F. 4. *Big data* y *not so big data*

Jorge Serrano-Cobos

30 noviembre 2012

Serrano-Cobos, Jorge (2013). "*Big data* y *not so big data*". *Anuario ThinkEPI*, v. 7, pp. 161-163.

Resumen: Se muestra el concepto de "big data" como sistema de análisis de grandes cantidades de datos para descubrir tendencias, patrones, correlaciones e incluso causalidades para generar predicciones. Se adapta esta disciplina al contexto del trabajo diario del profesional de la información, hablando de "small data" y de analítica web. Se ofrecen diversos ejemplos de herramientas que pueden permitir gestionar y analizar enormes cantidades de datos públicos.

Palabras clave: *Big data*, *Small data*, *Linked data*, Bases de datos, Analítica web, Análisis de datos, Visualización de información, Cibermetría, Webometría, *Search analytics*.

Title: Authority control ontologies in the field of linked open data

Abstract: The concept of "big data" is explained as a system of analyzing large amounts of data to discover trends, patterns, correlations and even causality to generate predictions. This discipline can be adapted to the context of the daily work of the information professional dealing with "small data" and web analytics. Examples are provided of tools that support the management and analysis of huge amounts of publicly available data.

Keywords: Big data, Small data, Linked data, Databases, Web analytics, Data analysis, Information display, Cybermetrics, Webometrics, Search analytics.

Big data

Big data es uno de esos conceptos que están con nosotros desde hace mucho, pero que periódicamente saltan a la palestra de los medios y gozan de su minuto de gloria (o semana, mes o año...).

El término hace referencia a sistemas que manipulan enormes cantidades de datos, sobre los que ejecutan diferentes tipos de análisis con técnicas propias de *business analytics*, *data mining* o *text mining* para buscar patrones. Podemos hablar de volúmenes de terabytes, petabytes (1.000.000.000.000.000 bytes, o más).

Las dificultades inherentes a gestionar semejantes cantidades de bytes son fácilmente imaginables: encontrar y obtener los datos, almacenarlos de forma que la organización tenga acceso rápido a ellos y/o compartirlos con clientes actuales o potenciales, encontrar una forma de buscar la aguja en el pajar y, por supuesto, analizar los datos y mostrarlos de forma resumida y visualmente clara para convertirlos en información apropiada para la toma de decisiones...

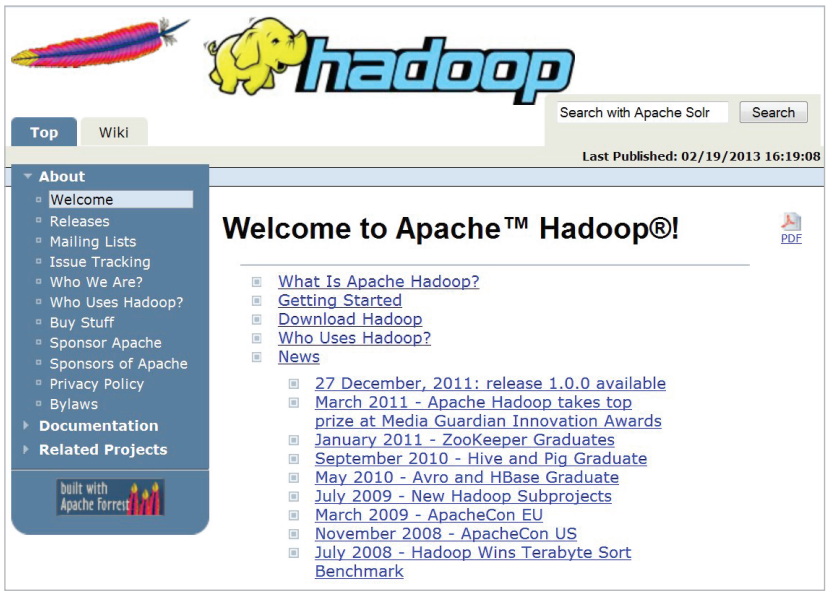
Grandes compañías como *Oracle*, *Google* o *IBM* se han apuntado al carro de proporcionar productos (con tendencia al *cloud computing*) que permitan a compañías no tan grandes utilizar y computar esa información. La herramienta reina actualmente es quizá *Hadoop*. *http://hadoop.apache.org*

> **"*Big data* hace referencia a sistemas que manipulan enormes cantidades de datos, sobre los que ejecutan diferentes tipos de análisis"**

Las aplicaciones de este tipo de técnicas son prácticamente infinitas: marketing, logística, gestión de recursos hídricos, estudio del lenguaje humano, genómica, investigación científica en general...

Big data en Información y Documentación

Al enfrentarnos al desafío del *big data* desde la perspectiva de Información y Documentación, una opción puede ser reunir un profesional encargado de conceptualizar los objetivos que



Hadoop

se quieren conseguir (qué indicadores buscamos, qué conclusiones esperamos, cuáles son las hipótesis de trabajo que queremos confirmar o refutar) y un especialista en datos, con una fuerte formación combinada en programación, gestión de sistemas y matemáticas de alto nivel.

En principio se puede pensar que hablamos de conceptos alejados de nuestra realidad profesional, pero no tiene por qué ser así. Aunque en muchos casos nuestros proyectos se circunscriban más al ámbito de "*not so big data*" (más conocido como *small data*), es interesante investigar y conocer mejor las posibilidades que las técnicas de análisis inherentes a *big data* pueden aportarnos.

Estamos más familiarizados con la disciplina de la analítica web (o *digital analytics*, es decir, análisis de datos no sólo de la Web, como las *apps* de los móviles), cuyo espejo a gran escala sería *data analytics*. Pasamos de *descriptive analytics* mostrando "lo que hay" -a veces a través de aplicaciones de *visualization analytics* en 2D ó 3D- a usar técnicas de modelización predictiva que permiten descubrir tendencias, patrones, correlaciones, e incluso más allá, causalidades para generar predicciones, lo que se denomina *predictive* o *prescriptive analytics*. La bibliometría y cibermetría tienen mucha relación con este tipo de analítica.

Aplicaciones

Algunas herramientas de *big data* o *small data* gratuitas son, por ejemplo:

– *Yahoo clues*

Muestra visualmente tendencias de búsqueda, muy útil para segmentación de mercados digitales.
http://clues.yahoo.com

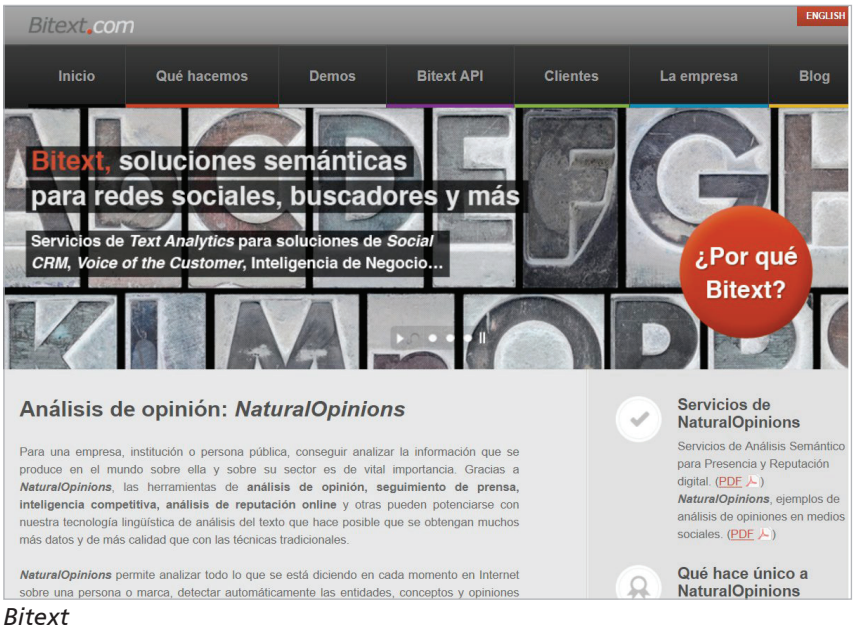
– *Google keywords tool*

Permite descubrir las palabras clave, frases de búsqueda o *queries* más buscadas en torno a un concepto utilizado como semilla. El sistema contabiliza millones de expresiones y datos históricos. Nosotros sólo recibimos la pequeña parte relacionada con lo que pedimos, pero podemos pedir mucho.

En un reciente informe realizado sobre *e-commerce* de vino, estudiamos más de 6.000 expresiones de búsqueda distintas en 6 idiomas, que correspondían a 128 millones de búsquedas de media mensual, para encontrar entre otras cosas cómo se busca el vino español en comparación con vinos de otros países. Podemos combinar esta herramienta con *Google trends* para ayudar en la visualización de la información (temporal o geográficamente).
https://adwords.google.com/o/KeywordTool
http://bit.ly/Twmvui

– *Microsoft academic search*

Buscador y visualizador de relaciones. Con sus limitaciones y sesgos está haciendo bastante por presionar a otros para mejorar las aplicaciones de



Bitext

http://oemv.es/esp/el-mejor-estudio-sobre-el-vino-en-internet-realizado-en-espana-681k.php

análisis y visualización bibliométricas.

http://academic.research.microsoft.com

– *Google correlate*

Permite buscar términos de búsqueda con similares patrones que uno dado, o introducir nuestros propios datos para que los analice.
http://www.google.com/trends/correlate

"Se puede pensar que hablamos de conceptos alejados de nuestra realidad profesional, pero no tiene por qué ser así"

– *Google fusion tables*

Aplicación con la que se pueden combinar y visualizar distintos conjuntos de datos, alojables en la nube.

Por ejemplo, en *MASmedios.com* se ha usado para visualizar mediante geolocalización acciones realizadas en ubicaciones físicas por millones de ciudadanos.
http://support.google.com/fusiontables/answer/2571232/?hl=en&
http://www.eleccions2011.gva.es/es/ciudadanos/mapas-municipales

– *Google n-gram viewer*

Ayuda a analizar históricamente las palabras más usadas en libros de *Google books,* menciones a autores, etc. Hay que tener precaución con los posibles fallos en la digitalización, que pueden dar resultados "divertidos".
http://books.google.com/ngrams

– *Google public data explorer*

Permite explotar las posibilidades de visualización de datos obtenidos de entidades públicas o nuestros propios datos, lo que nos lleva a conectar *big data* con *linked data* y *open data*.
http://www.google.com/publicdata/directory

– En el sector de la *social analytics* han surgido multitud de aplicaciones que usan las apis de distintas fuentes (*Google*, *Facebook*, *Twitter*, etc.) para obtener datos con los que realizar comparaciones de todo tipo y extraer tendencias en tiempo real, incluyendo lo que se denomina *sentiment analysis* para evaluar la opinión que los usuarios tienen de una marca, como realiza la compañía española *Bitext*.
http://www.datasciencecentral.com/video/real-time-analytics-for-small-data-big-data-and-huge-data
http://www.bitext.com/actividad/soluciones/sol_naturalopinions.html

– Las apis de estas fuentes son intensamente utilizadas en multitud de ámbitos, generando distintas herramientas, desde el seo al turismo, pasando por la webometría. Se utiliza en nuestro campo para realizar inteligencia competitiva, investigaciones longitudinales de *search analytics,* posicionamiento en buscadores, etc., en áreas tan dispares como el *e-commerce* y la exportación o los rankings universitarios, por poner ejemplos, pero las posibilidades están por explorar.

Más información

http://www.datasciencecentral.com
http://www.kdnuggets.com