novática
CEPIS UPGRADE Revista de la Asociación de Técnicos de Informática

Novática, revista fundada en 1975 y decana de la prensa informática española, es el órgano oficial de expresión y formación continua de **ATI** (Asociación de Técnicos de Informática), organización que edita también la revista **REICIS** (Revista Española de Innovación, Calidad e Ingeniería del Software). ***Novática*** edita asimismo **UP**GRADE, revista digital de **CEPIS** (*Council of European Professional Informatics Societies*), en lengua inglesa, y es miembro fundador de **UP**ENET (**UP**GRADE ***E**uropean **NET**work*).

<http://www.ati.es/novatica/>
<http://www.ati.es/reicis/>
<http://www.upgrade-cepis.org/>

ATI es miembro fundador de **CEPIS** (*Council of European Professional Informatics Societies*) y es representante de España en **IFIP** (*International Federation for Information Processing*); tiene un acuerdo de colaboración con **ACM** (*Association for Computing Machinery*), así como acuerdos de vinculación o colaboración con **AdaSpain**, **AI2, ASTIC, RITSI** e **Hispalinux**, junto a la que participa en **ProInnova**.

Coordinación Editorial, Redacción Central y Redacción ATI Madrid
Padilla 66, 3º, dcha., 28006 Madrid
Tlfn.914029391; fax.913093685 <novatica@ati.es>
Composición, Edición y Redacción ATI Valencia
Av. del Reino de Valencia 23, 46005 Valencia
Tlfn./fax 963330392 <secreval@ati.es>
Administración y Redacción ATI Cataluña
Via Laietana 46, ppal. 1ª, 08003 Barcelona
Tlfn.934125235; fax 934127713 <secregen@ati.es>
Redacción ATI Andalucía
Isaac Newton, s/n, Ed. Sadiel,
Isla Cartuja 41092 Sevilla, Tlfn./fax 954460779 <secreand@ati.es>
Redacción ATI Aragón
Lagasca 9, 3-B, 50006 Zaragoza.
Tlfn./fax 976235181 <secreara@ati.es>
Redacción ATI Asturias-Cantabria <gp-astucant@ati.es>
Redacción ATI Castilla-La Mancha <gp-clmancha@ati.es>
Suscripción y Ventas <http://www.ati.es/novatica/interes.html>, ATI Cataluña, ATI Madrid
Publicidad
Padilla 66, 3º, dcha., 28006 Madrid
Tlnf.914029391; fax.913093685 <novatica@ati.es>
Imprenta: Derra S.A., Juan de Austria 66, 08005 Barcelona.
Depósito legal: B 15.154-1975 -- ISSN: 0211-2124; CODEN NOVAEC
Portada: "Gaia gateway" / © Concha Arias Pérez
Diseño: Fernando Agresta / © ATI 2006

Nº 185, enero-febrero 2007, año XXXIII

sumario

editorial
El cambio generacional > 02

en resumen
Problemas complejos, respuestas inteligentes > 02
Llorenç Pagés Casas

monografía
Búsqueda en la Web del futuro
(*En colaboración con* UPGRADE)
Editores invitados: Ricardo Baeza-Yates, José María Gómez Hidalgo y Paolo Boldi

Presentación: buscando en la Web del futuro > 03
Ricardo Baeza-Yates, Paolo Boldi, José María Gómez Hidalgo

Búsqueda exploratoria: de encontrar a comprender > 05
Gary Marchionini

Aprendiendo a analizar textos en lenguaje natural > 10
Giuseppe Attardi

Snaket: un motor de agrupamiento de resultados de búsqueda > 16
Paolo Ferragina, Antonio Gullí

Naturaleza multimodal de la Web: nuevas tendencias en el acceso a la información > 23
Luis Alfonso Ureña López, Manuel Carlos Díaz Galiano, Arturo Montejo Raez, Mª Teresa Martín Valdivia

Recuperación de información con adversario en la Web > 29
Ricardo Baeza-Yates, Paolo Boldi, José María Gómez Hidalgo

GERINDO: Gestión y recuperación de información en grandes recopilaciones de documentos > 36
Nivio Ziviani, Alberto H. F. Laender Edleno Silva de Moura, Altigran Soares da Silva, Carlos A. Heuser, Wagner Meira Jr.

Líneas de investigación en Terrier: un motor de búsqueda para la recuperación avanzada en la Web > 43
Iadh Ounis, Christina Lioma, Craig Macdonald, Vassilis Plachouras

Yahoo! Research Barcelona: Recuperación de Información y Minería Web > 49
El Equipo de Investigación de Yahoo!

secciones técnicas
Lengua e Informática
Traducción automática y homosemantismo globalizador > 51
José Mayoralas García

Redes y servicios telemáticos
Las redes de comunicaciones ayer y hoy. Problemas a resolver para una red global > 55
Guillermo Ibáñez Fernández

Tecnologías y Empresa
Los Proyectos Europeos contribuyen al desarrollo del sector TIC > 61
Joan Batlle Montserrat

Referencias autorizadas > 64

sociedad de la información
Futuros emprendedores
SAPientes: la experiencia de participar en la FLL > 70
Alonso Alvarez García

Novática interactiva
La privacidad de los datos en Internet > 74
Foro de Debate

Programar es crear
El juego de los engranajes (CUPCAM 2006, problema B, enunciado) > 75
Manuel Abellanas Oar, Manuel Freire Morán

Polígonos en malla (CUPCAM 2006, problema A, solución) > 76
Dolores Lodares González, Angel Herranz Nieva

asuntos interiores
Coordinación editorial / Socios Institucionales > 77

Monografía del próximo número: "*Informática para deficientes visuales*"

Presentación: buscando en la Web del futuro

Ricardo Baeza-Yates[1], Paolo Boldi[2], José María Gómez Hidalgo[3]

[1]Yahoo! Research, Barcelona (España) y Santiago (Chile); [2]Università degli Studi di Milano, Milan (Italia); [3]Universidad Europea de Madrid (España)

<ricardo@baeza.cl>, <boldi@dsi.unimi.it>, <jmgomez@uem.es>

Editores invitados

Ricardo Baeza-Yates es el director de los nuevos laboratorios de investigación de Yahoo! en Barcelona y en Latinoamérica (Santiago, Chile). Previamente ha sido catedrático y director del Centro para la Investigación en la Web del Departamento de Informática de la Universidad de Chile, y Catedrático ICREA (*Institució Catalana de Recerca i Estudis Avançats*) en el departamento de Tecnología en la Universidad Pompeu Fabra en Barcelona. Ricardo es Doctor en Informática por la Universidad de Waterloo (Canada). Es coautor del libro *Modern Information Retrieval*, publicado en 1999 por Addison-Wesley, y también de la segunda edición del *Handbook of Algorithms and Data Structures* (Addison-Weesley, 1991). También fue coeditor del libro *Information Retrieval: Algorithms and data Structures* (Prentice-Hall, 1992). Es el primer científico informático elegido para la Academia de Ciencias de Chile, en 2003.

Paolo Boldi obtuvo su doctorado en informática en la Universidad de Milán, donde es actualmente profesor asociado en el Departamento de Ciencias de la Información. Sus intereses investigadores han tocado muy variados temas de la informática teórica y aplicada, tales como: la teoría de dominios, la teoría no clásica de la computabilidad, la computabilidad distribuida, las redes anónimas, el sentido de la dirección, y los sistemas auto-estables. Más recientemente, sus trabajos se han centrado en problemas relacionados con la World Wide Web, un campo de investigación en el que también ha aportado sistemas software utilizados por muchos otros especialistas en el tema. En particular, ha contribuido a escribir un motor de Recuperación de Información sobre texto altamente eficiente (MG4J), y una herramienta de compresión de grafos (WebGraph) que alcanza las tasas de compresión habituales en las herramientas actuales.

José María Gómez Hidalgo es Doctor en Matemáticas, y ha sido profesor e investigador en la Universidad Complutense de Madrid, y lo es en la Universidad Europea de Madrid desde hace 10 años, donde actualmente dirige el Departamento de Sistemas Informáticos. Sus principales intereses investigadores incluyen el Procesamiento del Lenguaje Natural y el Aprendizaje Automático, con aplicaciones al Acceso a la Información periodística y biomédica, y la Recuperación de Información con Adversario, con aplicaciones en el filtrado de correo basura y en la detección de pornografía en la Web. Ha participado en 10 proyectos de investigación, dirigiendo algunos de ellos. José María es coautor de múltiples artículos científicos centrados en los temas mencionados, que pueden accederse por medio de su página Web <http://www.esi.uem.es/~jmgomez/>. Es miembro del Comité de Programa del CEAS (*Conference on Email and Anti-Spam*) 2007, del Spam Symposium 2007 y de otras conferencias, y ha revisado artículos de JASIST (*Journal of the American Society for Information Science and Technology*), ECIR (*European Conference on Information Retrieval*) y otras. También es revisor de proyectos para la Comisión Europea.

Desde la edición de la monografía de ***Novática*** sobre "*Recuperación de la Información y la Web*" en junio de 2002, las dimensiones de la Web, los tipos de información que contiene y sus patrones de uso han evolucionado notablemente. Estos cambios plantean nuevos retos para sus puntos de entrada por excelencia, los motores de búsqueda, incluyendo:

■ **Eficiencia**. Desde sus inicios, los motores de búsqueda han sido diseñados para recuperar referencias Web ante consultas de usuarios en cuestión de milisegundos. Sin embargo, no es lo mismo gestionar unos millones de páginas Web, que recuperar con rapidez sobre centenares de miles de millones. Por ejemplo, el número de servidores Web se ha duplicado en los últimos 18 meses, de acuerdo con los informes de Netcraft. La cantidad de información en la Web crece con más rapidez que la potencia de los ordenadores, y es preciso rediseñar los algoritmos para seguir manteniéndola manejable.

■ **Web Semántica**. Los seres humanos son capaces de usar la Web para tareas tales como encontrar la palabra "automóvil" en holandés, reservar un libro en una biblioteca, o buscar el DVD más barato y comprarlo. Sin embargo, una computadora no puede realizar las mismas tareas sin que un humano las guíe, porque las páginas Web están diseñadas para ser leídas por personas y no por computadoras. La Web Semántica es la visión de una información comprensible por las computadoras, de modo que se pueda automatizar los aspectos más tediosos relacionados con encontrar, compartir y combinar información en la Web. Los elementos centrales de la Web Semántica son un modelo de datos denominado Marco de Descripción de Recursos (*Resource Description Framework*, RDF), una serie de formatos de de intercambio de datos (como RDF/XML, Turtle, etc.), y notaciones como el Esquema RDF (*RDF Schema*, RDFS) y el Lenguaje de Ontologías de la Web (*Web Ontology Language*, OWL), que facilitan la descripción formal de conceptos, términos y relaciones en un dominio determinado. La Web Semántica posibilitará nuevas formas de búsqueda, más simples y efectivas que las actuales, que deben construirse sobre el procesamiento inteligente de la información actual en la Web, incluyendo el análisis del lenguaje y de los datos multimedia.

■ **Redes Sociales interactivas**. Una de las razones más importantes del crecimiento de servidores y páginas Web es la enorme popularidad de los servicios de redes sociales interactivas, como Flikr, Blogger, Digg, MySpace, YouTube, la Wikipedia y muchos otros. Estos sitios permiten que sus usuarios publiquen y compartan fácilmente y con rapidez todos los tipos de información, incluyendo reflexiones personales, fotografías, vídeos, intereses y referencias, noticias, etc. La capacidad de compartir se posibilita especialmente a través del apoyo computacional a las relaciones interpersonales, con características como Amigo de un Amigo (*Friend of a Friend*, FOAF), que permiten a los usuarios compartir su red de relaciones personales. Los servicios de redes sociales dan apoyo a comunidades dinámicas interactivas emergentes, que toman decisiones sociales sobre la calidad de los contenidos Web. Estas comunidades pueden ser la clave de los motores de búsqueda del futuro, como el análisis de enlaces lo ha sido de la presente.

■ **Personalización y otras formas de contexto**. A medida que crece la potencia de las computadoras, ésta se puede aprovechar para implementar características más avanzadas en los buscadores. Explotar la información acerca del contexto del usuario (su ubicación física, las búsquedas previas y recientes, los clicks anteriores, etc.) ofrece la oportunidad de entregar información más precisa al mismo, en tanto que puede adaptarse a sus objetivos e intereses de búsqueda a corto y largo plazo. Tener presente el contexto del usuario es también crítico en los anuncios en la Web, un ámbito de importancia siempre creciente que permite sacar partido de la información de usuario para realizar segmentaciones más efectivas.

■ **Multimedia y multilingüe**. La Web es una comunidad con nacionalidades variadas que se expresan en diferentes idiomas, a los que los motores de búsqueda dan aún un soporte muy limitado. Incluso los aspectos más básicos de la internacionalización (como la elección de la codificación y del juego de caracteres) se cubren en los motores actuales de una manera parcial e insatisfactoria. Tan sólo los crecientemente efectivos sistemas de traducción dan soporte a la multiplicidad de lenguas, y los usuarios siguen demandando características translingües que les permitan superar las barreras del idioma,

recuperando documentos en múltiples lenguas ante consultas en su idioma materno. La potencia computacional y la calidad de los algoritmos de análisis multimedia también posibilitan mejores índices e interfaces, en los que los usuarios plantean consultas en forma de ejemplos de fotos o incluso videos, con el objetivo de obtener resultados multimedia.

■**Spam Web**. El que posiblemente sea el mayor valor de la Web, esto es, su capacidad para conectar desde fragmentos de información a personas, está siendo objeto de abuso de una manera creciente. Como ocurrió con el correo basura o *spam*, algunos proveedores de contenidos abusan de este valioso medio de comunicación, con el fin de obtener ventajas comerciales. Para ello, confeccionan páginas y preparan enlaces con el fin de obtener un posicionamiento privilegiado de manera inmerecida ante las consultas más populares de los usuarios. Más aún, violan la seguridad de sitios Web dinámicos (foros, redes sociales, etc.) con el objeto de insertar contenidos falseados y referencias a sus sitios Web. De este modo logran, en último término, dirigir el tráfico de usuarios a sus sitios Web, y el dinero a sus bolsillos, a veces disfrazando esta práctica como optimización de motores de búsqueda (*Search Engine Optimization*, SEO). Los operadores de motores de búsqueda, redes sociales, etc. se enfrentan a la exigencia de detener, o al menos reducir, este tipo de abuso.

Los autores invitados a esta monografía son investigadores de primera línea y representantes de la industria de los buscadores, y sus artículos cubren la mayoría de los aspectos anteriores, ofreciendo al lector una valiosa introducción a las técnicas y funcionalidades de los motores de búsqueda actuales y futuros.

El trabajo de ***Gary Marchionini*** presenta algunos de los modos de búsqueda avanzados a los que es preciso dar soporte en las herramientas de búsqueda Web actuales, con ejemplos concretos de cómo se viene realizando en proyectos como el Open Video Digital Library.

Giuseppe Attardi describe algunas de las técnicas de análisis del lenguaje natural que se encuentran en el centro de las aplicaciones más avanzadas de la Web Semántica. El análisis del lenguaje permite construir, mantener y explotar los recursos necesarios realizar la visión de la Web Semántica (especialmente las ontologías).

La personalización se cubre en esta monografía con el trabajo de ***Paolo Ferragina*** y ***Antonio Gullì***, que describen como obtener resultados más personalizados y precisos usando un motor de agrupamiento de resultados Web muy avanzado y efectivo, denominado Snaket.

Luis Alfonso Ureña López, ***Manuel Carlos Díaz Galiano***, ***Arturo Montejo Ráez*** y ***Mª Teresa Martín Valdivia*** presentan una serie de experimentos en recuperación multilingüe y multimedia (texto e imagen) basada en los contenidos, que dan soporte a nuevas formas de consulta en motores de búsqueda, con énfasis en la multimodalidad: consultas con tipos de información mixtos incluyendo texto e imágenes de ejemplo.

Ricardo Baeza-Yates, ***Paolo Boldi***, y ***José María Gómez Hidalgo*** han preparado una presentación de los problemas y soluciones actuales al spam Web y a otras formas de abuso, con énfasis en el análisis de enlaces y el filtrado de contenidos Web.

A continuación, se incluyen dos trabajos que representan esfuerzos de considerable magnitud y efectividad en el ámbito de la Recuperación de Información a gran escala y en la Web.

Por un lado, ***Nivio Ziviani***, ***Alberto H.F. Laender***, ***Edleno Silva de Moura***, ***Altigran Soares da Silva***, ***Carlos A. Heuser***, y ***Wagner Meira Jr.*** presentan una visión general de algunos de los resultados sobre búsqueda en la Web más relevantes de Gerindo, uno de los proyectos de mayores dimensiones y relevancia centrados en estos temas en los últimos años.

Por el otro, ***Iadh Ounis***, ***Christina Lioma***, ***Craig Macdonald***, y ***Vassilis Plachouras*** describen Terrier, una plataforma y motor de recuperación de altas prestaciones diseñado para permitir a los investigadores el desarrollo de nuevos modelos de recuperación, implementaciones eficientes, y la realización de investigaciones en muchos otros temas. Terrier puede ser fácilmente desplegado sobre colecciones de documentos a gran escala.

Finalizamos la monografía con una presentación de las actividades y líneas de investigación de Yahoo! Research, la única gran corporación dedicada a la búsqueda Web con laboratorios de investigación situados en España (concretamente, en Barcelona).

Referencias útiles sobre "Buscadores en la Web"

Además de las referencias y fuentes de información mencionadas en los artículos de esta monografía, los lectores interesados pueden examinar los siguientes libros, revistas científicas y actas de conferencias, y sitios web relacionados con el tema.

Libros:

■ **S. Abiteboul, P. Buneman, D. Suciu.** *Data on the Web: from Relations to Semistructured Data and XML*. Morgan Kauffman, 2000. ISBN:15586 0622X.

■ **M. Agosti, A. Smeaton (editors).** *Information Retrieval and Hypertext*. Kluwer, 1996. ISBN: 079239710X.

■ **R. Baeza-Yates, B. Ribeiro-Neto.** *Modern Information Retrieval*. Addison-Wesley, 1999. ISBN: 020139829X. Web site: http://sunsite.dcc.uchile.cl/irbook/.

■ **S. Chakrabarti.** *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann, 2003.

■ **D.A. Grossman, O. Frieder.** *Information Retrieval: Algorithms and Heuristics*. Springer, 2004. ISBN: 1402030045.

■ **Witten, A. Moffat, T. Bell.** *Managing Gigabytes*. Morgan Kauffman, 1999 (second edition). ISBN: 1558605703.

Revistas científicas:

■ **ACM Transactions on Information Systems,** <http://www.acm.org/pubs/tois/>.

■ **ACM Transactions on Internet Technology**, <http://www.acm.org/pubs/periodicals/toit/>.

■ **European Journal of Information Systems**, <http://www.palgrave-journals. com/ejis/index. html>.

■ **Electronic Library,** <http://www.emeraldinsight.com/info/journals/el/el.jsp>.

■ **IEEE Intelligent Systems,** < http://www.computer.org/portal/site/intelligent/>.

■ **IEEE Internet Computing,** < http://www.computer.org/portal/site/internet/>.

■ **IEEE Transactions on Information Theory,** <http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?puNumber=18>.

■ **IEEE Transactions on Knowledge and Data Engineering,** <www.computer.org/mc/tkde>.

■ **Information Processing & Management,** < http://ees.elsevier.com/ipm/>.

■ **Information Retrieval Journal,** <http://ees.elsevier.com/ipm/>.

■ **Journal of the Association for Information Systems**, < http://jais.aisnet.org/>.

■ **SIGIR Forum, <**http://www.acm.org/sigs/sigir/forum/>.

■ **SIGWEB Newsletter, <**http://www. sigweb.org/>.

■ **VLDB Journal, <**http://www. informatik. uni-trier.de/~ley/db/journals/vldb/index. html**>.**

■ **World Wide Web, <**http://vlib.org**/>.**

Conferencias:

■ **ACM DocEng**, <http://www. document engineering.org/>.

■ **ACM JCDL**, <http://www.acm.org/jcdl/>.

■ **ACM SIGIR,** <http://www.acm.org/sigir/>.

■ **CIKM,** <http://www.cs.umbc.edu/cikm/>.

■ **CLEF,** <http://www.clef-campaign.org/>.

■ **ECIR,** <http://irsg.bcs.org/ecir.php>.

■ **RIAO,** <http://www.riao.org/>.

■ **SPIRE,** <http://cn.net.au/>.

■ **TREC,** <http://trec.nist.gov/>.

Sitios Web:

■ **Centro de Investigación en la Web**, <http://www.ciw.cl>.

■ **Google Labs: laboratorio tecnológico de Google**, <http://labs.google.es>.

■ **Laboratorios de Investigación de Yahoo!**, <http://research.yahoo.com>.

■ **Página Personal de Ricardo Baeza-Yates**,<http://www.baeza.cl>.

■ **Página Personal de Paolo Boldi**, <http://boldi.dsi.unimi.it>.

■ **Página Personal de José María Gómez**, <http://www.esp.uem.es/jmgomez>.

■ **Programa de Investigación MAVIR**, <http://www.mavir.net>.

■ **Recursos sobre Recuperación en la Web**, <http://www.webir.org>.

■ **Search Engine Watch**, <http://www.searchenginewatch.com>.

■ **World Wide Web Consortium**, <http://w3c.org>.