conocimiento puedo extraer a partir de los datos que le permitan a mi institución lograr una ventaja competitiva frente a su competencia? Una alternativa muy recurrida cuando por problemas tecnológicos no se puede procesar toda la base de datos, es tomar una muestra representativa de ésta y alimentar algún algoritmo extractor de patrones que nos permita crear un modelo

→

predictivo. Lo anterior, obviamente corre el riesgo de dejar pasar datos que pueden ser claves a la hora de descubrir un nuevo conocimiento, pero al menos nos da una buena aproximación del fenómeno en estudio. La otra alternativa es derechamente preprocesar toda la base de datos, aplicar algoritmos de extracción de patrones y armarse de paciencia, suponiendo que tenemos un computador de amplias capacidades. Los resultados pueden ser realmente sorprendentes, sobre todo cuando se cruzan datos provenientes de fuentes diversas y que complementan el análisis.

IMPLICANCIAS DE BIG DATA

Privacidad: imaginemos una situación donde la aplicación de Big Data nos va proponiendo compras en distintas etapas de nuestras vidas y luego al detectar que fuimos a una clínica, nos propone un seguro de vida. ¿Qué pasa con la privacidad de nuestros datos personales? [2].

Backtracking de decisiones: ¿cómo se tomarían las decisiones si pudiésemos analizar todas las alternativas posibles, con solo cambiar las entradas del modelo que se originó a partir de los datos? [4].

Personalización de la oferta: un viejo sueño del *marketing one to one* se puede hacer real: conocer los gustos, preferencias, necesidades etc. de los clientes en forma individual, y orientar el mensaje ya no a la masa, sino a cada persona, logrando una efectividad sin precedentes.

Nuevos negocios basados en datos: ya hay empresas que venden el servicio de almacenamiento de datos. Otras que se dedican a su procesamiento. ¿Qué tal algunas cuyo rubro sea la limpieza de datos? U otras que solo se dediquen a la generación de reportes, etc. [3].

ALGUNAS APLICACIONES

Health Informatics: con el avance de la ciencia médica, la esperanza de vida ha aumentado considerablemente. La medicina preventiva está siendo cada vez más recurrida para asegurar la detección temprana de enfermedades. Para lograrlo, se debe mantener un registro histórico de todos los exámenes que se le ha realizado a un paciente. Nuevamente la cantidad de datos es enorme, lo que se traduce en un tremendo desafío.

Brain Informatics and neuromarketing: ¿cuál es la estructura y contenido correcto para que un sitio web atraiga y retenga a sus visitantes? Utilizando técnicas de la neurociencia, como los electroencefalogramas y dispositivos *eye tracking,* podemos conocer la respuesta de los usuarios web a estímulos visuales, tales como imágenes, colores, vídeo etc. presentes en una página web. Con los sistemas de *eye tracking* se realiza un seguimiento del movimiento ocular y análisis de dilatación pupilar, la cual está directamente relacionada con la aceptación/rechazo del estímulo por parte del usuario. Adicionalmente, los datos generados en el electroencefalograma nos permiten clasificar la respuesta emocional del usuario frente al estímulo visual [3]. Una sesión de 30 minutos utilizando estas técnicas puede generar varios terabytes de datos crudos (más detalles ver www.akoriproject.com). Estas técnicas también son utilizadas para conocer las preferencias de los consumidores en el retail, analizando cuáles son los niveles de emoción, atención y memoria, configurando el área del marketing conocida como *neuromarketing*.

Web opinión mining: a partir de la extracción de información desde las redes sociales se pueden

IMAGEN 1.
EQUIPO DE TRABAJO DEL PROFESOR JUAN VELÁSQUEZ.

IMAGEN 2.
DE IZQUIERDA A DERECHA: JUAN VELÁSQUEZ, YERKO COVACEVICH Y FRANCISCO MOLINA.

lograr sorprendentes análisis respecto de las percepciones, sentimientos, opiniones que tienen los internautas sobre un producto o servicio en tiempo real. El problema es que solo en Twitter se están generando más de 40.000 tweets por segundo, cifra que irá en aumento en los próximos años [1].

REFLEXIÓN FINAL

Hay algo muy interesante en casi todos los problemas relacionados a Big Data: los algoritmos y técnicas que se desarrollan no son para un tipo de dato en específico. Dicho de otra manera, si desarrollamos un algoritmo de análisis de serie de tiempo para datos generados en un radiotelescopio, y luego lo aplicamos a los generados por un electroencefalograma, con algunos ajustes claro está, podremos extraer patrones a partir de las ondas cerebrales de un paciente y quién sabe, detectar en forma temprana una anomalía. Como en todo nuevo concepto, hay mucho de mito y poco de realidad. Aparecen miles de expertos, gurúes del área, pero que están igual que todos nosotros: somos testigos del nacimiento de algo grande que comienza recién a dar sus primeros pasos en ciencia, en tecnología y en los negocios.

Big data: ¿la misma cerveza pero con otro envase? Aún no lo tenemos claro, pero conviene que esta vez se la tome con calma y muy helada, sino puede causarle una big indigestion. ■

REFERENCIAS

[1] "Detecting Trends on the Web: A Multidisciplinary Approach", Dueñas Fernández R., Velásquez, Juan D. and L'Huillier, Gastón. Information Fusion, 20:129-135, 2014.

[2] "Web Mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in Web-based environments", Velásquez, Juan D., Expert Systems with Applications, 40 (1): 5228-5239, 2013.

[3] "A neurology-inspired model of web usage", Pablo E. Román and Juan D. Velásquez, Neurocomputing, 131: 300-311 ,2014.

[4] "Are you ready for the era of 'big data'", B. Brown, M. Chui J. Manyika, McKinsey Quarterly, 4:24-35, 2011.

[5] "Big data: the management revolution", A. McAfee and E. Brynjolfsson, Harvard Business Review, 90(10):60-68, 2012.

LA NUEVA ERA DE DATOS EN ASTRONOMÍA

FAVIOLA MOLINA

Investigadora postdoctoral en el Departamento de Ciencias de la Computación, Universidad de Chile, donde trabaja con el profesor Alexandre Bergel. Doctora en Ciencia de la Universität Heidelberg y fellow del Instituto Max Planck para Astronomía en Heidelberg, Alemania (2013). Astrónoma de soporte en el Observatorio Europeo Austral (2008-2009). Magíster en Astronomía y Astrofísica de la Pontificia Universidad Católica de Chile (2008). Licenciada en Física de la Universidad de Los Andes, Venezuela (2004). Áreas de interés: Astro-computación, análisis estadístico del medio interestelar y formación de estrellas, poblaciones estelares y recientemente formación de discos planetarios y de transición.

fmolina@dcc.uchile.cl

El desarrollo de cualquier disciplina científica involucra el manejo de datos. En el caso particular de la Astronomía, el incremento de la cantidad y tamaño de los archivos de datos ha ido creciendo con el paso de los años, considerando así a ésta como una ciencia de datos intensivos [Hassan and Fluke, 2011].

Hasta mediados del siglo XX, en específico en la Astronomía óptica, los detectores eran placas fotográficas. La exposición de las mismas podía tomar horas de acuerdo a la intensidad del objeto que se quería observar. Más tarde, se dio paso a los fotómetros fotoeléctricos que ofrecían mayor sensibilidad, precisión, linealidad y un mayor rango dinámico para el análisis que las placas fotográficas[1].

1 http://star-www.rl.ac.uk/docs/sc5.htx/node7.html