Sumario

En noviembre del pasado año, Swets & Zeitlinger Publishers fue adquirido por Taylor and Francis Group.

Por lo tanto, a partir de 2004 el editor de El profesional de la información es Taylor and Francis Group.

Esta adquisición no tiene efectos en lo que a periodicidad y funcionamiento se refiere, por lo que las contribuciones deberán seguir las mismas pautas dictadas en «Información para los autores».

Tampoco se realizarán cambios inmediatos en las suscripciones y distribución.

Para más información sobre la revista:

http://www.szp.swets.nl/szp/journals/pi.htm

El profesional de la información

Los contenidos de **El profesional de la información** están referenciados en los siguientes servicios bibliográficos y bases de datos:

Bedoc
http://www.inforarea.es/bedoc.htm

Bulletin Board for Libraries (Bubl)
http://bubl1.lib.strath.ac.uk/archive/journals/epdli/

Compludoc
http://www.ucm.es/BUCM/complu

ConnectSciences (Pascal)
http://connectsciences.inist.fr

Consorci de Biblioteques Universitàries de Catalunya (Cbuc)
http://sumaris.cbuc.es/13866710.htm

Datathéke
http://milano.usal.es/dtt.htm

Dialnet
http://dialnet.unirioja.es

DoIS (Documents in Information Science)
http://dois.mimas.ac.uk/DoIS/data/julqtichq.html

Ebscohost Electronic Journals Service
http://ejournals.ebsco.com/direct.asp?JournalID=105302

Extenza e-publishing services
http://www.extenza-eps.com/extenza/contentviewing/viewJournal.do?journalId=65

Guíame
http://www.guiame.net/lista-fuentes.html

GVA
http://www.pre.gva.es/argos/docus/webbolsum/Sumarios/sumcast/iwe.htm

Índice Español de Ciencias Sociales y Humanidades (Isoc)
http://www.cindoc.csic.es/prod/isoc-cd.html

Information Science and Technology Abstracts (Ista)
http://www.infotoday.com/ISTA

Information Services in Physics, Electronics and Computing (Inspec)
http://www.iee.org.uk/publish/inspec

IWEb
http://www.doc6.es/iwe

Library and Information Science Abstracts (Lisa)
http://www.csa.com/csa/factsheets/lisa.shtml

Oclc Firstsearch
http://www2.oclc.org/oclc/fseco/topic_area.asp?topic=Z

Registros Bibliográficos para Bibliotecas Públicas Españolas (Rebeca)
http://www.mcu.es/REBECA/que.html

Resúmenes de Información y Documentación (ReID)
http://www.sisdoc.es/servicios/reid.htm

SwetsWise
http://www.swetswise.com/link/access_db?issn=1386-6710

Universidad de Castilla-La Mancha
http://biblioteca2.uclm.es/biblioteca/sumarios/pi.pdf

Universidad de Oviedo
http://librivision.uniovi.es/web/sumarios_web/Profesional-de-la-Informacion/

Excavando la web

Por **Ricardo Baeza-Yates**

Resumen: *La web es el fenómeno más importante de internet, demostrado por su crecimiento exponencial y su diversidad. Por su volumen y riqueza de datos, los buscadores de páginas se han convertido en una de las herramientas principales. Son útiles cuando sabemos qué buscar. Sin embargo, es seguro que la web tiene muchas respuestas a preguntas nunca imaginadas. El proceso de descubrir relaciones o patrones interesantes en un conjunto de datos se llama minería de datos (del inglés data mining) y en el caso de la web se llama minería de la web (web mining). En este artículo presentamos las ideas más importantes en minería de la web y algunas de sus aplicaciones.*

Ricardo Baeza-Yates es doctor en Computer science por la Univ. de Waterloo, Canadá, desde 1989. Actualmente es catedrático y director del Depto. de Ciencias de la Computación de la Universidad de Chile (http://www.dcc.uchile.cl). También dirige el Centro de Investigación de la Web (http://www.ciw.cl), es presidente de Clei (http://www.clei.cl), miembro de la Academia de Ciencias de Chile, coordinador del subprograma de electrónica e informática de Cyted (http://www.cyted.org) y miembro del directorio de Ieee Computer Society (http://www.computer.org), entre otras actividades. Sus áreas principales de investigación son la recuperación de información, la minería de la web y el diseño y análisis de algoritmos y sus aplicaciones, tales como el buscador de la web chilena TodoCL (http://www.todocl.cl).

Palabras claves: *Minería de la Web, Análisis de enlaces, Análisis de contenido, Análisis de uso, Buscadores, Ubicuidad.*

Title: Web mining

Abstract: *The web is the internet's most important phenomenon, as demonstrated by its exponential growth and diversity. Hence, due to the volume and wealth of its data, search engines have become among the web's main tools. They are useful when we know what we are looking for. However, certainly the web holds answers to questions never imagined. The process of finding relations or interesting patterns within a data set is called "data mining" and in the case of the web, "web mining". In this article we present the main ideas behind web mining and some of its applications.*

Keywords: *Web mining, Link analysis, Content analysis, Usage mining, Search engines, Findability.*

Baeza-Yates, Ricardo. *"Excavando la web". En: El profesional de la información, 2004, enero-febrero, v. 13, n. 1, pp. 4-10.*

1. Introducción

La web tiene actualmente al menos unas cuatro mil millones de páginas estáticas[1] y un número cientos de veces mayor de dinámicas (aquellas que sólo se crean producto de un clic o de una consulta en un sitio web). Además, tenemos que agregar toda la web invisible, en intranets o páginas con acceso restringido. La web oculta es seguramente miles de veces más grande que la pública. La figura 1 muestra estas distinciones, minimizando el tamaño tanto de la parte dinámica como de la oculta. Una última región, la web semántica, se muestra en un tono más claro. En la actualidad, se estima que las páginas con información semantica constituyen algo menos del 5%, aunque se espera que en el futuro sea mayor. Esta información semántica, representada principalmente en los metadatos de cada página, no es muy usada ya que existe un porcentaje mayor de páginas que tienen información no fidedigna o directamente falsa (*spamming* de metadatos). Finalmente, la región con rayas paralelas indica la zona que efectivamente poseen los buscadores web, que se corresponde en gran parte con la zona pública estática y un poco de la dinámica.

En este artículo presentamos una introducción al tema de minería de la web, indicando los tipos principales: de contenido, de estructura y de uso, en particular los dos últimos así como sus aplicaciones. Las referencias bibliográficas incluidas son las más relevantes a nuestro juicio y sirven como punto de partida para ahondar en el tema. Por otra parte, el artículo se cen-

Artículo recibido el 31-10-03
Aceptación definitiva: 24-01-04

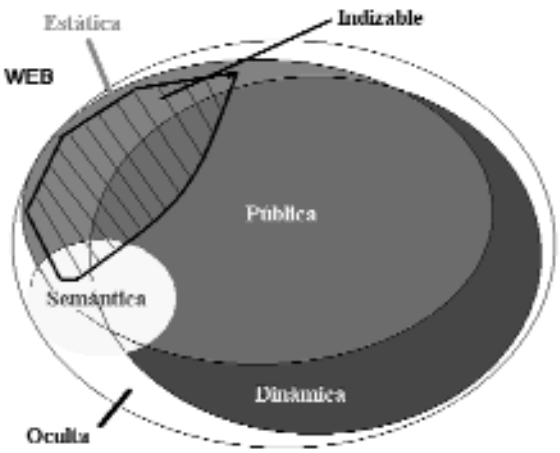

Figura 1: partes de la web

tra en resultados de nuestro propio trabajo de investigación, tanto en datos extraídos de buscadores web como en el rediseño de sitios web.

«Para conocer qué páginas apuntan a otra es necesario recorrer toda la web, algo que los grandes buscadores como Google o Alltheweb hacen periódicamente»

Comenzamos describiendo las principales características de la web, seguido de los distintos tipos de datos que pueden analizarse, detallando cada uno de ellos. Continuamos con el uso de estas técnicas en el desarrollo de sitios web y los requerimientos en su diseño para poder realizar minería de uso. Concluimos con el posible impacto futuro de esta nueva área de investigación.

2. Infometría de la web

Es el producto del trabajo colaborativo de millones de personas. Por ende, sus características representan su esfuerzo que, en la mayoría de los casos, es mínimo. **George Kipling Zipf**, un lingüista de Harvard, publicó su libro acerca de la ley del mínimo esfuerzo un año antes de su deceso (1939) a la prematura edad de 40 años. Su descubrimiento inicial fue que si uno contaba el número de veces que se usaba cada palabra en distintos textos en inglés, y las ordenaba de la más a la menos frecuente, se cumplía que la frecuencia *F* de la palabra *i*-ésima, multiplicada por *i*, era igual a una constante *C*, y la constante *C* dependía del texto escogido. Actualmente, es necesario elevar *i* a un exponente *t* mayor que 1 y cercano a 2 para muchos textos existentes, en particular de la web. Graficando esta curva mediante el uso de una escala logarítmica en ambos ejes, se convierte en una recta con pendiente negativa *t*.

Zipf prefirió explicar estos resultados empíricos como una condición humana, donde siempre es más fácil escribir una palabra conocida que usar una que lo es menos. Fenómenos similares aparecen en otros ámbitos como el número de citas bibliográficas a un artículo dado o las poblaciones de ciudades. Diversos autores, entre ellos **Mandelbrot** y **Miller**, argumentaron más tarde que, en realidad, la ley de **Zipf** representa la consecuencia de las leyes de las probabilidades en procesos asociados a variables ordenadas por frecuencia. Sin querer tomar partido en esta disputa científica, cierta o no, la ley de **Zipf** aparece frecuentemente en la práctica y refleja bien la actitud natural de minimizar el esfuerzo, exceptuando los casos extremos, que serían en el ejemplo inicial usar muy pocas o muchas palabras. Tal vez esta ley sólo explica la diversidad humana, la cual se inclina más por la pereza que por la erudición. De hecho, que *t* sea ahora alrededor de 1.8 para textos en inglés, indica un mayor sesgo en esa diversidad y una degradación en el tiempo de la riqueza del vocabulario que usamos al escribir en ese idioma.

La figura 2 muestra esta ley a la izquierda, mientras que a la derecha se muestra cómo crece el número de palabras distintas, es decir, el tamaño del vocabulario *V* usado en el texto. Esta es una ley experimental

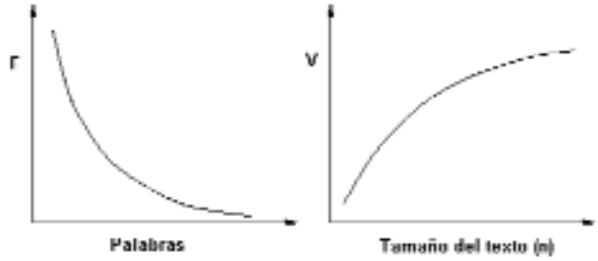

Figura 2: la ley de Zipf y la de Heaps

llamada de **Heaps**, que crece en forma sublineal con un exponente entre 0.4 y 0.7, siendo el límite superior el caso de la web. Más detalles se pueden encontrar en el capítulo 6 de **Baeza-Yates** y **Ribeiro-Neto** (1999).

Si hay algún fenómeno donde el principio del mínimo esfuerzo aparecería si existiera, es la web. Aparte de la distribución de palabras en la web, las siguientes medidas siguen una ley de **Zipf**:

—Tamaños de las páginas o de otros tipos de archivos (imágenes, audio, etc.). En este caso la ley no se ajusta bien al comienzo, porque hacer páginas con muy poco texto produce el pudor de la vergüenza que contrarresta al mínimo esfuerzo.

—Número de enlaces que salen de una página. La curva no se ajusta muy bien en los extremos, porque hacer una página con muy pocos vínculos cae en el caso

del punto anterior, y por otra parte, hay casos con muchos enlaces que son producidos de forma automática.

—Número de enlaces que llegan a una página. La mayoría de las páginas tienen sólo un enlace a ellas y hay pocas con muchos enlaces. Esto define una medida de popularidad de páginas web.

—Fecha de actualización. Existen más páginas nuevas o modificadas que viejas.

—Número de componentes conexos de distinto tamaño. Es decir, grupos de páginas en las que se puede navegar de cualquier página a otra. Esto representa en cierta medida el número de páginas de un sitio web: muchos sitios tienen pocas, pocos sitios tienen muchas.

—Uso de las palabras en las interrogaciones a los buscadores. El resultado es que la mayoría de las preguntas son muy simples y hay pocas complejas.

Lo anterior se extiende a otras medidas en internet, como tráfico en la Red, uso de proxies, etc. ¿Es todo esto una casualidad producto del azar o un fenómeno del comportamiento humano? La respuesta aún no es clara, pero la evidencia empírica no deja de sorprender. A continuación describiremos los distintos tipos de datos que existen en la web y clasificaremos los distintos casos de minería.

3. Minería de la web

Hay tres tipos de datos principales. El más importante y difícil de procesar es el contenido, que es multimedial, en el cual el texto juega un rol dominante. El segundo proviene de la estructura no lineal de la web: sus hiper-enlaces. Finalmente, el último procede del uso reflejado a través de los logs o bitácoras de los servidores Web[2] (ver por ejemplo Cooley et al, 1997). La figura 3 muestra los tres tipos de minería de la web, donde tanto las personas como agentes de software están involucrados en la generación o extracción de estos datos.

Estos datos pueden analizarse de forma estática o dinámica. En el primer caso se usan instantáneas de la web en un cierto momento. Sin embargo es más interesante analizar la dinámica de la web, es decir, sus cambios en el tiempo. Los datos pueden ser locales (un sitio web específico o de los sitios web de una institución) o globales (nos referimos a una fracción importante de la web, como un país completo u otra división de similar tamaño, ya sea cultural, temática, o política). En general, el análisis del uso es local, mientras que el de estructura es global. Otra distinción es que la minería puede ser genérica o específica a una aplicación web, por ejemplo un buscador de páginas.

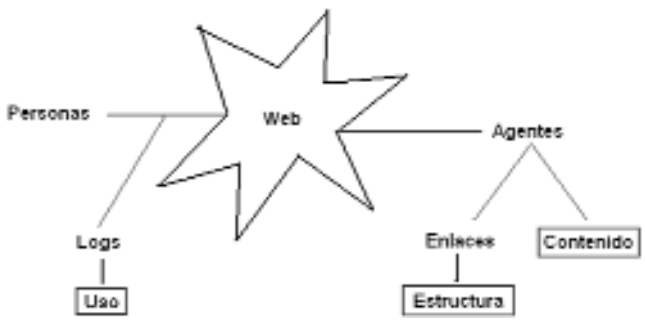

Figura 3: tipos de datos en la web

Los ejemplos de minería web de mayor impacto hoy en día con datos globales son del análisis de la estructura, en particular para encontrar las páginas más populares desde el punto de vista de la estructura de enlaces, técnicas usadas por buscadores globales.

4. Excavando el contenido

La forma más simple para recuperar información es a través de buscadores como *Google* o directorios como *Yahoo!*. Pero también es posible usar análisis de lenguaje natural para entender parcialmente la semántica del texto, extraer otros objetos como imágenes o audio, aprovechar las marcas de html para transformar el contenido o extraer datos específicos. Una aplicación puntual consiste en mejorar los resultados de los buscadores agrupando páginas similares. Uno de los problemas principales es cómo encontrar aquellas que poseen el contenido que necesitamos, pues sólo localizar las que son indizables ya es difícil, como mostramos en la figura 1. Las áreas de investigación relacionadas son minería de texto, minería multimedial, extracción de información, técnicas para resumir texto, traducción automática, etc. Para referencias en este tema ver el artículo de **Eíto Brun** y **Senso** (2004) a continuación.

5. Desenredando la estructura

La estructura de la web es compleja y evoluciona en el tiempo. Hay desde sectores altamente conectados hasta islas que sólo conocen algunos buscadores. La estructura puede ser usada por los buscadores para jerarquizar los resultados en base a las páginas más referenciadas utilizando heurísticas como *Pagerank* (**Brin**; **Page**, 1998) usado en *Google* o *Hits* (**Kleinberg**, 1998). También sirve para encontrar grupos de páginas que se apuntan entre sí y representan comunidades de personas con intereses similares. El problema principal en este caso es entender el proceso de evolución y su relación con las personas que participan en él. El uso y análisis de la estructura de la web se trata en un libro reciente de **Chakrabarti** (2002), mientras que de la parte dinámica es posible encontrar más información en un libro recién compilado por **Levene** y **Poulovassilis** (2004).

Para conocer qué páginas apuntan a otra es necesario recorrer toda la web, algo que los grandes busca-

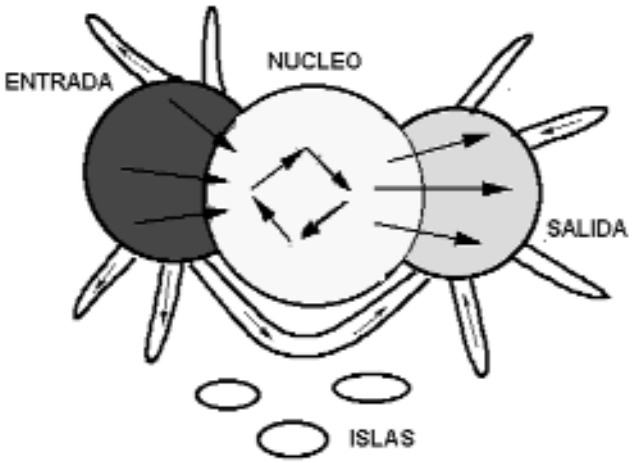

Figura 4: estructura macroscópica de la web

dores como *Google* o *Alltheweb* hacen periódicamente. Según dos recorridos de la web realizados por *Altavista* durante mayo y octubre de 1999, cada uno de más de 200 millones de páginas (alrededor de un 20% de la web de esa época) y 1.500 millones de enlaces. **Broder** (2000) presenta un estudio de la estructura macroscópica de la web, que es bastante intrincada y que resumimos a continuación.

Los resultados preliminares ya habían indicado que la distribución de los enlaces a y desde páginas seguían una ley de **Zipf**. Los nuevos resultados mostraron que la fracción de páginas de la web que son apuntadas por *i* páginas es proporcional a $1/i^{2.1}$, mientras que la fracción de páginas que tienen *i* enlaces es proporcional a $1/i^{2.7}$. Esto significa que el número de páginas muy apuntadas (populares) y la cantidad de páginas con muchos enlaces es muy pequeño. Estos valores son casi los mismos para los dos recorridos, pese a que entre ellos pasaron 6 meses.

Para analizar la estructura de la web se la modela como un grafo donde los nodos son páginas y los arcos son los enlaces entre páginas. Luego buscaron las partes del grafo que están conectadas entre sí. El núcleo o centro de la web lo constituían más de 56 millones de páginas, donde existe un camino para ir de cualquier página a otra, siendo el largo máximo (diámetro del núcleo) al menos 28. En otras palabras, el camino más corto entre dos páginas en el peor caso implica visitar 28 de ellas. Esto contrastó con el modelo del mundo pequeño que predecía un diámetro máximo de 20 páginas y una web completamente conexa. En este estudio se encontraron caminos de hasta largo 900, lo que indica que el diámetro es mucho mayor. De todos modos, este largo no es tan grande considerando que son cientos de millones de páginas.

«Posiblemente la aplicación más importante de minería se localiza en el diseño de sitios web»

La figura 4 muestra el resto de la estructura. A la izquierda había 43 millones de páginas desde las cuales se puede llegar al centro, pero no viceversa al revés. Del mismo modo, a la derecha existían otros 43 millones que pueden ser accedidas desde el centro pero que no enlazan páginas del núcleo. Alrededor de estos dos grupos encontraron tentáculos que contenían 44 millones de páginas y que son caminos sin salida, con la excepción de algunos tubos, que conectan el grupo de la izquierda con el de la derecha. Finalmente se encontraban 17 millones de páginas agrupadas en islas que no están conectadas al resto de la web. Muchos se preguntarán cómo *Altavista* conocía estas islas si no estaban conectadas al resto de la web. La explicación es simple: son sitios que fueron directamente enviados al buscador y por lo tanto están en su índice aunque el resto del mundo no los conozca. En la prác-

tica, para los buscadores sólo es fácil encontrar las páginas del núcleo y del componente de salida.

Los autores del estudio no hacen ninguna interpretación sobre esta estructura. Hemos realizado un análisis similar en la web chilena (**Baeza-Yates**; **Poblete**; **Saint-Jean**, 2003) y hemos encontrado que el grupo de la izquierda son páginas más nuevas que aún no son demasiado conocidas y que si tienen éxito pasarán al centro, donde están las más consolidadas. En cambio, en el grupo de la derecha se encuentran no sólo páginas antiguas, que no enlazan al centro de la web porque en su época no existían, sino también páginas mal hechas o mantenidas o que, por razones comerciales, no tienen enlaces externos. Los tentáculos son variaciones sobre el tema, incluyendo sitios web que no enlazan a nadie fuera de su sitio y revelan la complejidad dinámica de la web. Por otra parte hemos comprobado que las islas (sitios que no apuntan a nadie y que tampoco son apuntados) son más del 50% de los sitios en Chile y que, como son difíciles de encontrar, en general no están en los buscadores.

Recientemente también hemos analizado la evolución de la composición de la estructura de la web chilena, encontrando que un porcentaje alto de sitios desaparece en un año (cerca de un 20%) y que los sitios migran de un componente de la estructura a otro, siendo el núcleo de la web el componente más estable; por otra parte las islas es el componente más inestable (**Baeza-Yates**; **Poblete**, 2003). Éste es un ejemplo de minería de la parte dinámica de la estructura.

El análisis de estructura también sirve para sitios web, por ejemplo para encontrar enlaces no existentes o páginas desconectadas o para intranets. De hecho, un estudio reciente en todos los sitios web de una gran empresa (**Fagin** et al., 2003) mostró una estructura similar a la web global y permite mejorar las búsquedas internas. También es posible comparar las estructuras de un sitio Web en el tiempo para controlar cambios o entre sitios Web distintos para analizar duplicaciones o diferencias.

6. Analizando el uso

Analizar los archivos de acceso o bitácoras (logs) de un servidor web es de gran interés desde el punto de

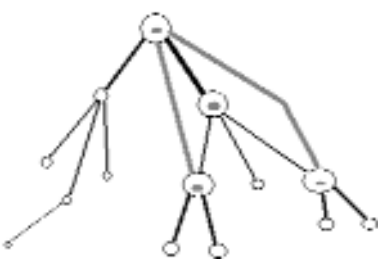

Figura 5: reorganización de enlaces en base a análisis de visitas al sitio

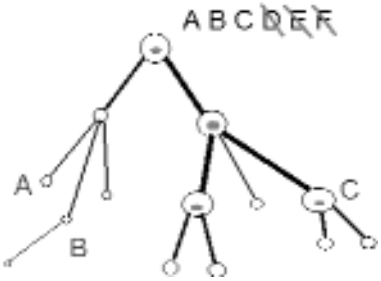

Figura 6: análisis de consultas en un sitio web

vista comercial (Srinivasan et al, 2000). Si una página nunca es visitada tal vez no tiene razón de ser o, por el contrario, si las muy visitadas no están en los primeros niveles de jerarquía del sitio web, esto sugiere mejorar su organización y navegación. Por lo tanto, es importante detectar patrones de acceso y sus tendencias. Esta detección puede ser genérica o para un usuario específico (lo que permite una personalización de forma dinámica) y los resultados pueden ser usados para recomendar servicios o productos. El problema principal en este caso es poder diferenciar a los usuarios y cuándo se conectan o desconectan (determinar sesiones). La figura 5 muestra un sitio donde dos páginas de tercer nivel son muy visitadas. Después de colocar enlaces directos a ellas (en tono más claro), las visitas a la página intermedia de segundo nivel disminuyen debido a los nuevos caminos directos. Estos enlaces se denominan *hotlinks*.

Un tipo de análisis de uso mucho más interesante y aún poco estudiado es el de consultas en el buscador del sitio web. Por supuesto esto sólo es posible si existe tal herramienta. En la figura 6 se muestran las 3 palabras más buscadas y que son encontradas en el sitio indicando dónde (A, B y C) y las tres más buscadas que no son encontradas en el sitio (D, E y F). Esta figura también incluye el análisis de visitas usando distinto grosor para los enlaces. A continuación explicamos cada caso:

—A, B: son palabras que la gente busca mucho, pero que no son asociadas por las personas al camino que existe desde la página principal. Esto quiere decir que las palabras que identifican estos caminos están mal escogidas y debieran ser reemplazadas por A y B donde corresponda.

—C: es muy buscada pero la gente llega a la página donde se encuentra. Esto indica que C es tan buena como la palabra usada en la página principal para llegar hasta donde se encuentra C. En este caso sería recomendable agregar también C en la página principal donde corresponda.

—D: es la más buscada que no se encuentra y que significa lo mismo que una palabra ya usada en el si-

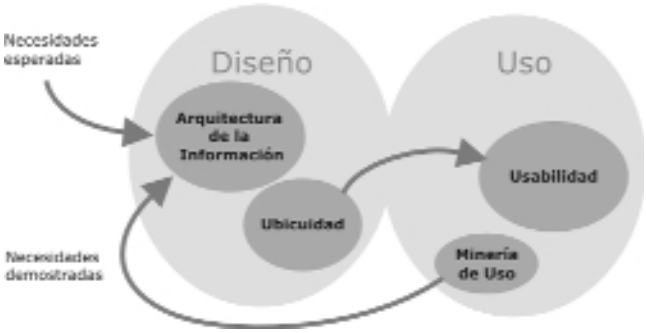

Figura 7: causalidad en el diseño y uso de sitios web

tio, por ejemplo A. Por lo tanto D debiera ser agregado al texto que indica el camino que representa a A.

—E: no está relacionada con el sitio y se busca por error.

—F: relacionada con el sitio, pero que su significado no está incluido en él. Esto puede ser un contenido o producto importante que podría ser un nuevo nicho de mercado.

La discusión anterior tiene que ver con lo que se llama esencia de la información (*information scent*), un término introducido por **Pirolli** (1997). Todas las palabras tienen alguna esencia, pero entre las que significan lo mismo, es decir polisémicas, hay algunas con más esencia que otras. Por ejemplo, para un banco ¿es mejor usar crédito o préstamo? Las palabras más consultadas son las de mayor esencia y dependen del lenguaje, el país, la cultura, la edad, entre otros factores. Un ejemplo de análisis de consultas en un buscador es presentado por **Baeza-Yates** y **Saint-Jean** (2003) y una recopilación de resultados se encuentra en **Baeza-Yates** (2004) incluyendo el uso de minería de consultas para mejorar el desempeño (índices) y los resultados de buscadores web (ranking). Otras aplicaciones son la agrupación de consultas similares para recomendar consultas similares y para mejorar los resultados.

7. Impacto en diseño de sitios web

Posiblemente la aplicación más importante de minería se localiza en el diseño de sitios web. De hecho, los dos ejemplos de uso mencionados en la sección anterior se refieren a esto. Sin embargo, para poder hacer minería de uso es necesario que la gente llegue al sitio, lo que implica el mantenimiento de un conjunto de requerimientos.

La figura 7 muestra los pasos causales del diseño de un sitio web que tiene cuatro fases relevantes:

—Arquitectura de la información: diseño del contenido y forma del sitio, por ejemplo siguiendo las normas de **Morville** y **Rosenfeld** (2002).

—Ubicuidad: si el sitio está bien diseñado puede ser encontrado. La facilidad para encontrarlo la llamamos ubicuidad (que extiende el término *findability* en inglés no sólo a las personas, sino también a los buscadores) la cual es muy relevante ya que un gran porcentaje de visitas llega a través de los buscadores.

—Usabilidad: si lo encontramos podemos usarlo, aunque para ello primero debemos poder verlo. La satisfacción de la persona al interactuar con el sitio se mide a través de la usabilidad.

—Minería de uso: si es usado, tenemos información de uso que puede ser aprovechada como hemos descrito anteriormente.

A continuación explicamos un poco más la segunda fase, la menos conocida de todas, para que la última, que ya hemos explicado, sea posible.

Un gran porcentaje de las visitas a un sitio, en especial si es nuevo, proviene de un buscador. Por lo tanto es importante poder buscar y encontrar el que nos interesa. Esto significa responder a las siguientes preguntas:

—¿Encontrará mi sitio un buscador? La respuesta es no, si no se ha registrado o si no posee un enlace desde un sitio conocido (es decir, del núcleo de la web).

—¿Pongo trabas a los buscadores para entrar a mi sitio? La respuesta es sí, si usa mapas de imágenes, muchas instancias de *Flash*, *Javascript* u otros mecanismos que no son html y que no permiten extraer los enlaces hacia páginas internas.

—¿Tengo el texto correcto en mi página principal?, ¿puedo encontrar mi sitio imaginando qué palabras usarán mis clientes?, ¿tengo las palabras con mayor esencia?

—¿Queda mi sitio bien ubicado en una búsqueda? Si la respuesta es no, debemos conseguir enlaces desde otros sitios y mejorar los títulos y metadatos de nuestras páginas.

Mas detalles sobre ubicuidad y usabilidad, en particular los temas de visibilidad y accesibilidad son tratados por **Baeza-Yates** y **Rivera** (2003). Nuestro modelo causal de diseño de sitios web es detallado en **Baeza-Yates** y **Velasco** (2004).

8. Epílogo

La minería de la web está recién en sus albores y tiene múltiples derivaciones, en particular el análisis de consultas permite complementar la navegación con lo que la persona está buscando. Esto impactará no sólo en el diseño de sitios web, sino también el mundo del comercio electrónico y la publicidad en internet.

Agradecimientos

Deseamos agradecer las sugerencias de redacción de **Helena Fernández**, **Mari Carmen Marcos** y los revisores anónimos. También a todos mis coautores, los cuales han hecho posible llevar a cabo los trabajos de investigación que permiten la exposición de ejemplos en minería de la web.

Notas

1 Estimación del autor de acuerdo al contenido actual de Google y AlltheWeb.

2 Usaremos la palabra bitácora para los archivos que guardan la información de las personas que visitan y que es lo que hacen en un sitio Web.

Bibliografía

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier. *Modern information retrieval.* Addison-Wesley, England, 1999.
http://sunsite.dcc.uchile.cl/irbook/

Baeza-Yates, Ricardo; Saint-Jean, Felipe. "Análisis de consultas a un buscador y su aplicación a la jerarquización de páginas web". En: *BID*, 2003, junio, n. 10.
http://www2.ub.es/bid/consulta_articulos.php?fichero=10baeza.htm

Baeza-Yates, Ricardo; Poblete, Bárbara; Saint-Jean, Felipe. *Evolución de la web chilena: 2001-2002*. Centro de Investigación de la Web, Universidad de Chile, enero 2003.
http://www.ciw.cl/recursos/estudio2002/estudio2002html.html

Baeza-Yates, Ricardo; Rivera, Cuauhtémoc. "Ubicuidad y usabilidad en la web". En: *Revista de gerencia tecnológica informática*, 2003, julio, v. 1, n. 2.
http://www.cidlisuis.org/aedo/RGTIN2V1/RGTI_04.pdf

Baeza-Yates, Ricardo; **Poblete, Bárbara**. "Evolution of the composition of the chilean web structure". En: *Proceedings of the 1st Latin American web conference*, Ieee CS Press, 2003.
http://www.la-web.org/stamped/02_baeza-yates-poblete.pdf

Baeza-Yates, Ricardo; Velasco, Javier. "The user experience from design to use, and back: a causal model". En: *5th Asis information architecture summit*, 2004.

Baeza-Yates, Ricardo. "Web usage mining in search engines". En: **Scime, Anthony** (ed.). *Web mining and applications and techniques*. Idea Group Publishing, 2004 (por aparecer).

Brin, Sergey; Page, **Lawrence.** "The anatomy of a large-scale hypertextual web search engine". En: *Proceedings of the 7th International world wide web conference*, 1998.
http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

Broder, Andrei; Kumar, Ravi; Maghoul, Farzin; Raghavan, Prabhakar; Rajagopalan, Sridhar; Stata, Raymie; Tomkins, Andrew; Wiener, Janet. "Graph structure in the web: experiments and models". En: *Proceedings of the 9th International world wide web conference*, 2000, pp. 247-256**.**
http://www9.org/w9cdrom/160/160.html

Chakrabarti, Soumen. *Mining the web: discovering knowledge from hypertext data*. San Francisco: Morgan Kaufmann Publishers, 2002.

Cooley, R.; Mobasher, B.; Srivastava, J. "Web Mining: Information and Pattern discovery on the World Wide Web". En: *ICTAI*, 1997, pp. 558-567.

Eíto Brun, Ricardo; Senso, José A. "Minería textual". En: *El profesional de la información*, 2004, enero-febrero, v. 13, n. 1.

Fagin, Ronald; Kumar, Ravi; McCurley, Kevin; Novak, Jasmine; Sivakumar, D.; Tomlin, John; Williamson, David. "Searching the workplace web". En: *Proceedings of the 12th International world wide web conference*, 2003.
http://www2003.org/cdrom/papers/refereed/p641/xhtml/p641-mccurley.html

Kleinberg, Jon. "Authoritative sources in a hyperlinked environment". En: *Proceedings 9th ACM-Siam Symposium on discrete algorithms*, 1998. Versión extendida en: *Journal of the ACM*, 1999, n. 46, pp. 604-632.

Levene, Mark; Poulovassilis, Alexandra (eds.). *Web dynamics*. Springer Verlag, 2004 (por aparecer).

Pirolli, Peter. "Computational models of information scent-following in a very large browsable text collection". En: *Human factors in computing systems: proceedings of the CHI'97 conference*, 1996, pp. 3-10.

Rosenfeld, Lou; Morville, Peter. *Information architecture for the world wide web: designing large-scale web sites* (2a.ed.). O´Reilly, 2002.

Srivastava, J.; Cooley, R.; Deshpande, M.; Tan, P. "Web Usage Mining: discovery and applications of usage patterns from web data". En: *SIGKDD Exploration*, 2000, n. 1, pp. 12-23.

Zipf, George. *Human behaviour and the principle of minimal effort.* Cambridge: Addison-Wesley, 1949.

Ricardo Baeza-Yates, *Centro de Investigación de la Web, Dpto. de Ciencias de la Computación, Universidad de Chile, Blanco Encalada 2120, Santiago, Chile.*
rbaeza@dcc.uchile.cl