OBSERVATORIO

BIG DATA Y ANALÍTICA WEB. ESTUDIAR LAS CORRIENTES Y PESCAR EN UN OCÉANO DE DATOS

Jorge Serrano-Cobos

Jorge Serrano-Cobos, licenciado en documentación por la *Universidad de Granada*, es profesor asociado de la *Universidad Politécnica de Valencia* donde imparte materias como SEO, analítica web y social media analytics. Ha trabajado en las empresas *MASmedios*, *Google*, *Planeta DeAgostini* y *Serikat*, y ha realizado docenas de proyectos de consultoría y asesoría en marketing académico y científico, estudio de mercados digitales, analítica web, análisis estratégico cualitativo y cuantitativo digital (cibermetría) para e-commerce, e-marketing, arquitectura de la información, diseño de interacción, SEO y search analytics, con vocación de internacionalización.
http://orcid.org/0000-0002-4394-4883

MASmedios
Garcilaso, 15. 46003 Valencia, España
jorgeserrano@gmail.com

Resumen

Se realiza un recorrido por las características, posibilidades, disciplinas científicas, técnicas y tecnologías que se recogen dentro del paraguas interdisciplinar del *big data* y la analítica web desde el punto de vista de su aplicación a la praxis. Se realiza una reflexión en torno a los retos, riesgos y problemas que las herramientas y los datos no resuelven por sí solos, así como sobre los contextos de uso de estas técnicas de tratamiento de datos para la toma de decisiones.

Palabras clave

Big data, Analítica web, Analítica predictiva, Analista, *Data mining*, *Text mining*, Análisis, Cibermetría, Marketing, Usuarios, Estadística, Aprendizaje, Informe de situación.

Título: Big data and web analytics. Studying the currents and fishing in an ocean of data

Resumen

A tour is provided of the features, possibilities, scientific, technical and technologies that are collected under the interdisciplinary umbrella of big data and web analytics from the point of view of its application in practice. A reflection is offered about the challenges, risks and problems that the tools and data cannot resolve on their own, as well as the contexts in which these data processing techniques are useful for decision making.

Keywords

Big data, Web analytics, Predictive analytics, Analyst, Analysis, Data mining, Text mining, Cybermetrics, Marketing, Users, Statistics, Machine learning, Situation report.

Serrano-Cobos, Jorge (2014). "*Big data* y analítica web. Estudiar las corrientes y pescar en un océano de datos". *El profesional de la información*, v. 23, n. 6, noviembre-diciembre, pp. 561-565.
http://dx.doi.org/10.3145/epi.2014.nov.01

Introducción

Recuerdo que allá por los 80, la primera vez que vi una computadora con tarjetas perforadas (y no era en un museo) que iba a ser sustituida por un disco (de considerable tamaño, eso sí), me dije "lo que nos queda para *HAL 9000*". En aquella época estaban en boca de todos los amantes de la tecnología las últimas novedades de la inteligencia artificial, y ya nos hacíamos la boca agua con la llamada "quinta generación de computadoras" con capacidad de resolución de problemas complejos imitando la forma de pensar de los seres humanos.

Años más tarde, y antes de otras burbujas, ya tuvimos un "bluff" de esa rama de la inteligencia artificial en este sentido, sustituida por la computación basada en insectos sociales, a su vez sustituida (en cierto modo) por la computación social, esta vez usando personas en lugar de insectos ("*the wisdom of crowds*"). Probablemente no era su momento.

Pero por otro lado, quién le hubiera dicho a Vannevar Bush la que se iba a liar después de la burbuja de internet de 2000 y su *rebranding* a través de la marca "Web 2.0". Viendo cómo han cambiado las cosas, y haciendo un poco de memoria histórica, uno puede tener la tentación de predecir

Artículo recibido el 19-10-2014

hacia dónde irán. Y eso es, hoy día, lo que todo el mundo intenta hacer: lo llaman "*predictive analytics*".

Variedad, volumen, velocidad, variabilidad

El signo de los tiempos actuales es que nunca en la historia los ciudadanos, instituciones y empresas de todo tamaño hemos tenido acceso a tantos y tan variados datos. La cuestión es qué hacer con ellos y cómo.

Hay una gran variedad de tecnologías, herramientas y disciplinas científicas que intentan lidiar con esa enorme cantidad de datos dispersos que están esperando a ser explotados. Pero para que el consumidor medio pueda explotarlo, o mejor dicho pagar para que se lo hagan, es necesario generar una marca fácilmente reconocible que prometa algo grande sin decir qué: de ahí la marca "*big data*", que en sí alude a los sistemas que gestionan grandes, enormes, con-

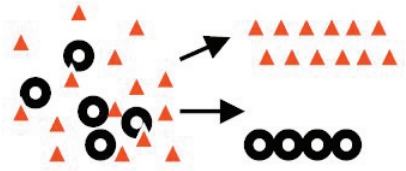
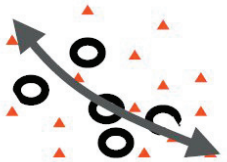
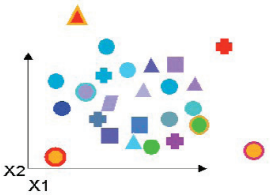
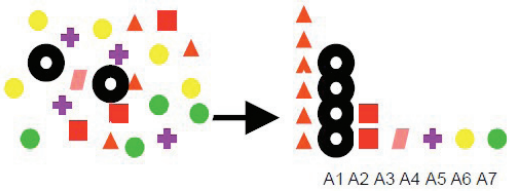
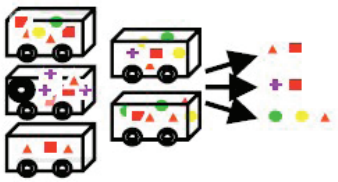
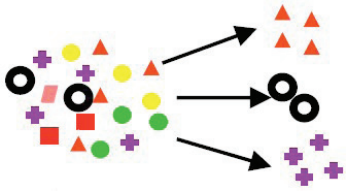
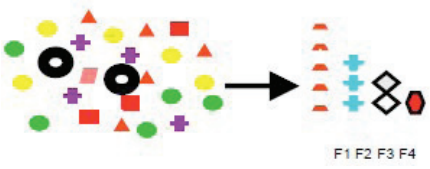
Operaciones y técnicas	Algoritmos	Aplicaciones
Clasificación 	-Regresión logística (modelos lineales generalizados, *GLM*) -Árboles de decisión -Bayes naïf -Máquinas de vectores soporte (*support vector machine*, *SVM*)	-Técnica estadística clásica -Popular / reglas / transparencia -App embebida -Amplios / datos estrechos / texto
Regresión 	-Regresión múltiple (*GLM*) -Máquinas de vectores soporte (*SVM*)	-Técnica estadística clásica -Amplios / datos estrechos / texto
Detección de anomalías 	-*SVM* de una clase	Falta de ejemplos del campo objetivo
Importancia de atributos 	-Longitud de descripción mínima (*minimum description length*, *MDL*)	-Reducción de atributos -Identificación de datos útiles -Reducción del ruido de los datos
Reglas de asociación 	*Apriori*	-Análisis del cesto del mercado -Análisis de enlaces
Clustering (agrupación en racimos) 	-Jerárquico K-media -Jerárquico O-cluster (*orthogonal partitioning clustering*, de *Oracle*)	-Agrupación de productos -Minería de textos -Análisis de genes y proteínas
Extracción de características 	Factorización no negativa de matrices (*non negative matrix factorization*)	-Análisis de textos -Reducción de características

Figura 1. Esquema de funciones y algoritmos. Adaptado de **Berger** (2012), p. 32.
http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataoracleadvanalytics11gr2-1930894.pdf

juntos de datos (*data sets*).

Dentro de esta gestión hay varios problemas de por sí importantes: cómo capturar esos datos y de qué fuentes; dónde almacenarlos, con el costo que conlleva; cómo encontrar la aguja de información en ese pajar de datos y a ser posible, en tiempo real; cómo analizarlo para generar conocimiento práctico con esas agujas de información y, finalmente, cómo representarlo, cómo visualizar ese análisis para que sea comprensible por quienes tienen que tomar las decisiones.

Para esa toma de decisiones, probablemente uno de los cambios experimentados más interesantes viene del hecho de que antes una empresa sólo tenía acceso a sus propios datos financieros, y acaso a algún informe realizado por una consultara sobre un sector completo o hasta cierto punto extrapolable. Ahora puede combinar su información interna con datos del mercado, de lo que se dice de los productos propios y la competencia en internet, de la conducta de los usuarios, de otros estudios sectoriales, de publicaciones científicas, de los datos que abastecen a esas publicaciones científicas, de...

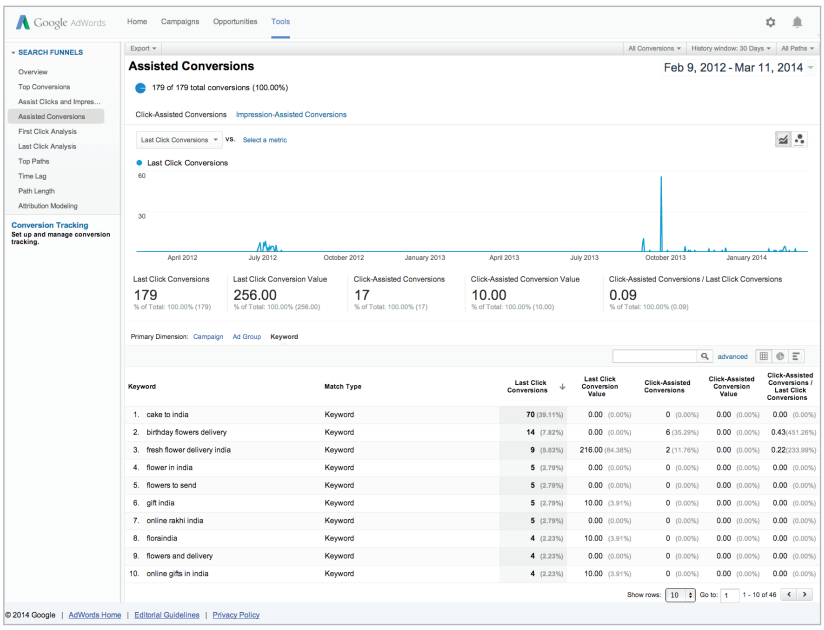


Figura 2. Ejemplo de análisis de conversiones asistidas por más de un canal. Fuente: *Google.com*

Analítica web

Imaginemos una tienda online. Las hay de todo tipo y tamaño. En principio pensaremos que sólo empresas de internet de la envergadura de *Amazon* o *eBay* harán uso de estas tecnologías (y lo hacen) pero no tiene por qué ser así. Una pequeña tienda online utiliza hoy día ya aplicaciones de analítica web para conocer qué hacen los usuarios que les visitan hasta que compran un producto, cuáles han sido los canales de publicidad y comunicación más eficaces y eficientes, o descubrir fallos de usabilidad o arquitectura de información, que impiden al usuario encontrar y utilizar lo que el responsable del sitio web desearía que encontrara y utilizara.

Sólo entender bien qué es lo que está ocurriendo dentro de su casa ya es un gran adelanto para esa pequeña tienda online, que utiliza programas a veces gratuitos como *Google Analytics* o *Piwik*. Éstos emplean la estadística más básica para analizar los datos de la interacción de los usuarios, acompañándose de gráficos para visualizar o resumir esas interacciones, normalmente en forma de series temporales, gráficos de barras, rankings de contenidos o flujos de navegación.

El problema de la analítica web en ese sentido, no es tanto saber cómo funciona el programa para explotar todas sus posibilidades (que también) sino saber qué buscar, cómo conectar unos datos con otros para encontrar las razones que llevan a los usuarios a clicar o no en un enlace, en un botón de compra, o a realizar las acciones que el dueño del sitio web desearía que se produjesen, lo que podríamos denominar en inglés *better queries* (mejores preguntas) en contraposición a más datos o mejores algoritmos, que parece ser un tema de discusión recurrente en la bibliografía relacionada. Porque el análisis se hace más rico al cruzar distintos indicadores, y de no hacerlo así, podemos fácilmente llegar a conclusiones arriesgadas.

Por ejemplo, sigamos imaginando que esa tienda online quiere mejorar su número de productos vendidos, y para ello realiza diversas acciones de marketing en distintos canales (redes sociales, publicidad en buscadores, publicidad gráfica, email marketing...), que va midiendo. Si contrasta el número de usuarios que llegan al sitio web desde cada canal con las ventas producidas directamente tras llegar de ese canal (lo que en *Google Analytics* se conoce como *last click attribution model*), y llega a la conclusión de que sólo uno de esos canales tiene correlación muy fuerte con un aumento número de ventas inmediatas, corre el riesgo de deducir que debe eliminar su gasto en los otros canales, debido a un típico error, asimilar correlación con causalidad. Y aunque ya sabemos que correlación no implica causalidad, los indicadores nada correlacionados ciertamente no parecen ser causa de ventas directas.

> Uno puede tener la tentación de predecir hacia dónde irán las cosas. Y eso es lo que todo el mundo intenta hacer: lo llaman '*predictive analytics*'

Y ese es el problema en este caso, analizar sólo las ventas directas (las que se realizan directamente tras clicar en un banner, por ejemplo). Quizá, si la tienda online utilizara un modelo de atribución que estudiara cómo cada canal asiste a los demás a convencerse de que es aquí donde debe realizar su compra, ayudando al usuario que viene al portal, se va a comparar un producto en otros sitios web, consulta foros o descubre opiniones en redes sociales, cambiaría su interés en desechar ciertos canales, por encontrar acaso

indicios de que los distintos canales se ayudaban más de lo que la correlación en ventas directas sugería.

Asimismo, otro problema añadido será adaptar los datos que aporta la aplicación al contexto de uso, a partir de unos objetivos contra los que medirse o unos competidores con los que compararse. Porque el mismo dato puede significar cosas distintas, dependiendo del contexto. Por ejemplo, tomemos el concepto de *engagement*, muy usado en marketing y en gestión de comunidades o redes sociales. En sí *engagement* (para esa tienda online) alude a la implicación o grado de relación entre usuario y marca. Según *Google Analytics*, el *engagement* se mide en función de "cuánto tiempo permanecen los usuarios en su sitio (en segundos) y el número de páginas que visualizan". Pero un sitio web puede entender que debe medir de otra forma más rica ese compromiso o implicación de los usuarios en su grado de interacción alrededor de un sitio web o de una marca, y por tanto, seleccionar qué fuentes (externas tipo *Twitter*, internas tipo logs de navegación), datos e indicadores va a utilizar para determinar si una marca, un servicio de información o una tienda online consigue que sus usuarios sean más fieles, e interactúen más y mejor con ésta, de forma que rentabilicen esas interacciones, dependiendo (de nuevo) en cada caso de cómo defina "rentabilidad".

> "Los datos de navegación dentro del sitio web (el reino de la analítica web) se suman ahora a los datos que hay en internet (cibermetría)"

De la analítica web a la cibermetría, *web science*…

Por tanto en un entorno tan competitivo como el actual, ese conocimiento interno probablemente ya no basta. El usuario puede tener implicación para con la marca o sitio web no sólo visitando éste, sino hablando bien (o mal) de él, compartiendo sus contenidos o productos, liderando opiniones… Hace falta pues dotarse de nuevas herramientas para integrar la diversidad de datos que pueden usarse en los distintos análisis, descubrir y gestionar mayor variedad de fuentes de datos, y mejorar la velocidad de procesamiento de los mismos, para entender mejor el escenario en el que ese servicio, proyecto o producto web se mueve.

Los datos de navegación dentro del sitio web (el reino de la analítica web) se suman ahora a los datos que hay en internet (cibermetría) respecto de:

- nichos de mercado online;
- conductas de búsqueda de los usuarios en otros sitios web y en distintos buscadores;
- microsegmentación de esos potenciales clientes;
- la competencia;
- relaciones entre esos competidores y otros sitios web medidas en forma de enlaces;
- relaciones entre las marcas y/o productos y los usuarios que los consumen medidas en forma de comentarios, retweets, respuestas, likes…;
- noticias que hay relativas a las temáticas de esos productos;
- innovación científica que permite intuir hacia dónde va la propuesta de valor única tecnológica de la competencia;
- etcétera, etcétera.

Aquí entramos en el reino de la inteligencia competitiva y la vigilancia tecnológica, en la que necesitaremos la combinación de disciplinas que trabajen con gran cantidad y variedad de datos (*big data*) como la cibermetría, *business intelligence*, *data mining* o *text mining*.

Esta variedad y variabilidad de los datos es un problema en constante evolución. Cuantas más fuentes de datos y más cambiantes, a menudo debido a razones difíciles de detectar, más ángulos de visión tendremos sobre un problema dado, más factores potencialmente causales podremos emplear en los estudios, y más probabilidades habrá de que los análisis sean más ricos, pero más difícil será detectar las fuentes de datos más útiles. La capacidad para lidiar con la complejidad inherente a esta conjugación de datos de fuentes heterogéneas es una de las áreas en las que más está avanzando, por ejemplo gracias no sólo a los repositorios de datos científicos, sino además, a que hoy día se comparten fuentes de datos a través de internet mediante literalmente miles de APIs, que ya incluyen *internet of things*, tomando como fuentes de datos los de productos *wearables* (vestibles) o los producidos por una ciudad en proyectos de *smart cities*.

Otra de las áreas que necesitan avanzar es la de la estandarización, como la que realiza el *Data Mining Group*, creando el lenguaje PMML (*predictive model markup language*) que permite que los modelos que se generen sean interoperables, con independencia del sistema con el que han sido construidos.

Mejores datos versus mejores algoritmos

Cisco estima que en 2012 cada día fueron creados cerca de 2,5 trillones de bytes de datos. Con tantos datos al alcance de la mano, uno en principio pensaría que el principal problema es tecnológico: ¿qué herramientas usar a un precio razonable, dónde custodiar tanta información, qué algoritmos emplear para obtener las conclusiones de mis análisis, y con cuánta velocidad antes de que éstas ya no sean útiles? Estas preguntas han dado lugar a una auténtica explosión de empresas dedicadas a crear herramientas de análisis, sistemas de almacenamiento, técnicas de programación…

Los campos y sectores económicos de aplicación son prácticamente cualesquiera que imaginemos, desde el periodismo de datos (conocido también como periodismo computacional) a la prevención de pandemias, pero permanece la discusión sobre si utilizar mejores algoritmos o más datos. De hecho, incluso se están estudiando algoritmos que usan muy pocos datos (*little data*) aunque lo hacen sobre segmentaciones muy específicas, a las cuales habrán llegado tras analizar gran cantidad de ejemplos o posibles segmentos, por lo que el problema permanece.

Para enfrentarse a las cantidades masivas de datos hay dos grandes escuelas, dos grandes disciplinas, la estadística y la informática con gran diversidad de técnicas:

- *behavioural analysis* o análisis de la conducta;
- *predictive analytics* o análisis predictivo;
- segmentación;
- *sentiment analysis* (mediante *text mining* a la que se aplica procesamiento del lenguaje natural);
- clasificación;
- agrupamiento o *clustering*;
- aprendizaje supervisado y no supervisado;
- regresión;
- árboles de decisión;
- inferencia;
- reconocimiento de patrones;
- representación del conocimiento;
- cálculo de probabilidades;
- reglas de asociación...

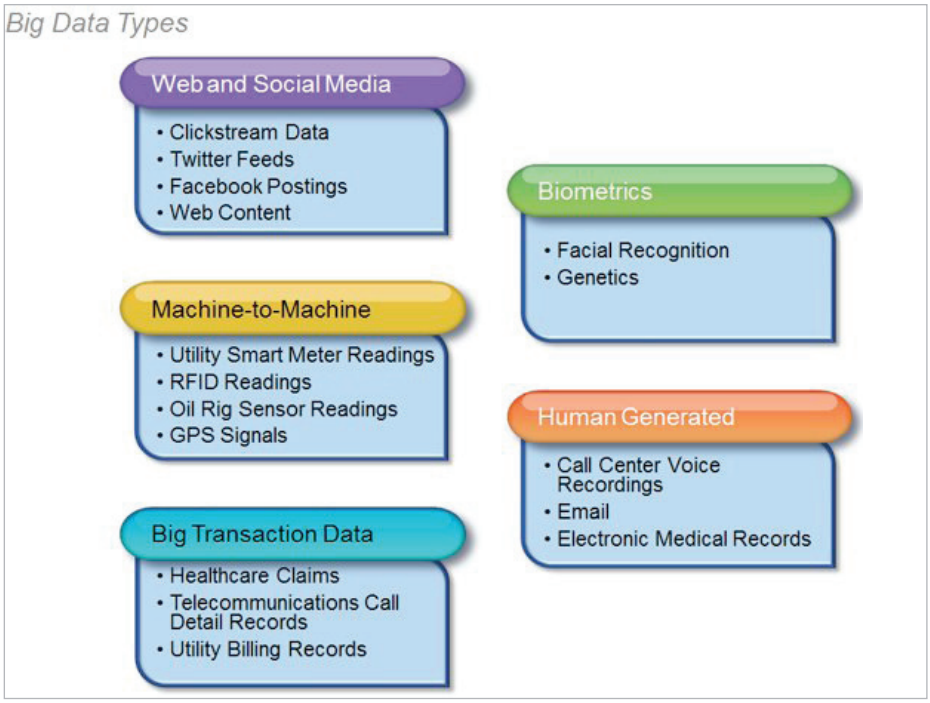


Figura 3. Algunos tipos de *big data*. Fuente: *http://ibm.com*

Actualmente la estadística y la informática se mezclan y tienden a entenderse para contestar distintos aspectos de un mismo problema, aunque no sea nada fácil encontrar a especialistas que abarquen ambas disciplinas enteramente, por lo que las líneas se desdibujan. Probablemente antes debemos entender qué técnicas contestan qué problemáticas, y de ahí aún antes, qué preguntas queremos que sean contestadas.

Conclusión

Hace falta un puente entre lo que necesita un cliente (sea un bibliotecario, un técnico de un ayuntamiento, el CEO de una *start-up* de internet…) y quien conoce cómo obtener respuestas de los datos. Cuando este cliente hable con un especialista en *machine learning* o en estadística, aquél necesita qué, o bien se tengan muy claras las preguntas para las cuales se vislumbra las posibilidades de solución mediante alguna o varias de las técnicas mencionadas, o bien se le forme en la especialidad sectorial o temática, para que desde la suya imagine qué soluciones podría dar sin esperar a la pregunta perfecta. Y es un puente que se podría acortar con más formación… por ambos lados.

Bibliografía

Barranco-Fragoso, Ricardo (2012). *¿Qué es big data?* IBM. *http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data*

Berger, Charlie (2012). *Big data analytics with Oracle advanced analytics in-database option*. *http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataoracleadvanalytics-11gr2-1930894.pdf*

Data Mining Group. *The Data Mining Group releases PMML v 4.2. http://www.dmg.org/DMGReleasesPMMLv4.2.pdf*

Garrett, Wu (2013). "Why more data and simple algorithms beat complex analytics models". *DataInformed. Big data and analytics in the enterprise*. *http://data-informed.com/why-more-data-and-simple-algorithms-beat-complex-analytics-models*

Google. *About the attribution models. https://support.google.com/analytics/answer/1665189*

Google. *Engagement. https://support.google.com/analytics/answer/1144430?hl=en*

MuleSoft. *The top 10 internet of things APIs. http://www.mulesoft.com/infographics/api/internet-things*

Winshuttle (2014). *Big data y la historia del almacenamiento de la información. http://www.winshuttle.es/big-data-historia-cronologica*