Tendencias en minería de datos de la Web

Por **Ricardo Baeza-Yates**

Resumen: *Panorámica general y tendencias de diferentes aspectos y aplicaciones de la minería de datos en internet, con referencia a la Web 2.0, el spam, análisis de búsquedas, redes sociales y la privacidad.*

Palabras clave: *Minería de datos, Minería Web, Análisis, Logs, Web 2.0, Redes sociales, Spam, Redes sociales, Privacidad.*

Title: Web data mining trends

Abstract: *Overview and trends of different aspects and applications of data mining on the Internet, in relation to Web 2.0, spam, analysis of searches, social networks and privacy.*

Keywords: *Data mining, Web mining, Analysis, Logs, Web 2.0, Social networking, Spam, Social networks, Privacy.*

Baeza-Yates, Ricardo. "Tendencias en minería de datos de la Web". *El profesional de la información*, 2009, enero-febrero, v. 18, n. 1, pp. 5-10.

DOI: 10.3145/epi.2009.ene.01

Ricardo Baeza-Yates *es Ph. D. en Computer Science (Univ. of Waterloo, Canadá, 1989), Magister en Ing. Eléctrica (1986) y Cs. de la Computación (1985), e Ingeniero Eléctrico de la Univ. de Chile. Vicepresidente de Yahoo! Research para Europa, Medio Oriente & Latinoamérica, basado actualmente en Barcelona. Además es profesor Icrea asociado a la Univ. Pompeu Fabra y director científico de la línea de gestión de información de la Fundación Barcelona Media. Sus áreas de investigación son recuperación de información, minería de la Web, algoritmos y visualización de información. Es co-autor de un libro en recuperación de información (Addison-Wesley, 1999), de un manual de referencia en algoritmos y estructuras de datos (Addison-Wesley, 1991) y co-editor de un libro en recuperación de la información (Prentice-Hall, 1992). Fue presidente del CLEI (Centro Latinoamericano de Estudios en Informática) y Coordinador Internacional del Programa Iberoamericano en Ciencia y Tecnología (Cyted) en las áreas de electrónica e informática aplicadas, 2000-04. En 2002 fundó en Chile el Centro de Investigación de la Web (www.ciw.cl), del cual fue su primer director, y fue la primera persona de su área científica en ser incorporada a la Academia de Ciencias de Chile en 2003. En el 2007 obtuvó la medalla J. W. Graham de la Univ. de Waterloo que se otorga a ex-alumnos por innovación en computación.*

Introducción

ESTE AÑO LA WEB CUMPLE VEINTE AÑOS y casi parece que ha estado con nosotros desde siempre. Hace sólo 16 años comenzó a hacerse popular y hoy hay más de 187 millones de servidores web según el estudio mensual de *Netcraft*.

http://news.netcraft.com/

La Web inicial era eminentemente estática, es decir ficheros que estaban en un servidor de un sitio web. Hoy tenemos una Web completamente dinámica donde las páginas se generan dependiendo de las interacciones que realizan las personas en todo el mundo, conectados en muchos casos a través de teléfonos móviles.

Esta Web en la práctica es infinita, pues podemos generar un número arbitrario de páginas en un calendario o distintas representaciones de todas las bases de datos detrás de todos los sitios de comercio digital del mundo. Hoy hay más 1.500 millones de personas que usan internet, el 40% de ellas en Asia y con tasas de penetración en la población cercanas al 75% en América del Norte. Todas ellas la usan para enviar mensajes, buscar información o participar en redes sociales, entre millares de otras aplicaciones disponibles.

La Web como tal no tendría mucho sentido sin los buscadores que permiten acceder en forma directa a los distintos sitios y páginas. En una Web casi infinita lo importante no es el volumen sino la calidad, y tiene varios submundos importantes, como la Web semántica, o la Web 2.0 de contenido generado en forma cooperativa y directa por los usuarios.

El diagrama de la figura 1 intenta mostrar todo esto, donde lo que cubren los buscadores son principalmente la Web estática y una parte de la dinámica. La Web oculta es aquella que no podemos ver (salvo sus propietarios, evidentemente) y por lo tanto depende de cada observador. Cada persona tiene acceso a una Web distinta, y lo mismo los buscadores.

A continuación explicaremos cómo aprovechar la Web como fuente de datos y las tendencias principales en el análisis de los mismos, esto último en mi opinión personal.

"La Web es un ecosistema de contenidos, estructuras de enlaces y datos de su uso"

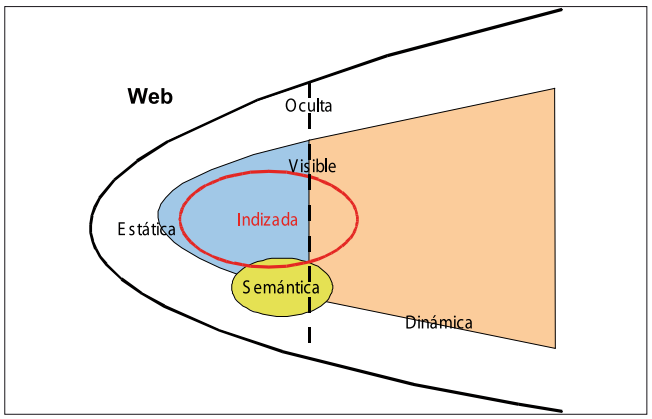

Figura 1. Segmentación de la Web

Minería de datos

La Web como ecosistema contiene y genera un universo de datos, tanto provenientes del propio contenido de sus páginas y la estructura de sus enlaces como de su uso por parte de las personas. Estos datos tienen una importancia crucial para el mejoramiento de la misma desde un punto de vista social y también comercial. Por esta razón la minería de datos de la Web ha crecido rápidamente y es una herramienta vital para entenderla y dar valor económico a los datos que obtenemos de ella[10]. Se distinguen tres tipos de minería Web:

– Minería de contenido: texto, imágenes, etiquetas (*tags*), metadatos, etc.;

– Minería de estructura: enlaces y sus relaciones; y

– Minería de uso: interacción de las personas con la Web.

Los dos primeros tipos de datos se obtienen recolectando todo el contenido de los sitios web usando software especial llamado recolector (*crawler*). Un recolector comienza con un conjunto de sitios iniciales y sigue todos los enlaces que encuentra en las páginas según un cierto conjunto de reglas predeterminadas (por ejemplo, qué dominios o tipos de ficheros recorrer). Los datos de uso provienen de los registros (*logs*) de los servidores web y de aplicaciones específicas, como las de búsqueda.

En cada uno de estos tipos de minería podemos mencionar decenas de ejemplos, pero los que son más interesantes incluyen todos ellos, es decir la Web completa. Los objetivos también son diversos, desde caracterización de la Web hasta problemas muy específicos de una aplicación dada. Un buen ejemplo de caracterización es el estudio de la Web española que realizamos en el 2005[2], que es el más completo a la fecha.

Con respecto a las técnicas usadas, sin duda la más popular es lo que se llama aprendizaje automático[8, 12, 17, 28]. Consiste en aprender como predecir variables en función de otras variables a través de subconjuntos de datos completos y luego evaluar cuán buena es la predicción en otro subconjunto de datos. El algoritmo resultante se usa en los datos reales con la suposición de que su desempeño será similar. Este proceso se repite en el tiempo para ir mejorando la herramienta con casos difíciles. Para esto se pueden utilizar árboles de decisión, máquinas de soporte vectorial o redes neuronales, entre otros.

"La participación de los usuarios en la Web 2.0 genera un círculo virtuoso que permite mejorar los contenidos y acercarnos a una búsqueda semántica real"

La Web 2.0

Sin duda uno de los temas más fascinantes es el contenido generado por los usuarios en la llamada Web 2.0. Esto incluye tanto recursos como la *Wikipedia* o el *Open Directory Project*, como blogs, *Yahoo! Answers* (respuestas de personas), *Flickr* (fotos), *YouTube* (vídeos) y *Del.icio.us* (favoritos), entre muchos otros. Uno de los casos más interesantes es *Flickr*, que permite etiquetar fotos con palabras o frases de manera libre, generando lo que se llama una folksonomía.

Entre los resultados recientes debemos mencionar el análisis de la calidad de las respuestas en *Yahoo! Answers*[1], la clasificación de etiquetas de *Flickr* usando *Wordnet* y *Wikipedia* que duplica la cobertura del vocabulario[21] y las anotaciones visuales de partes de una imagen y técnicas de reconocimiento de patrones para etiquetar imágenes sin etiquetas[19]. Estas técnicas generan un círculo virtuoso que permite mejorar el contenido de la Web usando los recursos de la Web 2.0 y también nos permite acercarnos a una búsqueda semántica real[6].

"El spam es el gran reto de la Web, y para combatirlo *Yahoo! Research* lidera el proyecto *Web Spam Challenge*"

Spam de Web

Los incentivos económicos en la Web están basados en el tráfico y la principal herramienta para generar tráfico son los buscadores. Esto ha dado lugar a lo que

se llama *spam* de la Web, es decir acciones que intentan obtener una ubicación mejor en los resultados de búsqueda que la que correspondería a la calidad intrínseca del sitio que la realiza. Ejemplos son poner texto falso o enlaces sin sentido en las páginas. Incluso hay spam de Web asociado a la interacción, por ejemplo clics realizados sobre la publicidad de la competencia (y que la competencia tiene que pagar a los buscadores) o sobre los resultados propios (para simular un aumento de tráfico).

El análisis de spam incluye resultados sólo basados en la estructura de la web (spam de enlaces[7, 9]) o en el contenido del texto (spam de contenido).

Para fomentar la investigación en esta área, *Yahoo! Research* con el apoyo de otras instituciones comenzó un desafío, el *Web Spam Challenge*[27], en el año 2006. Para esto se cuenta con más de una colección de webs del Reino Unido –donde se constata que el spam es cada año más complejo– y miles de juicios de personas sobre si un sitio web contiene spam o no (ver figura 2). Cada año el desafío es más difícil. En 2007 se logró un AUC de 0,96 que bajó al 0,85 en 2008, donde AUC (*area under curve*) es el área debajo de la curva de predicción de que un sitio sea spam.

Análisis de búsquedas

Las búsquedas que las personas hacen en el buscador de un sitio web da información importante sobre lo que quieren, qué cosas no encuentran o cuáles les faltan. Esta información tiene un gran valor comercial

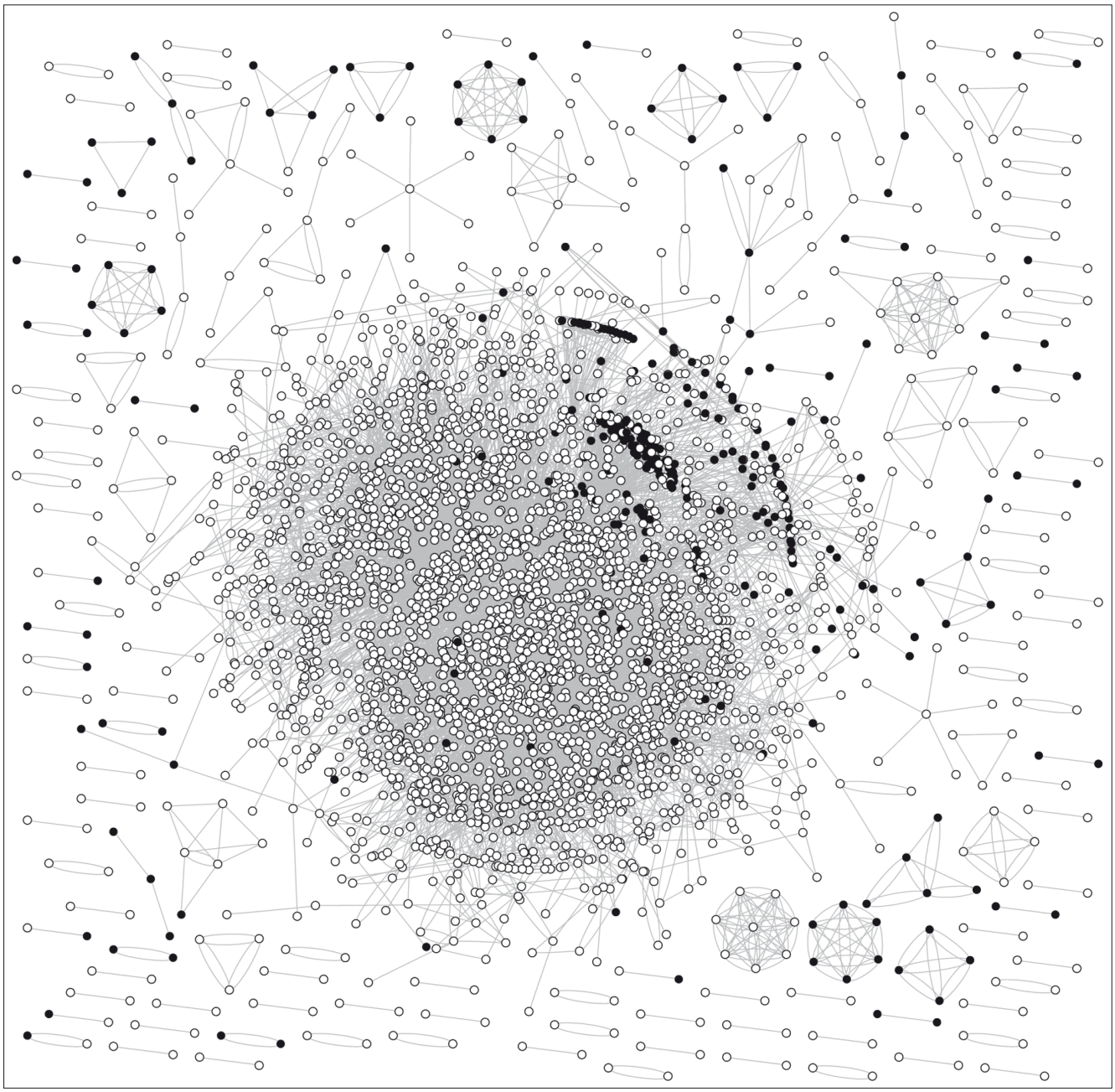

Figura 2. Ejemplo de la estructura de enlaces del Reino Unido. Los sitios en negro son los que hacen spam (fuente[9])

"Las búsquedas de información son minoritarias frente a la navegación o las transacciones (descargas, compras o reservas)"

pues permite no sólo optimizar el contenido actual de un sitio, sino también agregar nuevos contenidos[3, 21].

Uno de los trabajos seminales en el área describe un modelo para realizar análisis de consultas en un sitio y clasificar los distintos tipos de páginas en base a esto[3]. Este estudio también es importante para predecir la intención detrás de una consulta en un buscador y así modificar la presentación o el orden de los resultados[4]. Hoy en día las búsquedas relacionadas con información son un porcentaje minoritario, pues la mayoría buscan una web específica para navegar (consultas navegacionales) o para interactuar (consultas transaccionales), descargar ficheros, comprar objetos o reservar un pasaje. Por ejemplo es posible predecir para el 24% de las búsquedas –típicamente navegacionales– cual será la siguiente web que será explorada con una precisión superior al 90%[18].

Más aún, es posible encontrar relaciones semánticas entre consultas de un buscador, por ejemplo consultas equivalentes, en base a las páginas que la gente selecciona[5]. Un ejemplo de relaciones entre consultas sobre Chile se muestra en la figura 3.

Esto permite generar recursos semánticos, que pueden ser usados en otras aplicaciones, el volumen de los cuales es mucho más grande que la Web 2.0 ya que se hacen cientos de millones de consultas al día.

"En 2008 las redes sociales han tenido un gran auge en Latinoamérica, en particular en Chile, Colombia y Argentina"

Redes sociales

Las redes sociales explícitas como *Facebook* o *MySpace* ya relacionan a cientos de millones de personas. En 2008 han tenido un gran auge en Latinoamérica con la versión en castellano de *Facebook*, en particular en Chile y Colombia y recientemente en Argentina. Las redes sociales permiten que personas influencien a personas y poder medir esa influencia es unos de los objetivos principales de este tipo de minería.

El primer escollo de las redes sociales es que la misma persona puede tener múltiples identidades. Así que predecir la influencia de una red social a otra sólo es posible parcialmente en aquellas identidades que son

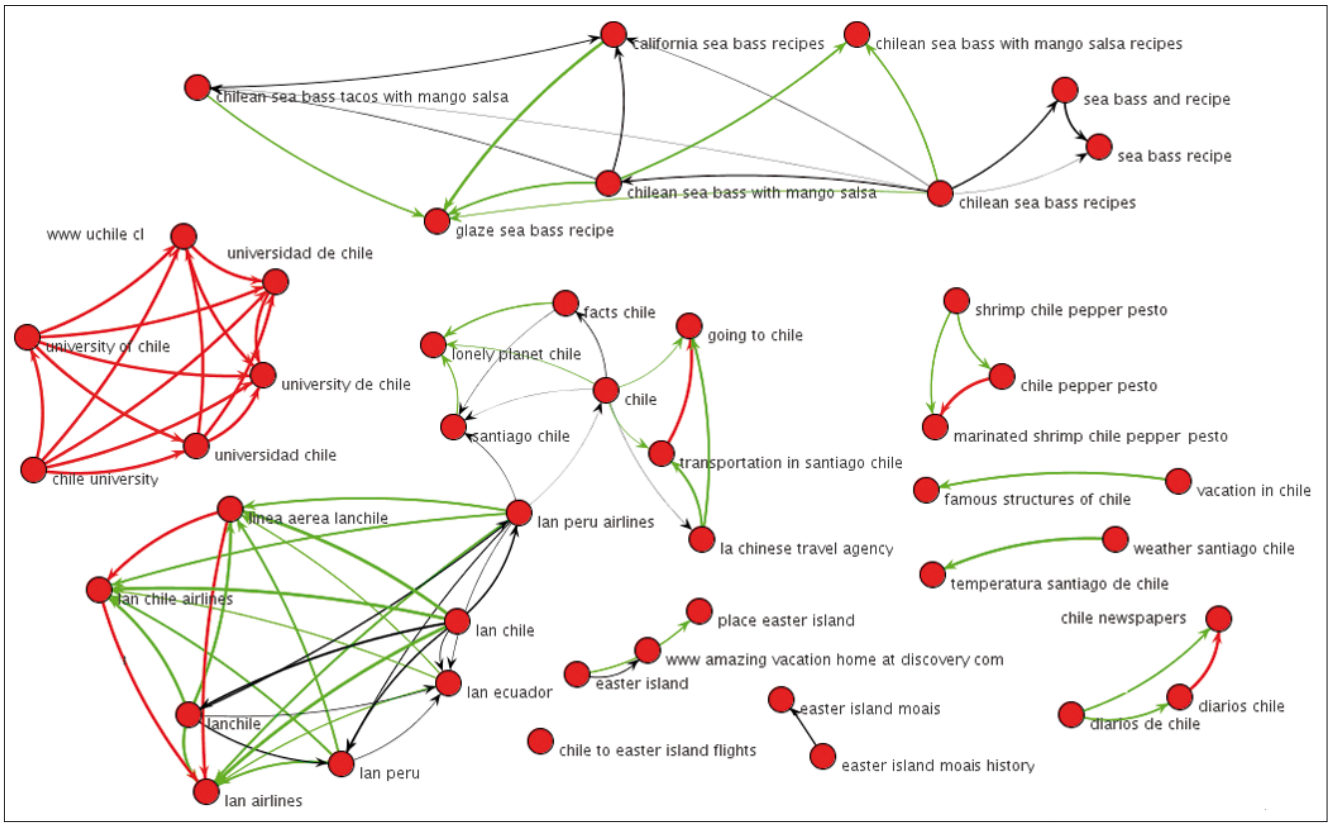

Figura 3. Relaciones semánticas en Chile. Los nodos unidos en rojo son preguntas equivalentes y los en verde son preguntas más específicas

las mismas, aunque en algunos casos podrían ser también falsas equivalencias (por esto es mejor usar la dirección de correo electrónico como identificador único en vez de un nombre de usuario que podría estar en dos redes sociales asociado a personas distintas)"

Otro problema importante en redes sociales es su evolución en el tiempo, conocer si las relaciones se mantienen, si hay comunidades y cómo es su dinámica, etc. Se han estudiado datos de redes de correo electrónico[16], llamadas entre teléfonos móviles[20] y mensajes de chat[13]. Sin embargo internet permite realizar también experimentos sociales a gran escala, algo que es nuevo en la sociología tradicional. Esta segunda línea de trabajo incluye trabajos aún no publicados de **Duncan Watts** de *Yahoo! Research*, como el análisis de encuestas en *Facebook* o de la cantidad de trabajo realizado por personas en el "Turco mecánico" de *Amazon http://www.mturk.com*

Entre los resultados más interesantes tenemos que la percepción personal de nuestros amigos puede ser diferente a la realidad, o que pagar más por unidad de trabajo hace que la gente trabaje más pero no mejor. También es importante el estudio de la influencia de una persona en una red social, ya sea para escoger música[23] o una película[14], y así encontrar a los líderes de cada grupo.

Los desafíos actuales son analizar datos de mayor volumen y que se extiendan más en el tiempo, para inventar una nueva ciencia social[15, 24].

"La minería de datos puede identificar personas concretas"

Privacidad

Uno de los temas más importantes relacionados con la minería de datos es su privacidad, y para mantenerla muchas veces se anonimizan identificadores de usuario, direcciones IP o cualquier otro dato que pueda identificar a una persona. Sin embargo con el uso de IPs dinámicas, distintas personalidades, computadores compartidos, etcétera, es difícil poder identificar a una persona, más aún cuando esos datos están distribuidos entre el proveedor de internet y el sitio web. Una técnica muy usada para preservar la privacidad es que los datos sean k-anónimos, es decir que no se pueda distinguir los datos de una persona de al menos otras (k-1) personas[26]. Por ejemplo, si k=10, habrá subconjuntos de datos de personas de al menos tamaño 10 que son iguales.

En algunos tipos de datos garantizar que sean k-anónimos no es trivial, cosa que pasó cuando *AOL* publicó en la Web un registro de su buscador que incluía consultas con sesiones anónimas[25]. Con estos datos un periodista identificó a una persona usando una sesión que contenía en las preguntas un código postal y un medicamento poco usual: cruzó esos datos con información pública de los hospitales correspondientes a esa zona. Por esta razón los buscadores han decidido no publicar este tipo de información y limitar el tiempo de almacenamiento de este tipo de datos (18 meses en el caso de *Google* y *Microsoft Live* y sólo 13 meses en *Yahoo!*) y guardarlos usando técnicas de anonimización más poderosas. Un buen resumen reciente de estos problemas fue presentado por **Cooper**[11].

Epílogo

Las tendencias en minería de datos son las de la misma Web. Estamos viendo sólo su comienzo, y queda mucho por hacer. Existe una gran diversidad de datos, en conjuntos cada día más voluminosos, y que abarcan periodos de tiempo más largos. En cada uno de ellos hay innumerables preguntas a responder y casos extraños a encontrar. Estas preguntas pueden ser desde medidas específicas hasta modelos para el comportamiento de millones de personas.

Bibliografía

1. **Agichtein, Eugene**; **Castillo, Carlos**; **Donato, Debora**; **Gionis, Aristides**; **Mishne, Gilad**. "Finding high-quality content in social media". *WSDM 2008*, 2008, pp. 183-194.

2. **Baeza-Yates, Ricardo**; **Castillo, Carlos**; **López, Vicente**. *Estudio de la Web española*, 2005. *http://www.catedratelefónica.upf.edu*

3. **Baeza-Yates, Ricardo**; **Poblete, Bárbara**. "A website mining model centered on user queries". En: *Semantics, Web and mining*, Ackermann, M. et al. (eds.), Springer LNAI 4289, 2006, pp. 1–17.

4. **Baeza-Yates, Ricardo**; **Calderón, Liliana**; **González, Cristina**. The intention behind web queries. *Spire 2006*, 2006, LNCS Springer, Glasgow, Scotland.

5. **Baeza-Yates, Ricardo**; **Tiberi, Alessandro**. Extracting semantic relations from query logs. *ACM KDD 2007*, 2007, San Jose, California, EUA, pp. 76-85.

6. **Baeza-Yates, Ricardo**; **Ciaramita, Massi**; **Mika, Peter**; **Zaragoza Hugo**. Towards semantic search. *NLDB 2008*, E. Kapetanios, V. Sugumaran y M. Spiliopoulou (eds.). Springer LNCS 5039, 2008, London, UK, pp. 4-11.

7. **Becchetti, Luca**; **Castillo, Carlos**; **Donato, Debora**; **Baeza-Yates, Ricardo**; **Leonardi, Stefano**. "Link analysis for web spam detection". *ACM transactions on the Web*, 2008, v. 2, n. 1.

8. **Bishop, Christopher M.** *Pattern recognition and machine learning*, Springer, 2007.

9. **Castillo, Carlos**; **Donato, Debora**; **Gionis, Aristides**; **Murdock, Vanessa**; **Silvestri, Fabrizio**. "Know your neighbors: web spam detection using the web topology". En: *Proc. of Sigir*, 2007, pp. 423-430. Amsterdam: ACM Press).

10. **Chakrabarti, Soumen**. *Mining the Web: discovering knowledge from hypertext data*. Morgan-Kaufmann Publishers, 2002.

11. **Cooper, Alissa**. "A survey of query log privacy-enhancing techniques from a policy perspective". *ACM transactions on the Web*, 2008, v. 2, n. 4, pp. 1-27.

12. **Duda, Richard O.**; **Hart, Peter E.**; **Stork, David G.** *Pattern classification* (2nd edition), 2001, New York: Wiley.

13. **Golder, Scott A.**; **Wilkinson, Dennis**; **Huberman, Bernardo**. "Rhythms of social interaction: messaging within a massive online network". *Proc. of Third international conference on communities and technologies*. C. Steinfield, B. Pentland, M. Ackerman, & N. Contractor (Eds.), 2007, Springer, pp. 41-66.

14. **Goyal, Amit**; **Bonchi, Francesco**; **Lakshmanan, Laks V. S.** "Discovering leaders from community actions". *Proc. of ACM 17th Conference on information and knowledge management* (CIKM), 2008, Napa Valley, California, EUA.

15. **Hedström, Peter**. "Experimental macro sociology: predicting the next best seller". *Science*, 2006, n. 331, pp. 786-787.

16. **Kossinets, Gueorgui**; **Watts, Duncan J.** "Empirical analysis of an evolving social network". *Science*, 2006, n. 311, pp. 88-90.

17. **Mitchell, Tom**. *Machine learning*, McGraw Hill, 1997.

18. **Piwowarski, Benjamin**; **Zaragoza, Hugo**. "Predictive user click models based on click-through history". En: *Proc. of the Sixteenth Conference on information and knowledge management* (CIKM 2007), 2007, Lisboa, Portugal, pp. 175–182.

19. **Olivares, Ximena**; **Ciaramita, Massi**; **Van Zwol, Roelof**. "Boosting image retrieval through aggregating search results based on visual annotations". *ACM Multimedia*, 2008, Vancouver, Canada.

20. **Onnela, Jukka-Pekka**; **Saramaki, Jari**; **Hyvonen, Jorkki**; **Szabó, Gabo**; **Lazer, David**; **Kaski, Kimmo**; **Kertesz, Janos**; **Barabási, Albert L**. "Structure and tie strengths in mobile communication networks". *Proc. of the National Academy of Sciences*, 2007, v. 104, n. 18, pp. 7332-7336.

21. **Overell, Simon**; **Sigurbjornsson, Borkur**; **Van Zwol, Roelof**. "Classifying tags using open content resources". En: *Second ACM international conference on web search and data mining* (*WSDM 2009*), 2009, Barcelona, Spain.

22. **Rosenfeld, Louis**; **Hurst, Marko**. *Search analytics: conversations with your customers*. Rosenfeld Media, por aparecer en 2009.

23. **Salganik, Matthew J.**; **Dodds, Peter S.**; **Watts, Duncan J.** !Experimental study of inequality and unpredictability in an artificial cultural market". *Science*, 2006, n. 331, pp. 854-856.

24. **Salganik, Matthew J.**; **Watts, Duncan J.** Leading the herd astray: an experimental approach to self-fulfilling prophecies in cultural markets". *Social psychology quarterly*, 2008, n. 71, pp. 338-355.

25. **Shen, Xuehua**. *Chronicle of AOL search query log release incident*, 2006. *http://sifaka.cs.uiuc.edu/xshen/aol_querylog.html*

26. **Sweeney, Latanya**. "k-Anonymity: a model for protecting privacy". *International journal of uncertainty, fuzziness, and knowledge-based systems*, 2002, v. 10, n. 5, pp. 557-570.

27. *Web spam challenge*, 2008. *http://webspam.lip6.fr/wiki/pmwiki.php*

28. **Witten, Ian H.**; **Frank, Eibe**. *Data mining: practical machine learning tools and techniques* (2nd edition). Morgan-Kaufmann, 2005.

***Ricardo Baeza-Yates**, Yahoo! Research*
ricardo.baeza@barcelonamedia.org
ricardo.baeza@upf.edu