UPGRADE
The European Online Magazine for the IT Professional
http://www.upgrade-cepis.org
Vol. III, No. 3, June 2002

UPGRADE is the European Online Magazine for the Information Technology 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 CEPIS society ATI (Asociación de Técnicos de Informática, Spain <http://www.ati.es/>)
UPGRADE is also published in Spanish (full issue printed, some articles online) by Novática, and in Italian (online edition, abstracts only) by the Italian CEPIS society ALSI <http://www.alsi.it> and the Italian IT portal Tecnoteca <http://www.tecnoteca.it/>
UPGRADE was created in October 2002 by CEPIS and was first published by Novática and by Informatik/Informatique, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch>)

Chief Editors
François Louis Nicolet, Zurich <nicolet@acm.org>
Rafael Fernández Calvo, Madrid <rfcalvo@ati.es>

Editorial Board
Prof. Wolffried Stucky, CEPIS President
Fernando Piera Gómez and
Rafael Fernández Calvo, ATI (Spain)
François Louis Nicolet, SI (Switzerland)
Roberto Carniel, ALSI - Tecnoteca (Italy)

English 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 Antonio Crespo Foix, © ATI 2002

Layout: Pascale Schürmann

E-mail addresses for editorial correspondence: <nicolet@acm.org> and <rfcalvo@ati.es>

E-mail address for advertising correspondence: <novatica@ati.es>

Coming issue:
"**XML, eXtensible Markup Language**"

2 Editorial: UPGRADE grows and matures
– Prof. Wolffried Stucky (President of CEPIS)

Information Retrieval and the Web

Guest Editors: Ricardo Baeza-Yates, Peter Schäuble

Joint issue with NOVÁTICA

3 Presentation – Retrieving Information: A Discipline with a Tradition
– Ricardo Baeza-Yates, Peter Schäuble
Includes a list of useful references for those interested in knowing more about Information Retrieval (IR)

5 Information Retrieval for Enterprise Content – *Prabhakar Raghavan*
The author describes a broad set of commercial applications of IR techniques.He also elaborates on the differences between Internet and Intranet retrieval.

9 Information Retrieval on the Web: A New Paradigm – *Jacques Savoy*
A survey of information retrieval on the Web is presented with emphasis on distributed retrieval, link-based ranking, and evaluation of search engines.

12 An Analysis of Query Languages for XML
– Adelaida Delgado Domínguez, Ricardo Baeza-Yates
Several query languages for XML are analysed emphasizing the W3C proposal, Xquery, from the perspective of semistructured data as well as text retrieval.

25 Methodologies to develop Web Information Systems and Comparative Analysis
– M. José Escalona, Manuel Mejías, and Jesús Torres
This article presents a comparative analysis of development methodologies for Web Information Systems.

37 Distributed Information Retrieval from Web-Accessible Digital Libraries using Mobile Agents – *J. Alfredo Sánchez, Sandra Nava Muñoz, Lourdes Fernández Ramírez, and Griselda Chevalier Dueñas*
An agent-based framework to support distributed information retrieval from heterogeneous Web-accessible digital libraries is described.

44 Automatic Extraction of Semantically-Meaningful Information from the Web – *Rafael Corchuelo, José Luis Arjona, and Antonio Ruiz*
This paper introduces another agent-based framework for automatic extraction of semantic information of Web pages.

52 Ontologies for Database Federation – *Nieves R. Brisaboa, Miguel R. Penabad, Ángeles S. Places, and Francisco J. Rodríguez*
Presents an architecture to integrate Web databases based on ontologies.

62 System for Compressing and Retrieving Structured Documents
– Joaquín Adiego, Pablo de la Fuente, Jesús Vegas, and Miguel Villarroel
A system that uses compressed inverted indices to retrieve documents considering content and structure of SGML or XML documents is described.

70 TEXRET: An Interactive TEXture RETrieval System
– Javier Ruiz-del-Solar, Pablo Navarrete, and Patricio Parada
Describes an interactive system to retrieve textures from image databases using soft-computing techniques.

78 The CLEF Campaigns: Evaluation of Cross-Language Information Retrieval Systems – *Martin Braschler, Carol Peters*
The authors describe the European Cross Language Evaluation Forum CLEF. Beside the US forum Trec and the Japanese forum NTCIR,CLEF is among the major venues to advance information retrieval technology.

82 The Web of Spain – *Ricardo Baeza Yates*
An analysis of the Spanish Web comparing it to the Brazilian and Chilean Web is presented, concluding that the results should be similar for the Webs of other European countries.

Presentation

Retrieving Information: A Discipline with a Tradition

Ricardo Baeza-Yates, Peter Schäuble

Information Retrieval (IR) is often associated with search engines and Internet; however, it evolved from an academic discipline which has its roots back in the fifties. During the first decades research activities usually took place in a Computer Science department and simple approaches based on occurrence statistics were shown surprisingly effective in retrieving relevant documents. Nevertheless, a small number of Information Retrieval research groups achieved important results in three respects:

1. *Theory*: Probabilistic retrieval models were developed that imply optimal retrieval effectiveness (see publications by Cooper, Robertson, and others). Later, retrieval was extended to other media, not only text.
2. *Systems*: Various algorithms and data structures were intended and integrated inpractical text retrieval systems (e.g. SMART, Topic, and Inquiry system) as well as multimedia retrieval systems recently.
3. *Evaluation*: Test collections were built consisting of documents, queries and – mosti mportantly – of relevance assessments that determine which documents are relevant to which queries. These test collections facilitate the comparison of different retrieval methods in respect of recall and precision (e.g. Cranfield, SMART, TREC collections).

When the Internet started growing, these Information Retrieval building blocks were ready to be used. The large amount of data as well as the federation of the Internet opened space for new and exciting concepts, such as link based raking, XML retrieval, heterogeneous data source integration, etc.

Some of these concepts are covered by the authors of this special issue on "Information Retrieval and the Web", who come from several countries; we partly use the classification above to present their papers.

Editorial Pointers

- State of the Art

Prabakhar Raghavan does a good job at describing a broad set of commercial applications of IR techniques. He also elaborates on the differences between Internet and Intranet retrieval.

Jacques Savoy presents a survey of information retrieval on the Web with emphasis on distributed retrieval, link-based ranking, and evaluation of search engines.

Adelaida Delgado and *Ricardo Baeza-Yates* present an analysis of query languages for XML emphasizing the W3C proposal, Xquery, from the perspective of semistructured data as well as text retrieval.

María José Escalona, *Manuel Mejías*, and *Jesús Torres* present a comparative analysis of development methodologies for Web information systems.

- Theory and Systems

Alfredo Sanchez, *Sandra Nava*, *Lourdes Fernández*, and *Griselda Chevalier* present an agent-based framework to support distributed information retrieval from heterogeneous Web-accessible digital libraries.

Rafael Corchuelo, *José Luis Arjona*, and *Miguel Toro* introduce another agent-based framework for automatic extraction of semantic information of Web pages.

Ricardo Baeza-Yates is Ph.D. in Computer Science (University of Waterloo, Canada). Magister in Electrical Engineering from the Universidad de Chile, and Computer Science and Electrical Engineer by the same university. He is currently Tenured Professor in the Computer Science Department of the Universidad de Chile, and Director of the Center for Research of the Web <http://www.ciw.cl>. His fields of research are information retrieval, Web mining, algorithms and information visualization. He is coauthor of the 2nd edition of the Handbook of Algorithms and Data Structures, Addison-Wesley, 1991; and coeditor of Information Retrieval: Algorithms and Data Structures, Prentice-Hall, 1992. He has also contributed several papers to journals published by professional organizations such as ACM, ATI, IEEE, and SIAM. Currently he is president of CLEI (Centro Latinoamericano de Estudios en Informática), member of the IEEE Computer Society Board of Governors and international coordinator of an Iberoamerican project on models and techniques for searching the Web financed by the Spanish agency CYTED (Programa de Cooperación Iberoamericano). In 2000 he began a startup Internet company to search the Chilean web <http://www.todocl.cl>. He can be reached by e-mail at <rbaeza@dcc.uchile.cl>

Peter Schäuble is CEO of Eurospider Information Technology AG, i.e. the leading Swiss expert in Information Retrieval and providing software for News Monitoring and Corporate Retrieval <http://www.eurospider.com>. Prior to this position, he was Assistant Professor of Computer Science at the Swiss Federal Institute of Technology (ETH) Zurich and headed the Information Retrieval research group. Peter Schäuble has a M.S. (Dipl. Math. ETH) in mathematics and a PhD (Dr. sc. techn.) in computer science both from ETH. He has been a technical staff member of the European Space Agency (ESA) and a visiting scientist at Hewlett-Packard Laboratories in Palo Alto. He published various research papers and books on Information Retrieval.
<Peter.@eurospider.com>

Nieves Brisaboa, *Miguel Penabad*, *Angeles Places*, and *Francisco Rodríguez* present an architecture to integrate Web databases based on ontologies.

Joaquín Adiego, *Pablo de la Fuente*, *Jesús Vegas*, and *Miguel Villarroel* present a system that uses compressed inverted indices to retrieve documents considering content and structure of SGML or XML documents.

Javier Ruiz del Solar, *Pablo Navarrete*, and *Patricio Parada* present an interactive system to retrieve textures from image databases using soft-computing techniques.

- Evaluation

Martin Braschler and *Carol Peters* describe the European Cross Language Evaluation Forum CLEF. Beside the US forum Trec and the Japanese forum NTCIR, CLEF is among the major venues to advance information retrieval technology.

Ricardo Baeza-Yates presents an analysis of the Spanish Web comparing it to the Brazilian and Chilean Web. The results and conclusions should be similar for the Web of other European countries.

We sincerely thank all of them for their valuable contribution, as well as the co-editors of Upgrade for their initiative.

Useful References

Collected by Ricardo Baeza Yates

In addition to the references and sources mentioned in the articles of this issue, interested readers may look at the following books, journals and conference proceedings, as well as some of the many web sites available relevant to Web standards (<http://w3c.org>), search engines (<http://www.searchenginewatch.com>), etc.

Books

Abiteboul, S., Buneman, P. & Suciu, D. Data on the Web: from Relations to Semistructured Data and XML, Morgan Kauffman, 2000.

Agosti, M. & Smeaton, A. (editors) Information Retrieval and Hypertext, Kluwer, 1996.

Baeza-Yates, R. & Ribeiro-Neto, B. Modern Information Retrieval, Addison-Wesley 1999.
Web site: <http://sunsite.dcc.uchile.cl/irbook/>

Witten, I., Moffat, A. & Bell, T. Managing Gigabytes, Morgan Kauffman, 1999 (second edition).

Journals

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

Conferences

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>