

Novática, revista fundada en 1975, es el órgano oficial de expresión y formación continua de ATI (Asociación de Técnicos de Informática). Novática publica también *Upgrade*, revista digital de CEPIS (Council of European Professional Informatics Societies), en lengua inglesa.

NOVÁTICA

CEPIS **UPGRADE**

Revista de la Asociación de Técnicos de Informática

Edición digital

MAYO - JUNIO 2002

157

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

ATI es miembro de CEPIS y tiene un acuerdo de colaboración con ACM (Association for Computing Machinery). Tiene asimismo acuerdos de vinculación o colaboración con AdaSpain, Al² y ASTIC

CONSEJO EDITORIAL

Antoni Carbonell Noguera, Rafael Fernández Calvo, Francisco López Crespo, Julián Marcelo Cocho, Celestino Martín Alonso, Josep Molas i Bertrán, Roberto Moya Quiles, César Pérez Chirinos, Mario Piattini Velthuis, Fernando Píera Gómez (Presidente del Consejo), Miquel Sàrries Griñó, Carmen Ugarte García, Asunción Yturbe Herranz

Coordinación Editorial
Rafael Fernández Calvo <rfcavlo@ati.es>

Composición y autoedición
Jorge Llácer

Administración
Tomás Brunete, María José Fernández

SECCIONES TÉCNICAS: COORDINADORES

Arquitecturas
Jordi Tubella (DAC-UPC) <jordit@ac.upc.es>

Bases de Datos
Coral Calero Muñoz, Mario G. Piattini Velthuis (Escuela Superior de Informática, UCLM) <Coral.Calero@uclm.es>, <mpiattin@inf-cr.uclm.es>

Calidad del Software
Juan Carlos Granja (Universidad de Granada) <jcgranja@goliat.ugr.es>

Derecho y Tecnologías
Isabel Hernando Collazos (Fac. Derecho de Donostia, UPV) <ihernando@legaltek.net>

Enseñanza Universitaria de la Informática
Cristóbal Pareja Flores (Dep. Sistemas Informáticos y Programación-UCM) <cpareja@sip.ucm.es>

Informática Gráfica
Roberto Vivó (Eurographics, sección española) <rvivo@dsic.upv.es>

Ingeniería del Software
Luis Fernández (PRIS-E.I./UEM) <lufern@dpris.esi.uem.es>

Inteligencia Artificial
Federico Barber, Vicente Botti (DSIC-UPV) <fvbotti_fbarber@dsic.upv.es>

Interacción Persona-Computador
Julio Abascal González (FI-UPV) <julio@si.hu.es>

Internet
Alonso Álvarez García (TID) <alonso@ati.es>
Llorenç Pagés Casas (Atlante) <pages@ati.es>

Lengua e Informática
M. del Carmen Ugarte (IBM) <cugarte@ati.es>

Lenguajes informáticos
Andrés Marín López (Univ. Carlos III) <amarin@it.uc3m.es>
J. Ángel Velázquez (ESCET-URJC) <a.velazquez@escet.urjc.es>

Libertades e Informática
Alfonso Escolano (FIR-Univ. de La Laguna) <aescolan@ull.es>

Lingüística computacional
Xavier Gómez Guinovart (Univ. de Vigo) <xgomez@uvigo.es>
Manuel Palomar (Univ. de Alicante) <mpalomar@dlsi.ua.es>

Profesión informática
Rafael Fernández Calvo (ATI) <rfcavlo@ati.es>
Miquel Sàrries Griñó (Ayto. de Barcelona) <msarries@ati.es>

Seguridad
Javier Areatio (Redes y Sistemas, Bilbao) <jareatio@orion.deusto.es>

Sistemas de Tiempo Real
Alejandro Alonso, Juan Antonio de la Puente (DIT-UPM) <jaalonso.jp puente@dit.upm.es>

Software Libre
Jesús M. González Barahona, Pedro de las Heras Quirós (GSYC, URJC) <jgb.pheras@gsyc.es>, <gustavo@sol.info.unpl.edu.ar>

Tecnología de Objetos
Esperanza Marcos (URJC) <e.marcos@escet.urjc.es>

Tecnologías para la Educación
Benita Compostela (F. CC. PP. UCM) <benitu@dia.edunet.es>

Tecnologías y Empresa
Josep Sales Rufí (ESPIRAL) <jsales@pie.xtec.es>

TIC para la Sanidad
Pablo Hernández Medrano <phmedrano@terra.es>

TIC para la Sanidad
Valentín Masero Vargas (DI-UNEX) <vmasero@unex.es>

Las opiniones expresadas por los autores son responsabilidad exclusiva de los mismos. Novática permite la reproducción de todos los artículos, salvo los marcados con © o copyright, debiéndose en todo caso citar su procedencia y enviar a Novática un ejemplar de la publicación.

Coordinación Editorial y Redacción Central (ATI Madrid)
Padilla 66, 3º, dcha., 28006 Madrid
TIF: 914029391; fax: 913093685 <novatica@ati.es>

Composición, Edición y Redacción ATI Valencia
Palomino 14, 2º, 46003 Valencia
TIF/fax: 963918531 <secreval@ati.es>

Administración, Suscripciones y Redacción ATI Cataluña
Via Laietana 41, 1º, 1º, 08003 Barcelona
TIF: 934125235; fax: 934127713 <secregen@ati.es>

Redacción ATI Andalucía
Isaac Newton, s/n, Ed. Sadiel, Isla Cartuja 41092 Sevilla
TIF/fax: 954460779 <secreand@ati.es>

Redacción ATI Aragón
Lagasca 9, 3-B, 50006 Zaragoza
TIF/fax: 976235181 <secreara@ati.es>

Redacción ATI Asturias-Cantabria <gp-astucant@ati.es>

Redacción ATI Castilla-La Mancha <gp-clmancha@ati.es>

Redacción ATI Galicia
Recinto Ferial s/n, 36540 Silleda (Pontevedra)
TIF: 986581413; fax: 986580162 <secregal@ati.es>

Publicidad: Padilla 66, 3º, dcha., 28006 Madrid
TIF: 914029391; fax: 913093685 <novatica@ati.es>

Imprenta: 9-Impressió S.A., Juan de Austria 66, 08005 Barcelona.
Depósito Legal: B 15.154-1975
ISSN: 0211-2124; CODEN NOVAEC

Portada: Antonio Crespo Foix / © ATI 2002

SUMARIO

En resumen: Filtrando la avalancha <i>Rafael Fernández Calvo</i>	3
Monografía: «Recuperación de la información y la Web» (En colaboración con Upgrade , revista digital de CEPIS) Editores invitados: <i>Ricardo Baeza Yates y Peter Schäuble</i>	
Presentación. Recuperación de información: una disciplina con tradición <i>Ricardo Baeza Yates, Peter Schäuble</i>	4
Recuperación de información de contenidos empresariales <i>Prabhakar Raghavan</i>	5
Recuperación de información en la Web: nuevos paradigmas <i>Jacques Savoy</i>	8
Un análisis de lenguajes de consulta para XML <i>Adelaida Delgado Domínguez, Ricardo Baeza Yates</i>	11
Recuperación de información distribuida de bibliotecas digitales vía Web utilizando agentes móviles <i>J. Alfredo Sánchez, Sandra Nava Muñoz, Lourdes Fernández Ramírez, Griselda Chevalier Dueñas</i>	21
Extracción automática de información con semántica de la Web <i>Rafael Corchuelo, José Luis Arjona, Antonio Ruiz</i>	27
Sistema para la compresión y recuperación de documentos estructurados <i>Joaquín Adiego, Pablo de la Fuente, Jesús Vegas y Miguel Villarroel</i>	34
Las campañas CLEF: evaluación de Sistemas de Recuperación de Información Multilingüe <i>Martin Braschler, Carol Peters</i>	41
La Web de España <i>Ricardo Baeza Yates</i>	45
Secciones Técnicas	
Enseñanza Universitaria de la Informática Computing Curricula 2001 <i>Carlos Gregorio Rodríguez, Ángel Herranz Nieva, Raquel Martínez Unanue</i>	47
Informática gráfica Tutorial sobre Detección de Colisiones en Informática Gráfica <i>Juan J. Jiménez Delgado, Rafael J. Segura Sánchez, Francisco R. Feito Higuera</i>	55
Interacción Persona-Computador Ocultos pero no ausentes: los ciegos y la Informática (I) <i>Víctor M. Maheux</i>	59
Referencias autorizadas	63
Sociedad de la Información	
Personal y transferible LSSICE - Proyecto de Ley de Servicios de la Sociedad de la Información y de Comercio Electrónico: una ley, siete riesgos <i>Ignacio Boixo Pérez-Holanda, Darío Álvarez Gutiérrez</i>	67
Programar es crear Crucigramas <i>25º Concurso Internacional de Programación de ACM (2001): problema C</i> ¡Queso!: solución <i>Manuel Carro, Pablo Sánchez, Julio Mariño</i>	72
Asuntos Interiores	
Coordinación editorial / Programación de Novática	76
Normas de publicación para autores / Socios Institucionales	77

Monografía del próximo número: «XML, eXtended Mark-up Language»

Recuperación de la información y la Web

Ricardo Baeza Yates¹, Peter Schäuble²
¹ Departamento de Ciencias de la Computación,
 Universidad de Chile; ² Eurospider, Zürich (Suiza)

<rbaeza@dcc.uchile.cl>, <Peter.Schauble@eurospider.com>

Traducción: Manuel Galindo (Socio de ATI)

La Recuperación de Información (IR, *Information Retrieval*) se asocia a menudo con motores de búsqueda en Internet. Sin embargo, deriva de una disciplina académica cuyas raíces datan de los 50. Durante su primera década las actividades de investigación normalmente tenían lugar en los departamentos de Ciencias de la Computación y las aproximaciones más simples se basaban en las estadísticas de ocurrencia, que tenían una efectividad sorprendente en la recuperación de documentos relevantes. No obstante, un pequeño número de grupos de investigación en Recuperación de Información consiguieron resultados importantes en tres aspectos:

1. *Teoría*. Se desarrollaron modelos de recuperación pro-babilística que implicaban una óptima eficacia de recuperación (ver publicaciones de Cooper, Robertson y otros). Más tarde, la recuperación se extendió a otros medios, no sólo texto.
2. *Sistemas*. Recientemente se han intentado varios algoritmos y estructuras de datos, e integrado sistemas de recuperación impracticables (p.e., SMART, Topic, y Sistema Inquiry), así como sistemas de recuperación multimedia.
3. *Evaluación*. Se han construido colecciones de pruebas consistentes en documentos, consultas y --más importante aún-- de aseveraciones de relevancia que determinan qué documentos son relevantes para qué consultas. Estas colecciones de pruebas facilitan la comparación de distintos métodos de recuperación en lo concerniente a llamada y precisión (p.e. colecciones Cranfield, SMART y TREC).

Con el crecimiento de Internet, estos sistemas de Recuperación de Información constituyeron bloques de construcción listos para ser usados. La gran cantidad de datos tanto como la federación del espacio abierto de Internet para nuevos y excitantes conceptos, como enlaces basados en ranking, recuperación XML, integración de fuentes de datos heterogéneas, etc. Algunos de estos conceptos son tratados por los autores de esta monografía de *Novática* y de *Upgrade* sobre Recuperación de Información y la Web, y se utilizan parcialmente a continuación para presentar los artículos que la componen.

Punteros Editoriales

- «Estado del Arte»

Prabakhar Raghavan muestra algunas aplicaciones comerciales de las técnicas IR. También elabora las diferencias entre la recuperación en Internet y en Intranet. **Jacques Savoy** presenta una encuesta sobre recuperación de información en la Web, con énfasis en recuperación distribuida, ranking basado en enlaces y evaluación de motores de búsqueda.

Adelaida Delgado y **Ricardo Baeza Yates** presentan un análisis de lenguajes de consulta para XML, enfatizando en la propuesta del W3C, Xquery, desde la perspectiva de datos estructurados, así como de recuperación de texto.

- *Teoría y Sistemas*

Alfredo Sánchez, **Sandra Nava**, **Lourdes Fernández**, y **Griselda Chevalier** presentan un marco basado en agentes como soporte de recuperación de información distribuida desde bibliotecas digitales heterogéneas accesibles vía Web.

Rafael Corchuelo, **José Luis Arjona**, y **Miguel Toro** introduce otro marco basado en agentes para extracción automática de información semántica de páginas web.

Joaquín Adiego, **Pablo de la Fuente**, **Jesús Vegas**, y **Miguel Villarreal** presentan un sistema que usa índices invertidos comprimidos para recuperación de documentos considerando el contenido y la estructura de los documentos SGML o XML.

- *Evaluación*

Martin Braschler y **Carol Peters** describen el CLEF (*European Cross Language Evaluation*). Junto con el foro americano Trec y el japonés NTCIR, CLEF se

Notas del Editor de Novática:

1. Con objeto de incluir el máximo número posible de artículos en la monografía, se ha reducido el tamaño del cuerpo de letra.
2. A pesar de lo anterior, por razones de espacio no se incluyen en esta monografía los siguientes artículos: «*Ontologías en Federación de Bases de Datos*», de **Nieves R. Brisaboa**, **Miguel R. Penabad**, **Ángeles S. Places** y **Francisco J. Rodríguez**; «*Metodologías de desarrollo de Sistemas de Información en la Web y análisis comparativo*», de **M. José Escalona**, **Manuel Mejías** y **Jesús Torres**; y «*TEXTRET: un sistema interactivo de Recuperación de Texturas (TEXTURE RETrieval)*», de **Javier Ruiz del Solar**, **Pablo Navarrete** y **Patricio Parada**.

Dichos artículos serán publicados, en inglés, en el número 3/2002 de *Upgrade*, <<http://www.upgrade-cepis.org>>, y en próximos números de *Novática*, en castellano.

Presentación. Recuperación de información: una disciplina con tradición

encuentra entre los mayores puntos de encuentro para el avance de la tecnología de recuperación de información.

Ricardo Baeza Yates presenta un análisis de la Web española en comparación con la brasileña y la chilena. Los resultados y conclusiones deberían ser similares a los de otros países europeos.

A todos ellos expresamos nuestro agradecimiento por su valiosa colaboración, así como a los editores de *Novática* y *Upgrade* por su iniciativa.

Referencias útiles (lista elaborada por Ricardo Baeza Yates)

Además de las referencias que aparecen en los artículos de esta monografía, los lectores interesados pueden echar un vistazo a los siguientes libros, periódicos y actas de congresos, así como a muchos sitios web relativos a estándares web <<http://w3c.org>>, motores de búsqueda <<http://www.searchenginewatch.com>>, etc.

Libros

- Abiteboul, S., Buneman, P. & Suciu, D. *Data on the Web: from Relations to Semistructured Data and XML*, Morgan Kaufman, 2000.
- Agosti, M. & Smeaton, A. (editores) *Information Retrieval and Hypertext*, Kluwer, 1996.
- Baeza-Yates, R. & Ribeiro-Neto, B. *Modern Information Retrieval*, Addison-Wesley 1999. Sitio Web: <<http://sunsite.dcc.uchile.cl/irbook/>>
- Witten, I., Moffat, A. & Bell, T. *Managing Gigabytes*, Morgan Kaufman, 1999 (segunda edición).

Publicaciones periódicas

- Information Processing & Management
- Information Retrieval Journal
- ACM transactions in office information systems

Conferencias

- ACM SIGIR <<http://www.acm.org/sigir/>>
- JCDL <<http://www.acm.org/jcdl/>>
- CIKM <<http://www.cs.umbc.edu/cikm/>>
- SPIRE <<http://cn.net.au/>>
- TREC <<http://trec.nist.gov/>>
- CLEF <<http://clef.iei.pi.cnr.it:2002/>>
- NTCIR <<http://research.nii.ac.jp/~ntcadm/index-en.html>>

Editores invitados

Ricardo Baeza Yates es Ph. D. en Computer Science (Univ. of Waterloo, Canadá, 1989), Magister en Ingeniería Eléctrica (1986) y Ciencias de la Computación (1985) de la Universidad de Chile; e Ingeniero Civil Eléctrico de la misma universidad. Actualmente es Catedrático en el Depto. de Ciencias de la Computación de la Universidad de Chile y Director del Centro de Investigación de la Web <<http://www.ciw.cl>>. 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). Ha sido dos veces presidente de la Sociedad Chilena de Ciencia de la Computación y ha recibido premios de la Organización de Estados Americanos y el Instituto de Ingenieros de Chile. Actualmente, entre otros cargos, es presidente del CLEI (Centro Latinoamericano de Estudios en Informática), miembro del directorio de IEEE-CS y es coordinador internacional del subprograma de informática y electrónica aplicadas de CYTED (Programa de Cooperación Iberoamericana). Durante el año 2000 puso en marcha una empresa de Internet para buscar en la Web chilena <<http://www.todocl.cl>>. Su página personal está en <<http://www.dcc.uchile.cl/~rbaeza/spanish.html>>.

Peter Schäuble es CEO de la empresa Eurospider Information Technology AG, empresa suiza líder en recuperación de la información y que suministra software de monitorización de noticias y recuperación corporativa <<http://www.eurospider.com>>. Anteriormente fue Profesor Asistente de Ciencias de la Computación en el Instituto Federal Suizo de Tecnología (ETH, Zürich, Suiza) y dirigió el grupo de investigación de recuperación de la información. Es Licenciado en Matemáticas por el ETH y Doctor en Ciencias de la Computación por el mismo centro. Trabajó en el departamento técnico de la ESA (European Space Agency) y miembro invitado de los laboratorios de Hewlett-Packard en Palo Alto (California, EE.UU.). Ha publicado diversos artículos y libros de investigación sobre recuperación de la información.