

UPGRADE is the European Journal for the Informatics Professional, published bimonthly at <http://www.upgrade-cepis.org/>

Publisher

UPGRADE is published on behalf of CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by **Novática** (<http://www.ati.es/novatica/>), journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*, <http://www.ati.es/>)

UPGRADE monographs are also published in Spanish (full version printed; summary, abstracts and some articles online) by **Novática**

UPGRADE was created in October 2000 by CEPIS and was first published by **Novática** and **INFORMATIK/INFORMATIQUE**, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>)

UPGRADE is the anchor point for UPENET (UPGRADE European NETWORK), the network of CEPIS member societies' publications, that currently includes the following ones:

- **Informatik-Spektrum**, journal published by Springer Verlag on behalf of the CEPIS societies GI, Germany, and SI, Switzerland
- **ITNOW**, magazine published by Oxford University Press on behalf of the British CEPIS society BCS
- **Mondo Digitale**, digital journal from the Italian CEPIS society AICA
- **Novática**, journal from the Spanish CEPIS society ATI
- **OCG Journal**, journal from the Austrian CEPIS society OCG
- **Pliroforiki**, journal from the Cyprus CEPIS society CCS
- **Pro Dialog**, journal from the Polish CEPIS society PTI-PIPS

Editorial Team

Chief Editor: Llorenç Pagés-Casas, Spain, pages@ati.es

Associate Editors:

François Louis Nicolet, Switzerland, nicolet@acm.org

Roberto Carniel, Italy, rcarniel@dgf.uniud.it

Zakaria Maamar, Arab Emirates, Zakaria.Maamar@zu.ac.ae

Soraya Kouadri Mostéfaoui, Switzerland,

soraya.kouadrimostefaoui@gmail.com

Rafael Fernández Calvo, Spain, rfcervo@ati.es

Editorial Board

Prof. Wolfried Stucky, CEPIS Former President

Prof. Nello Scarabottolo, CEPIS Vice President

Fernando Píera Gómez and

Llorenç Pagés-Casas, ATI (Spain)

François Louis Nicolet, SI (Switzerland)

Roberto Carniel, ALSI - Tecnoteca (Italy)

UPENET Advisory Board

Hermann Engesser (Informatik-Spektrum, Germany and Switzerland)

Brian Runciman (ITNOW, United Kingdom)

Franco Filippazzi (Mondo Digitale, Italy)

Llorenç Pagés-Casas (Novática, Spain)

Veith Risak (OCG Journal, Austria)

Panicos Masouras (Pliroforiki, Cyprus)

Andrzej Marciniak (Pro Dialog, Poland)

Rafael Fernández Calvo (Coordination)

English Language Editors: Mike Andersson, Richard Butchart, David Cash, Arthur Cook, Tracey Darch, Laura Davies, Nick Dunn, Rodney Fennemore, Hilary Green, Roger Harris, Michael Hird, Jim Holder, Alasdair MacLeod, Pat Moody, Adam David Moss, Phil Parkin, Brian Robson

Cover page designed by Concha Arias Pérez

"Gaia gateway" / © ATI 2007

Layout Design: François Louis Nicolet

Composition: Jorge Llácer-Gil de Rames

Editorial correspondence: Llorenç Pagés-Casas pages@ati.es

Advertising correspondence: novatica@ati.es

UPGRADE Newsletter available at

<http://www.upgrade-cepis.org/pages/editinfo.html#newsletter>

Copyright

© Novática 2007 (for the monograph)

© CEPIS 2007 (for the sections UPENET and CEPIS News)

All rights reserved under otherwise stated. Abstracting is permitted with credit to the source. For copying, reprint, or republication permission, contact the Editorial Team

The opinions expressed by the authors are their exclusive responsibility

ISSN 1684-5285

Monograph of next issue (April 2007)

**"Information Technologies
for Visually Impaired People"**

(The full schedule of UPGRADE
is available at our website)

Monograph: Next Generation Web Search

(published jointly with Novática*)

Guest Editors: *Ricardo Baeza-Yates, José-María Gómez-Hidalgo, and Paolo Boldi*

- 2 Presentation. The Future of Web Search — *Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo*
- 5 Efficient Sparse Linear System Solution of the PageRank Problem — *Gianna M. Del Corso, Antonio Gullì, and Francesco Romani*
- 12 Learning to Analyze Natural Language Texts — *Giuseppe Attardi*
- 19 SNAKET: A Personalized Search-result Clustering Engine — *Paolo Ferragina and Antonio Gullì*
- 27 The Multimodal Nature of the Web: New Trends in Information Access — *Luis-Alfonso Ureña-López, Manuel-Carlos Díaz-Galiano, Arturo Montejo-Raez, and M^a Teresa Martín-Valdivia*
- 33 Adversarial Information Retrieval in the Web — *Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo*
- 41 GERINDO: Managing and Retrieving Information in Large Document Collections — *Nivio Ziviani, Alberto H. F. Laender, Edleno Silva de Moura, Altigran Soares da Silva, Carlos A. Heuser, and Wagner Meira Jr.*
- 49 Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web — *Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras*
- 57 Yahoo! Research Barcelona: Web Retrieval and Mining — *The Yahoo! Research Team*

UPENET (UPGRADE European NETWORK)

- 59 From **Novática** (ATI, Spain)
Informatics Profession
The Maturity of IT Professionalism in Europe — *Sean Brady*
- 68 From **Pro Dialog** (PTI-PIPS, Poland)
Graphical Interfaces
Portable Declarative Format for Specifying Graphical User Interfaces — *Zbigniew Fryźlewicz and Rafał Gierusz*
- 75 From **Novática** (ATI, Spain)
Next-generation Web
Blogs: On the Cutting Edge of the Next-generation Web — *Antonio Miguel Fumero-Reverón and Fernando Sáez-Vacas*

CEPIS NEWS

- 83 Harmonise Project: Building up to the Final Report—*François-Philippe Dragnet*
- 84 News & Events: European Funded Projects and News Updates

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>.

Presentation

The Future of Web Search

Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo

Since the publication of the UPGRADE issue on "*Information Retrieval and the Web*" on June, 2002, the dimension of the Web, and the kind of information on the Web and its usage, have clearly evolved, posing new challenges for their most prominent entry points, Search Engines. Among such challenges are:

1. Advanced search modes. Text data retrieval (such as what is the capital of France) is one of the most popular search activities on the Web. However, there are other search activities with more ambitious and sophisticated goals, such as searching to learn or to investigate. As more and more users access the Web, it is increasingly necessary to provide support to ever more sophisticated search strategies.

2. Efficiency. Since their very beginning, Search Engines have been designed to return Web references to user queries in milliseconds. However, dealing with millions of Web pages is not the same as achieving fast retrieval from

thousands of billions of pages. For instance, according to Netcraft Surveys the number of Web servers has doubled in the last 18 months. Information on the Web is increasing faster than computing power, and algorithms have to be rethought to keep them affordable.

3. Semantic Web. Humans are capable of using the Web to carry out tasks such as finding the Finnish word for "car", to reserve a library book, or to search for the cheapest DVD and buy it. However, a computer cannot accomplish the same tasks without human direction because web pages are designed to be read by people, not machines. The Semantic Web is a vision of information that is understandable by computers, so that they can automate more of the tedium involved in finding, sharing and combining information on the web. At its core, the Semantic Web consists of a data model called Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, Tur-

The Guest Editors

Ricardo Baeza-Yates is Director of the new Yahoo! Research laboratories in Barcelona and Latin America (Santiago, Chile). Before that he was professor and director of the Center for Web Research at the Computer Science department of the University of Chile, and also ICREA (*Institució Catalana de Recerca i Estudis Avançats*) Professor at the Department of Technology of the *Universitat Pompeu Fabra* in Barcelona, Spain. He holds a Ph.D. in Computer Science from the University of Waterloo, Canada. He is co-author of the book *Modern Information Retrieval*, published in 1999 by Addison-Wesley, as well as co-author of the 2nd edition of the *Handbook of Algorithms and Data Structures*, Addison-Wesley, 1991; and co-editor of *Information Retrieval: Algorithms and Data Structures*, Prentice-Hall, 1992. Among other awards, he received the Organization of American States award for young researchers in exact sciences (1993). In 2003 he was the first computer scientist to be elected to the Chilean Academy of Sciences. <ricardo@baeza.cl>.

Paolo Boldi obtained his Ph.D. in Computer Science at the University of Milano, where he is currently Associate Professor at the *Dipartimento di Scienze dell'Informazione*. His research interests touched many different topics in theoretical and applied computer science, such as: domain theory, non-classical computability theory, distributed computability, anonymous networks, sense of direction, self-stabilizing systems. More recently, his works focused on problems related to the World-Wide Web, a field where his research has also produced soft-

ware packages used by many people working in the same area. In particular, he contributed to write a highly efficient full-text IR engine (MG4J), and a graph compression tool (WebGraph) that is state-of-art as far as compression ratio is concerned. <boldi@dsi.unimi.it>.

José-María Gómez-Hidalgo holds a Ph.D. in Mathematics, and has been a lecturer and researcher at the *Universidad Complutense de Madrid* (UCM) and the *Universidad Europea de Madrid* (UEM), for 10 years, where he is currently the Head of the Department of Computer Science. His main research interests include Natural Language Processing (NLP) and Machine Learning (ML), with applications in Information Access in newspapers and biomedicine, and Adversarial Information Retrieval with applications in spam filtering and pornography detection on the Web. He has taken part in around 10 research projects, heading some of them. José María has co-authored a number of research papers in the topics above, which can be accessed at his home page <<http://www.esi.uem.es/~jmgomez/>>. He is Program Committee member for CEAS (*Conference on Email and Anti-Spam*) 2007, the Spam Symposium 2007 and other conferences, and he has reviewed papers for JASIST (*Journal of the American Society for Information Science and Technology*), ECIR (*European Conference on Information Retrieval*), and others. He has also reviewed research project proposals for the European Commission. <jmgomez@uem.es>.

tle, etc.), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL) that facilitate formal description of concepts, terms, and relationships within a given domain. The Semantic Web will enable new forms of Web Search, simpler and more accurate than the present ones, and it needs to be built around the intelligent processing of current Web information, including Language and Multimedia Analysis.

4. Online Social Networks. One of the most important reasons for the growth of the number of Web servers and pages is the increasing popularity of online social networking services, such as Flickr, Blogger, Digg, MySpace, YouTube, Wikipedia, and many others. These sites allow Web users to publish and share easily and quickly all forms of information, including their personal thoughts, pictures, videos, interests and references, news items, etc. The digital expression of personal relationships particularly facilitates sharing, with features such as Friend of a Friend (FOAF), which allow users to share their network of personal relationships. Social Networking services foster the emergence of online dynamic communities that make social decisions about the quality of Web content, which will be the key to the next generation of search engines (just as link analysis was the key to the current generation).

5. Personalization and other forms of context. As computational power increases, it must be converted into more advanced search engine features. The exploitation of information about user context (location, previous and recent searches, previous and recent clicks, etc.) may deliver more accurate information to the user as it can be tailored to his or her long and short search goals and information needs. Context awareness is also central to Web advertising, a field of ever-increasing importance that can exploit user information to identify targets in a more effective way.

6. Multimedia and multilingualism. The Web is still a community of diverse nationalities with different languages which have yet to be given more than limited support by search engines. Even very basic internationalization issues (such as the choice of charset and encoding) are still only covered in a very partial, unsatisfactory (and western-centric) way by current generation search engines. With only (increasingly effective) translation services as multi-language support tools, users are demanding cross-language features that allow them to cross the language barrier, retrieving results from queries in their mother tongue in many languages. The computational capabilities and the quality of multimedia analysis algorithms also allows better search interfaces and indexes, in which users pose queries in the form of pictures, audio files or even videos, in order to obtain multimedia material.

7. Web Spam. What is probably the Web's most valuable asset, the possibility of making connections from pieces of information to persons, is increasingly being the subject of abuse. Just as email spam erupted on the scene some years ago, so some content providers are now abusing this valuable means of communication to obtain an illegal commercial advantage by preparing pages and links with the

aim of getting an undeservedly high rank from a variety of popular user queries. Moreover, they hack dynamic websites (forums, social networks, etc.) in order to insert fake references and content which is ultimately targeted to delivering traffic and rank to their Web sites, and putting money in their pockets, in something which is sometimes disguised as Search Engine Optimization Web search engine operators, social networking services, etc. are required and committed to stopping, or at least reducing, this kind of abuse.

The authors invited to this special issue are prominent researchers and representatives of the search engine industry, and their papers cover most of these issues, providing the reader with a valuable overview of current and upcoming Web search engines techniques and functionalities.

The work by *Gianna del Corso, Antonio Gullì* and *Francesco Romani* focuses on the efficient computation of PageRank measures on an ever increasing Web graph. Only improvements such as those described in this paper will enable us to continue to use valuable link analysis techniques for Web site ranking.

Giuseppe Attardi describes some of the Natural Language analysis techniques which are at the core of advanced Semantic Web applications. Language Analysis makes it possible to build, maintain and exploit the resources needed by the Semantic Web (in particular, ontologies).

Personalization is covered by the work presented by *Paolo Ferragina* and *Antonio Gullì*, who describe how to obtain more personalized and accurate results by using an advanced and effective Web result clustering engine, which goes by the name of Snaket.

Luis-Alfonso Ureña-López, Manuel-Carlos Díaz-Galiano, Arturo Montejo-Raez and *M^a Teresa Martín-Valdivia* present a list of experiments on content-based multilingual and multimedia (images and text) retrieval, which support new ways of querying search engines, with an emphasis on multimodality: mixed media queries involving text and sample images.

Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo, have prepared an overview of the current problems and solutions to Web spam and other forms of abuse, focusing on link analysis and Web content filtering.

Next up, two papers present major and effective research efforts in the field of Web and large-scale information retrieval. On the one hand, *Nivio Ziviani, Alberto H.F. Laender, Edleno Silva de Moura, Altigran Soares da Silva, Carlos A. Heuser, and Wagner Meira jr.* present an overview of some of the Web search related results of Gerindo, one of the biggest and most prominent research projects on information retrieval in recent years. On the other, *Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras* describe Terrier, a high performance framework and engine designed to allow researchers to look into new information retrieval models, efficient implementations, and many other relevant topics, easily deployable on large-scale document collections.

We close this special issue with a description of the direction of research activities carried out by Yahoo! Research.

Useful References on Web Search Engines

In addition to the references and sources mentioned in the articles of this issue, interested readers may like to take a look at the following Web sites, books, journals, and conference proceedings.

Books

- S. Abiteboul, P. Buneman and D. Suciu. Data on the Web: from Relations to Semistructured Data and XML, Morgan Kaufman, 2000. ISBN: 155860622X.
- M. Agosti and A. Smeaton (editors) Information Retrieval and Hypertext, Kluwer, 1996. ISBN: 079239710X.
- R. Baeza-Yates and 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 and O. Frieder. Information Retrieval: Algorithms and Heuristics. Springer, 2004. ISBN: 1402030045.
- Witten, A. Moffat and T. Bell. Managing Gigabytes, Morgan Kaufman, 1999 (second edition). ISBN: 1558605703.

Journals

- 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/>>.

Conferences

- ACM DocEng, <<http://www.documentengineering.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/>>.

Web Sites

- Center for Web Research, <<http://www.cwr.cl>>.
- Google Labs, <<http://labs.google.com>>.
- José María Gómez home page, <<http://www.esp.uem.es/jmgomez>>.
- MAVIR Research Program, <<http://www.matir.net>>.
- Paolo Boldi home page, <<http://boldi.dsi.unimi.it>>.
- Ricardo Baeza-Yates home page, <<http://www.baeza.cl>>.
- Search Engine Watch, <<http://www.searchenginewatch.com>>.
- Web Information Retrieval resources, <<http://www.webir.org>>.
- World Wide Web Consortium, <<http://w3c.org>>.
- Yahoo! Research, <<http://research.yahoo.com>>.