IV
JORNADA PROFESIONAL DE LA RED
DE BIBLIOTECAS
DEL INSTITUTO CERVANTES:
Big Data y bibliotecas: convertir
Datos en conocimiento

MADRID 11 DE DICIEMBRE DE 2014

Big Data *versus* Better Queries: analítica web práctica para servicios de información

Jorge Serrano Cobos
Director I+D en MASmedios.com
Director del Máster de Marketing Digital de INEDE (UCV)
Profesor Asociado Universidad Politécnica de Barcelona

Resumen:
Se perfilan distintas etapas y herramientas en el análisis cuantitativo de servicios de información, tanto sobre sus usuarios como sus fondos, y la integración de los distintos análisis en la gestión de servicios de información y la optimización de las prácticas bibliotecas utilizando los datos internos y externos que el servicio tiene a su disposición, con posibles aplicaciones prácticas de la Analítica Web, la Cibermetría, Data Mining y Text Mining, lo que se denomina "Bibliomining".

Palabras clave: Bibliomining, Data Mining, Analítica Web, Big Data, bibliotecas, gestión de servicios de información

Aproximadamente desde 2011 se empieza a escuchar el concepto de "Big Data", definido por Gartner como el "conjunto de activos de información caracterizados por su gran volumen, velocidad y variedad, que exigen formas innovadoras y rentables de procesamiento de la información para mejorar la comprensión y la toma de decisiones".

Su aplicación llega a todos los sectores, desde la salud a la banca, y por tanto aplica preguntarse ¿puede resultar útil para la Gestión de la Información, y en concreto para el desempeño de los Servicios de Información? En un principio la respuesta sería "depende", obviamente de si hay a disposición del Servicio de Información un conjunto de datos o activos de información con las características antes mencionadas. Pero lo primero que necesita el profesional de la Información es entender o visualizar las aplicaciones que puede tener el análisis de los datos cuantitativos a su disposición, y después ya verá si eso entra en el rango del "Big Data" o del "Small Data".

Así, tomemos un servicio de información prototípico, por ejemplo una biblioteca que tiene a disposición pública de sus usuarios un sitio web con un OPAC para recuperar información de los contenidos que puede utilizar (ya sean libros, artículos, DVDs, ebooks, pdfs, etc.) Partimos de un hecho básico: sin datos que analizar, estamos ciegos, no sabemos bien qué hacemos mal y qué o cómo mejorar; con datos,

podremos inferir mejor qué dirección tomar, o al menos en qué experimentar con mayor probabilidad de impacto en nuestros objetivos, para después evaluar. Por tanto, la cuestión es qué y cómo analizar.

Veamos hasta dónde puede llegar el análisis cuantitativo desde dentro (del sitio web) hacia fuera. Primero analizaremos el sitio web y el OPAC mediante herramientas de Analítica Web. Alzaremos luego la mirada para abarcar el ecosistema que rodea a nuestro sitio web bibliotecario, conformado por competidores, colaboradores, usuarios que mantienen presencia digital en redes sociales, etc., lo que entra en los análisis propios de la disciplina de la Cibermetría, donde entonces también podremos usar distintas técnicas de análisis cuantitativos más propios del Big Data.

1. Analítica Web en Bibliotecas y Servicios de Información

La primera pregunta que debemos hacernos es para qué analizar. Para un bibliotecario, cabe preguntarse para qué necesita la Analítica Web si ya tiene las estadísticas de uso que le da el OPAC. Bien, esas estadísticas se fijan sobre todo en el final del proceso, ofreciendo en su mayoría rankings o tops (los libros más prestados, más descargados, los lectores más frecuentes…) pero no nos informa del principio del proceso de búsqueda de la información hasta llegar a ese final, ni lo de lo que sucede durante ese proceso.

Si el servicio de información se ofrece a través de Internet, los sistemas actuales de Análisis Web permiten capturar toda la traza de acciones que llevaron a ese usuario a llegar a un contenido concreto o a darse de alta como socio de la entidad. Y si ese servicio de información, por ejemplo una biblioteca, quiere mejorar sus objetivos (por ejemplo tener más usuarios y socios) hoy día es necesario invertir en mejorar tanto las formas de conseguir más usuarios, como de convencer y fidelizar a los que llegan. Por tanto, la Analítica Web se constituye una herramienta utilísima para descubrir qué hacemos mal o qué hacemos bien, en esa misión.

Si ya estamos convencidos de para qué usar la Analítica Web, ¿qué se ha de analizar? Se desglosa aquí una propuesta de 5 grandes macrodimensiones que sería preferente analizar en cualquier sitio web: objetivos de la institución, usuarios, competencia, canales, contenidos.

Fig. 1: Macrodimensiones de análisis en sitios web

¿Cuándo analizar? Hay varios momentos, antes de tener el sitio web publicado (en vivo y de cara al público, cuando lo estamos construyendo) y/ o después. En una situación ideal, se deberían hacer distintos análisis antes y después, por ejemplo:

- Antes de publicar:
  - Análisis de mercado: estudiaremos a los usuarios potenciales y sus necesidades
  - Análisis de la Competencia: comprobaremos sus puntos fuertes y débiles, y los contrastaremos con los nuestros
  - Análisis de la Usabilidad (Contenidos): confrontando lo que vemos en el análisis de mercado con los contenidos que les podemos ofrecer.

- Después de publicar:
  - Analítica Web: servirá para entender qué hacen los usuarios que ya vienen al sitio web
  - CRM (Customer Relationship Management): si la hay, es una herramienta valiosísima para añadir el análisis cualitativo al análisis cuantitativo, y ayudar a entender el "por qué" de la conducta de nuestros usuarios.

Así, lo primero que haremos será establecer los objetivos que el servicio de información se va a marcar, para contrastar en el futuro las mejoras hechas con los resultados obtenidos. Algunos ejemplos serían:

- Objetivo 1: incrementar el número de usuarios registrados
- Objetivo 2: Incrementar el número de suscriptores del newsletter del servicio en un XX%.
- Objetivo 3: aumentar los préstamos a usuarios en un X%

Después, debemos analizar a nuestros usuarios, tanto los actuales si los hay, como a los potenciales, los que podrían ser y no son. Debemos intentar segmentarlos, tipificarlos, y entender sus necesidades, y cruzarlas con las necesidades de nuestro sitio web, para lograr hacer coincidir nuestros objetivos con los suyos, y que encuentren en nuestro servicio de información el lugar idóneo.

Hay distintas herramientas para hacerlo, desde herramientas para generar encuestas online a herramientas para conocer qué buscan los usuarios y cómo buscan sobre los contenidos que el servicio de información ofrece, como Google Keywords Planner, que nos permite entender cuánto y cuándo se buscan más ciertas expresiones de búsqueda relativas a los contenidos sobre los que queramos saber el lenguaje natural de búsqueda de los usuarios potencialmente interesados.

A partir de ahí, y una vez hayamos generado el sitio web y abierto el OPAC al público, podemos empezar a analizar lo que ocurre mientras los usuarios van utilizando la información almacenada en el mismo. Hay varias herramientas para hacerlo, desde las que analizan logs a las que utilizan tags, siendo éstas las más completas desde el punto de vista del análisis de interacciones de los usuarios. De éstas, tenemos ejemplos como el software Open Source Piwik (http:(//www.piwik.org) aunque muy probablemente utilizaremos Google Analytics, herramienta gratuita con múltiples posibilidades.

Si es así, podemos empezar preparando nuestra cuenta de Google Analytics para segmentar el público al que se va a analizar, creando cohortes de usuarios con diversos criterios de segmentación.

Podemos filtrar por ejemplo al público de la IP o IPs que utilicen los trabajadores del servicio, para así no contaminar los resultados (es posible que nosotros naveguemos muchas veces por nuestro propio sitio web, y a buen seguro nuestra conducta de navegación no será similar a la de nuestro público)

También podemos crear dos vistas, una incluyendo sólo a un cierta IP para analizar a aquellos usuarios que utilizan los equipos conectados al sitio web desde el propio edificio del servicio o biblioteca, y por otro lado otra vista excluyendo a esta IP, para sólo analizar a los usuarios que provengan de internet (que consulten el sitio web desde sus hogares u otros lugares públicos o privados)

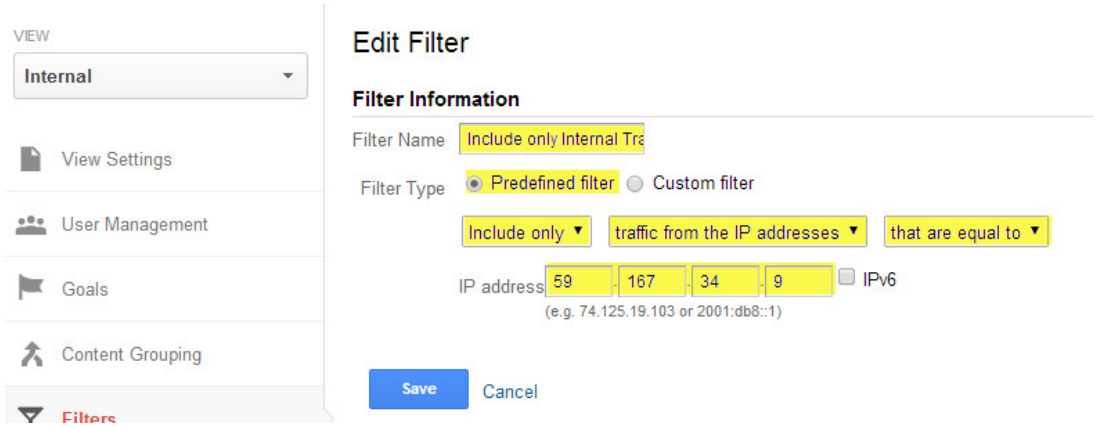

Fig.2: Filtro de IP con Google Analytics

Uno de los primeros análisis que aplica muy especialmente al caso de sitios web que utilicen OPACs es precisamente su estudio, lo que en Google Analytics se realiza con la opción "site search".

Esta funcionalidad permite decirle a Google Analytics cuáles son las variables de búsqueda que utiliza el buscador interno del OPAC, y así comunicarle qué palabras clave, qué expresiones de búsqueda son utilizadas por los usuarios del sistema, y desde qué páginas o contenidos del sitio web (incluido el catálogo) se produce esa búsqueda.

La información cualitativa se une aquí a la cualitativa, permitiéndonos descubrir qué contenidos provocan la incertidumbre en el usuario, o despiertan su curiosidad, o si describimos las fichas del catálogo de forma que se entiendan con el lenguaje natural del usuario o no, y hay que mejorar esa descripción incluyendo sinónimos o conceptos relacionados semánticamente.

Después, investigaremos las tendencias, contrastando la evolución de nuestros objetivos en función de los datos obtenidos para saber si (para cada objetivo) vamos mejor o peor. Por ejemplo, podemos analizar la tendencia o evolución del denominado embudo de conversiones, en el cual nos preguntaremos, de cada 100 usuarios que entran en el sitio web ¿qué porcentaje se transforma en conversiones? Entendiendo conversiones por acciones como registrarnos como usuarios, descargarnos algo, o reservar un libro, por ejemplo. ¿Y los que no convirtieron, dónde se fueron, dónde desistieron?

Directamente esos datos de evolución no nos darán la respuesta a "por qué", pero si nos permitirán intuir dónde puede estar el problema. Por tanto, podemos fijarnos en lo que genera más éxitos como en lo que menos, correlacionando diferentes métricas (páginas vistas, conversiones, tiempo en el sitio web…) de distintas dimensiones (usuarios, contenidos…) para observar por ejemplo qué canal de tráfico de usuarios es el que ha atraído más usuario que han llegado a convertirse en socios del servicio de

información, o por otro lado, qué contenidos son los que han provocado la misma reacción.

Una vez detectemos qué correlaciona mejor, podemos buscar las causas de esa fuerte correlación, realizando modificaciones medidas a lo largo del tiempo y experimentos como el testeo A/B o multivariante, en el que cambiamos ciertas variables de análisis para ver su efecto en las métricas analizadas para poder concluir las razones más probables. Por ejemplo, podemos cambiar el tamaño del botón de registro de socios (A/B testing) o hacer variaciones de color, tamaño y posición de ese botón (testeo multivariante) y ayudarnos de Google Analytics para que nos permita calcular la mejor combinación de factores.

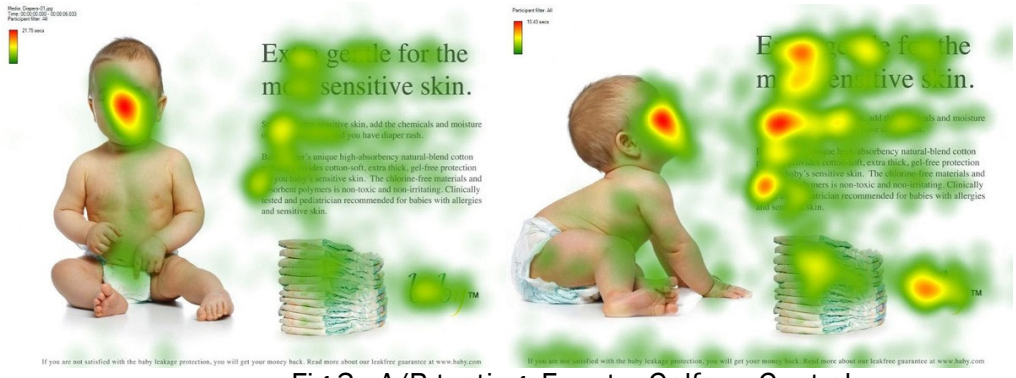

Fig.3.: A/B testing. Fuente: Galfano G, et.al.

Y más allá de la interacción interna de esos usuarios, ¿qué ocurre después? ¿Cuál es la interacción externa de los usuarios, lo que se denomina en marketing medios ganados? ¿Los usuarios se convierten en medios de propaganda en sus cuentas sociales mediante sus tweets, retweets, recomendaciones, conversaciones, etc., aumentando nuestra repercusión mediática, haciendo aumentar el número de enlaces hacia nuestro sitio web (denominados en Google Analytics "referrers") atrayendo así más usuarios potencialmente interesados en lo que tengamos que ofrecerles, o no es ése el caso?

Lo ideal con toda esta información es generar uno o varios cuadros de mando que permitan seguir en el tiempo los objetivos propuestos desde un inicio, acompañándolos de alarmas e informes enviados regularmente de forma automática, para estar al tanto de si estamos por encima o por debajo de las expectativas propuestas. Se puede llegar al extremo de preparar cuadros de mando especializados, como el que está realizando el servicio Mimas como parte de la Fundación JISC para las bibliotecas universitarias del Reino Unido, de forma que ese cuadro de mando se fije sólo en los indicadores de desempeño clave o KPI (Key Process Indicators) más importantes para los bibliotecarios.

1. Cibermetría

Pero no estamos solos en el ciberespacio. Hay más factores que delimitan las posibles razones del éxito o fracaso de nuestro servicio de información, en realidad como mínimo tantos como sitios webs tengan algo que ver con cualquiera de las palabras que contengan los contenidos de nuestro sitio web y que luchan con todos los demás por conseguir un mejor posicionamiento en las páginas de respuesta de los motores de búsqueda (lo que se denomina Search Engine Optimization o Posicionamiento en Buscadores y que se ha convertido en una disciplina del Marketing Digital)

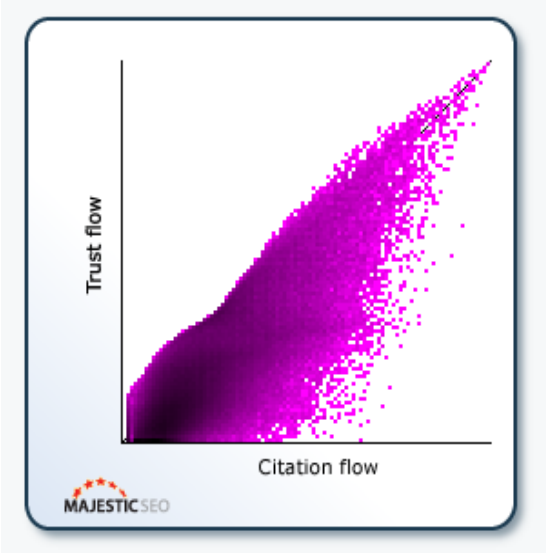

Fig.4.: Análisis de enlaces. Fuente: Majestic.com.

Todos esos sitios web son lo que se denominan nodos, formando parte una malla o red de sitios web que pueden estar intercomunicados por enlaces, tanto entre unos y otros como entre sitios web de terceros que hacen de enlace entre ellos, donde las razones para que un sitio web se enlace con otro son muy diversas. Normalmente unos están más interconectados que otros, por lo que es interesante descubrir cuáles son los más interconectados, y ver si hay formas de relacionarnos con ellos. ¿Por qué? Porque cuantos más enlaces haya hacia nuestro sitio web, y más aún si es desde sitios relacionados temáticamente, más probabilidades habrá para que un usuario potencialmente interesado en el contenido que podamos ofrecer llegue a conocer nuestro sitio web, lo que es además uno de los principios básicos del posicionamiento en buscadores desde el 98.

El estudio de métricas asociadas a los grafos como la centralidad, la densidad, el grado de intermediación o cercanía u otros, es propio de la disciplina científica denominada “Análisis de Redes Sociales” o Social Network Analysis (SNA) muy anterior a la aparición de medios sociales como Twitter o Facebook, con los que muchas veces se confunde.

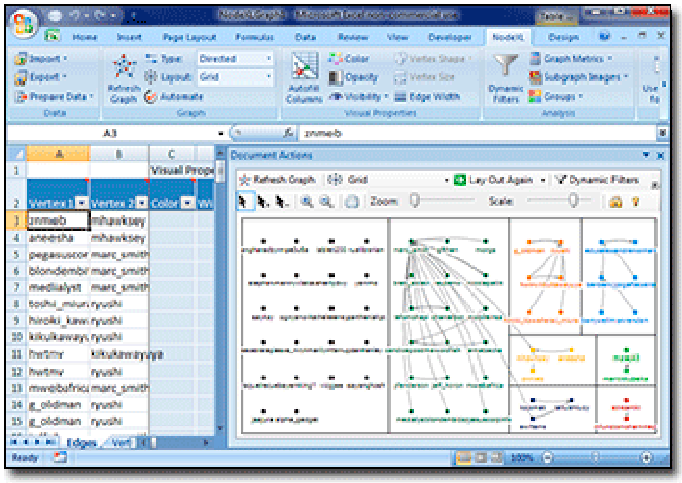

Fig.5.: Software de estudio de grafos en redes sociales. Fuente: NodeXL.

Pero este estudio sí puede utilizarse también en el campo de los social media o medios sociales, permitiéndonos encontrar a aquellos de nuestros usuarios que más frecuentemente comparten con otros usuarios sus quejas o su beneplácito para con nuestros contenidos y servicios, y que más y mejor interconectados están, con lo que tendrán más probabilidades de convertirse en medios de comunicación ganados que propaguen la "buena nueva" a otros usuarios de nuestra existencia.

La información externa que hay sobre otros sitios web, bien sean competidores directos (porque nos quieren hacer la competencia) o indirectos (porque utilizan las mismas palabras que nosotros usamos y aparecen por delante de los resultados de búsqueda para esa palabra o palabras/frases de búsqueda) puede ser recabada con distintas herramientas, y normalmente, cuanto más exhaustivo sea nuestro análisis, más herramientas deberemos utilizar, puesto que cada una fija su atención en cuestiones diferentes. Por ejemplo, Woorank (http://www.woorank.com) se fija en elementos internos del sitio web, Alexa (http://www.alexa.com) en el tráfico de paneles de usuarios en torno al sitio web, o ahrefs (http://www.ahrefs.com) o Majestic (http://www.majestic.com) en los enlaces que apuntan o parten de un determinado nodo o sitio web. Conociendo mejor a esos competidores, a nuestros usuarios, a los que podrían serlo, a nuestros contenidos y a las relaciones entre todo ellos, estaremos en mejores condiciones de tomar decisiones con mayores probabilidades de ayudarnos a llegar a los objetivos fijados.

1. Data Mining en bibliotecas: Bibliomining

Si nos hemos fijado, hemos visto cómo se ha evolucionado del simple análisis estadístico de porcentajes, tops o rankings de indicadores, y evolución de tendencias,

a buscar correlaciones, análisis causal bivariante o multivariante, e incluso el análisis de redes sociales.
Al subir el nivel en la panoplia de herramientas y fuentes de análisis de datos, aumenta el nivel de complejidad de los cálculos posibles. De hecho, podremos llegar a utilizar otros cálculos propios de la estadística inferencial, y a combinar esos cálculos con otros que vengan del mundo de la Computación, en concreto de las disciplinas asociadas con el Data Mining y el Text Mining.

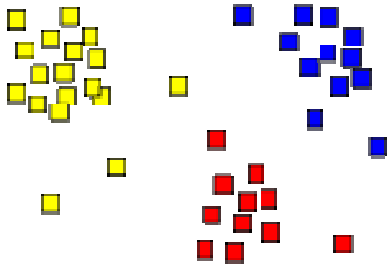

Fig.6.: Gráfico de clusters para segmentación de ítems. Fuente: Wikipedia.

La Minería de Datos o Data Mining busca patrones en los datos, lo que puede ser utilizado en diferentes aplicaciones y sectores. Aplicado al sector de las bibliotecas se ha dado en llamar Bibliomining. Podemos usar reglas de asociación para agrupar ítems de nuestra base de datos relacionados, bien de forma genérica, bien de forma personalizada o segmentada para grupos o clusters de usuarios detectados con o sin supervisión humana.

Se puede descubrir a aquellos usuarios con más probabilidades de venir ciertos días a la biblioteca física, y generar conjuntos o subconjuntos de usuarios por sus gustos, para organizar con ellos clubes de lectura que se reúnan en los días con más probabilidades de que la mayor parte de los miembros interesados puedan venir, predecir qué libros necesitarán más ejemplares y cuándo, etc.

1. Conclusiones

Las bibliotecas y servicios de información pueden acudir a las tecnologías de análisis de datos, ya sean internos o externos, con una visión no sólo de evaluación de presupuestos, sino que les pueden sacar mucho más partido.
Primero se deben diseñar objetivos que después se puedan medir, pero alineándolos con prácticas realizadas a lo largo de determinado rango de tiempo, para determinar qué se hace bien y qué no, a fin de revisar esas acciones en procesos iterativos de experimentación y mejora.

Para eso, los servicios de información deben y pueden profundizar más en su conocimiento del usuario al que dan servicio y sus necesidades, tanto actuales como probables, y luego los contenidos que pueden utilizar en esos servicios, al ir más allá de los indicadores básicos.
Si el servicio se ofrece vía web, hay que constatar que el servicio no está sólo, por lo que hay que considerar una gama mayor de factores a tener en cuenta, integrando la Cibermetría con la Analítica Web en nuestra búsqueda de las razones que puedan ayudar a optimizar el servicio.
En la aplicación de nuevas tecnologías de Data Mining y en la explotación de cantidades masivas de datos (Big Data) es aconsejable maridar bibliotecario más especialista en análisis de datos, para complementarse y aportar visiones integradores de un mismo problema.

Bibliografía

Aguillo, Isidro F. (2003). Cibermetría. Introducción teórico-práctica a una disciplina emergente. http://internetlab.cindoc.csic.es/cursos/cibermetria.pdf
Galfano G et. Al (2012). Eye gaze cannot be ignored (but neither can arrows). En: Quarterly journal of experimental psychology. 65(10):1895-910. doi: 10.1080/17470218.2012.663765.
http://www.ncbi.nlm.nih.gov/pubmed/22512343
Gartner. Gartner IT Glossary > Big Data. http://www.gartner.com/it-glossary/big-data/
Nicholson, S. (2003) The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. Information Technology and Libraries 22 (4).
http://www.bibliomining.com/nicholson/biblioprocess.htm
Wikipedia. Social Network Analysis. http://en.wikipedia.org/wiki/Social_network_analysis