The Information Retrieval Series Volume 29

Mihai Lupu • Katja Mayer • John Tait •
Anthony J. Trippe
Editors

Current Challenges in Patent Information Retrieval

Springer

Editors
Mihai Lupu
Information Retrieval Facility
Donau-City Straße 1
Vienna 1220
Austria
m.lupu@ir-facility.org

John Tait
Information Retrieval Facility
Donau-City Straße 1
Vienna 1220
Austria
john.tait@ir-facility.org

Katja Mayer
Information Retrieval Facility
Donau-City Straße 1
Vienna 1220
Austria
k.mayer@ir-facility.org

Anthony J. Trippe
3LP Advisors
Post Rd. 7003 Suite 409
43016 Dublin, OH
USA
tony@trippe.com

ISSN 1387-5264
ISBN 978-3-642-19230-2 e-ISBN 978-3-642-19231-9
DOI 10.1007/978-3-642-19231-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011926006

ACM Computing Classification (1998): H.3, I.7, J.1

Cover design: deblik

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Patent Information Retrieval is an economically important activity. Today's economy is becoming increasingly knowledge-based and intellectual property in the form of patents plays a vital role in this growth. Between 1998 and 2008, the number of patent applications filed worldwide grew by more than 50 percent. The number of granted patents worldwide continues to increase, albeit at a slower rate than at its peak in 2006 (18%), when some 727,000 patents were granted. The substantial increase in patents granted is due, in part, to efforts by patent offices to reduce backlogs as well as the significant growth in the number of patents granted by China and, to a lesser extent in the more recent years, by the Republic of Korea. According to these statistics, the total number of patents in force worldwide at the end of 2008 was approximately 6.7 million (WIPO report 2010). A prior art search might have to cover as many as 70 million patents. By combining data from Ocean Tomo's Intangible Asset Market Value Survey, and Standard and Poor's 1200 Index we can estimate that the global value of patents exceeds US$10 trillion in 2009.

A patent is a bargain between the inventor and the state. The inventor must teach the community how to make the product, and use the techniques he/she has invented in return for a limited monopoly which gives him a set time to exploit his invention and realise its value. Patents are used for many reasons, e.g. to protect inventions, to create value and to monitor competitive activities in a field. Much knowledge is distilled through patents, which is never published elsewhere. Thus patents form an important knowledge resource—e.g. much technical information represented in patents is not represented in scientific literature—and are at the same time important legal documents.

Despite the overall increase in patent applications and grants, a situation of economic downturn, such as the one the world has experienced in 2008, leads to a reduction in patent applications and grants (as indicated by preliminary figures published by WIPO for 2009). This is, to some extent, explained by the high costs involved in applying for a patent, particularly for small enterprises. The costs of the pre-application process, the long duration of the application process and the corresponding uncertainty in the long-term economy in such periods of economic downturn need to be addressed by changing the way we search the patent and non-patent

literature. Both the Intellectual Property (IP) professionals and the Information Retrieval (IR) scientists can see this book as a challenge: for the former, in terms of adapting to new tools; for the latter, in terms of creating better tools for an obviously difficult task; for both, in terms of engaging in exchange and cooperation.

In the past 10 or 15 years, general information retrieval and Web search engines have made tremendous advances. And still, we see a huge gap between the technologies which, on the one hand, were emerging from research labs and in use by major internet search engines, in e-commerce, and in enterprise search systems, and, on the other, the systems in day-to-day use by the patent search communities.

It has been estimated that since 1991, when the US Federal National Institute of Standards and Technology (NIST) began its Text Retrieval Conference (TREC) evaluation campaign, the available information retrieval and search systems have improved 40% or more in their ability to find relevant documents. And yet the technologies underlying the patent search system were largely unaffected by these changes. Patent searchers generally use the same technology as in the 1980s. Boolean specification of searches and set-based retrieval are the norm rather than the ranked retrieval systems used by Google and the like. Tools in some areas have moved on significantly: some providers have semantic analysis tools, others effective visualisation mechanisms for patent documents. And yet there has not been the kind of revolution in patent search which Google had represented for Web search.

In the past few years, the Information Retrieval Facility (a not-for-profit research institution based in Vienna, Austria) has organised a series of events to bring together leading researchers in IR with those who practice and use patent search, to establish the interdisciplinary dialogue between the IR and the IP communities and to create a discursive as well as empirical space for sustainable discussion and innovation.

In the first Information Retrieval Facility Symposium in Vienna in 2007 (www.irfs.at), a distinguished audience of information retrieval scientists and patent search specialists started to explore the reasons for the knowledge gap. It turned out that academic researchers were often unaware of the specialised needs of the patent searchers: for example, they needed a degree of transparency quite unlike the casual Web searchers, upon which the academics mainly focussed. The patent searchers were often unaware of the advances made in other areas, and how they had been achieved. There were difficulties in finding (and using) a common, comprehensible vocabulary. In the course of that first Symposium, and through subsequent IRF symposia and other joint activities, such as the CLEF-IP and TREC-CHEM tracks, the PaIR and Aspire workshops, major progress has been made in developing a common understanding, and even an agenda between search researchers and technologists and the patent search community.

This book is part of the development of that joint understanding. Its origins lie in the idea of producing post-proceedings for the first IRF Symposium. That idea was not fully followed up, in part because of pressure to produce more practical, action-oriented work, and in part because many of the participants felt their approaches were at too early a stage for formal publication. In the course of the following years it became apparent there really was a demand to produce a volume which was accessible to both the patent search community and to the information retrieval research

community; to provide a collected and organized introduction to the work and views of the two sides of the emerging patent search research and innovation community; and to provide a coherent and organised view of what has been achieved and, perhaps even more significantly, of what remains to be achieved.

We have already noted the need for transparency (or at least defensibility) of search processes from the patent search community. We hope this book will allow the IR researchers to better understand why such transparency is needed, and what it means in practise. Furthermore, it is our hope that this book will also be a valuable resource for IP professionals in learning about current approaches of IR in the patent domain. It has often been difficult to reconcile the focus on useful technological innovation from the IP community, with the demands for scientific rigour and to proceed on the basis of sound empirical evidence, which is such an important feature of IR (in contrast to some other areas of computer science).

Moreover, patent search is an inherently multilingual and multinational topic: the novelty of a patent may be dismissed by finding a document describing the same idea in any language anywhere in the world. Patents are complex legal documents, even less accessible than the scientific literature. These are just some of the characteristics of the patent system, which make it an important challenge for the search, information retrieval and information access communities.

The book has had a lengthy and difficult gestation: the list of authors has been revised many times as a result of changes in institutional, occupational and private circumstances. Although we, the editors, do feel we have succeeded in producing a volume which will provide important perspectives of the issues affecting patent search research and innovation at the time of writing, as well as a useful, brief introduction to the outlook and literature of the community accessible to its members, regardless of their background, we would have liked to cover several topics not represented here.

In particular it was disappointing we could not include a chapter on NTCIR, the first of the evaluation campaigns to focus seriously on patents. Also, a chapter on the use of Latent Semantic Indexing for the patent domain had been planned, which ultimately could not appear in this book.

Several of the chapters have been written jointly by intellectual property and information retrieval experts. Members of both communities with a background opposite to the primary author have reviewed all the chapters. It has not always been easy to reconcile their differing viewpoints: we must thank them for taking the time to resolve their differences and for taking the opportunity to exchange their knowledge across fields and disciplinary mind-sets and to engage in a mutual discourse that will hopefully foster the understanding in the future.

Finally, we would like to thank the IRF for making this publication possible, the publisher, Springer; and in particular Ralf Gerstner, for the patience with which he accepted the numerous delays, as well as the external reviewers who read each chapter and provided the authors with valuable advice.

The editors are very grateful to the following persons, who agreed to review the manuscripts: Stephen Adams, Linda Andersson, Geetha Basappa, John M. Barnard, Shariq Bashir, Helmut Berger, Katrien Beuls, Ted Briscoe, Ben Carterette, Paul

Clough, Bruce Croft, Szabolcs Csepregi, Barrou Diallo, Karl A. Froeschl, Norbert Fuhr, Eric Gaussier, Julio Gonzalo, Allan Hanbury, Christopher G. Harris, Ilkka Havukkala, Bruce Hedin, Cornelis H.A. Koster, Mounia Lalmas, Patrice Lopez, Teresa Loughbrough, Marie-Francine Moens, Henning Müller, Iadh Ounis, Florina Piroi, Keith van Rijsbergen, Patrick Ruch, Philip Tetlow, Henk Thomas, Ingo Thon, Steve Tomlinson, Anthony Trippe, Suzan Verberne, Ellen M. Voorhees, Peter Willett, Christa Womser-Hacker.

Mihai Lupu
Katja Mayer
John Tait
Anthony Trippe

Contents

Contributors

Doreen Alberts Theravance Inc., 901 Gateway Blvd., South San Francisco, CA, USA

Robert Arens Nuance Communications, Burlington, MA, USA, robert.arens@nuance.com

Niraj Aswani Department of Computer Science, University of Sheffield, Sheffield, UK, N.Aswani@dcs.shef.ac.uk

Hany Azzam Queen Mary University of London, London, UK, hany@eecs.qmul.ac.uk

Richard Bache Department of Computer and Information Sciences, University of Strathclyde, Glasgow G4 1XH, Scotland, UK, richard.bache@gmail.com

Jean G. Beney Dept. Informatique, LCI, INSA de Lyon, Lyon, France, jean.beney@insa-lyon.fr

Karim Benzineb SIMPLE SHIFT, Ruelle du P'tit-Gris 1, 1228 Plan-les-Ouates, Switzerland, karim@simple-shift.com

Helmut Berger max.recall information systems, Vienna, Austria, h.berger@max-recall.com

Harald Bosch Institute for Interactive Systems and Visualization, Universität Stuttgart, Stuttgart, Germany

Ted Briscoe University of Cambridge, Cambridge, UK, Ted.Briscoe@cl.cam.ac.uk; iLexIR Ltd, Cambridge, UK

Ben Carterette University of Delaware, Newark, DE 19716, USA, carteret@cis.udel.edu

Hamish Cunningham Department of Computer Science, University of Sheffield, Sheffield, UK, H.Cunningham@dcs.shef.ac.uk

Dominic DeMarco DeMarco Intellectual Property, LLC, 1111 16th Street, South Arlington, VA, USA

Barou Diallo European Patent Office, Patentlaan 2, 2288 EE Rijswijk Zh, Netherlands, bdiallo@epo.org

Michael Dittenbach max.recall information systems, Vienna, Austria, m.dittenbach@max-recall.com

Denise Fobare-DePonio Camarillo, CA, USA

Mark A. Greenwood Department of Computer Science, University of Sheffield, Sheffield, UK, M.Greenwood@dcs.shef.ac.uk

Jacques Guyot SIMPLE SHIFT, Ruelle du P'tit-Gris 1, 1228 Plan-les-Ouates, Switzerland

Christopher G. Harris Informatics Program, The University of Iowa, Iowa City, IA, USA, christopher-harris@uiowa.edu

Karl Harrison University of Cambridge, Cambridge, UK, Harrison@hep.phy.cam.ac.uk

Bruce Hedin H5, 71 Stevenson St., San Francisco, CA 94105, USA, bhedin@h5.com

John D. Holliday Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK

Jimmy Huang York University, Toronto, Canada, jhuang@yorku.ca

Hideaki Kamaya Hiroshima City University, 3-4-1 Ozukahigashi, Asaminami-ku, Hiroshima 731-3194, Japan, kamaya@ls.info.hiroshima-cu.ac.jp

Iraklis A. Klampanos University of Glasgow, Glasgow, UK, iraklis@dcs.gla.ac.uk

Steffen Koch Institute for Interactive Systems and Visualization, Universität Stuttgart, Stuttgart, Germany

Cornelis H.A. Koster Computing Science Institute ICIS, Univ. of Nijmegen, Nijmegen, The Netherlands, kees@cs.ru.nl

Ken Koubek Koubek Information Consulting Services LLC, Wilmington, DE, USA

Mihai Lupu Information Retrieval Facility, Vienna, Austria, m.lupu@ir-facility.org

Andrew Naish Camtology Ltd, Cambridge, UK, A.Naish@gmail.com

Hidetsugu Nanba Hiroshima City University, 3-4-1 Ozukahigashi, Asaminami-ku, Hiroshima 731-3194, Japan, nanba@hiroshima-cu.ac.jp

Manabu Okumura Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8503, Japan, oku@pi.titech.ac.jp

Peter Parapatics Department of Software Technology and Interactive Systems, Vienna University of Technology, Favoritenstr. 9-11/188, 1040 Vienna, Austria, p.parapatics@gmail.com

Andy Parker University of Cambridge, Cambridge, UK, Parker@hep.phy.cam.ac.uk; Camtology Ltd, Cambridge, UK

Andreas Pesenhofer max.recall information systems, Vienna, Austria, a.pesenhofer@max-recall.com

Florina Piroi Information Retrieval Facility, Vienna, Austria, f.piroi@ir-facility.org

Marek Rei University of Cambridge, Cambridge, UK, Marek.Rei@hep.phy.cam.ac.uk

Ian Roberts Department of Computer Science, University of Sheffield, Sheffield, UK, I.Roberts@dcs.shef.ac.uk

Suzanne Robins Patent Information Services, Inc., Westborough, MA, USA

Matthew Rodgers Landon IP, Alexandria, VA, USA

Thomas Roelleke Queen Mary University of London, London, UK, thor@eecs.qmul.ac.uk

Ian Ruthven Department of Computer and Information Sciences, University of Strathclyde, Glasgow, G12 8DY, UK, ir@cis.strath.ac.uk

Akihiro Shinmori INTEC Systems Institute Inc., 1-3-3 Shinsuna, Koto-ku, Tokyo 136-0075, Japan, shinmori_akihiro@intec-si.co.jp

Advaith Siddharthan University of Aberdeen, Aberdeen, UK, Advaith@abdn.ac.uk

Edlyn Simmons Simmons Patent Information Service, LLC, Mason, OH, USA

David Sinclair Camtology Ltd, Cambridge, UK, David.Sinclair@imense.co.uk

Mark Slater University of Cambridge, Cambridge, UK, Slater@hep.phy.cam.ac.uk

Padmini Srinivasan Computer Science Department and Informatics Program, The University of Iowa, Iowa City, IA, USA, padmini-srinivasan@uiowa.edu

Veronika Stefanov Information Retrieval Facility, Vienna, Austria, v.stefanov@ir-facility.org

Valentin Tablan Department of Computer Science, University of Sheffield, Sheffield, UK, V.Tablan@dcs.shef.ac.uk

John I. Tait Information Retrieval Facility, Techgate, Donau City Strasse 1, Vienna, 1220, Austria, john.tait@ir-facility.org

Toshiyuki Takezawa Hiroshima City University, 3-4-1 Ozukahigashi, Asaminami-ku, Hiroshima 731-3194, Japan, takezawa@hiroshima-cu.ac.jp

Hidekazu Tanigawa IRD Patent Office, 8th floor, OMM Building, 1-7-31, Otemae, Chuo-ku, Osaka 540-0008, Japan, htanigawa@ird-pat.com

Stephen Tomlinson Open Text Corporation, Ottawa, Ontario, Canada, stomlins@opentext.com

Anthony Trippe 3LP Advisors, Dublin, OH, USA, tony@trippe.com

Suzan Verberne Computing Science Institute ICIS, Univ. of Nijmegen, Nijmegen, The Netherlands, s.verberne@cs.ru.nl

Merijn Vogel Computing Science Institute ICIS, Univ. of Nijmegen, Nijmegen, The Netherlands, merijnv@cs.ru.nl

Ellen M. Voorhees NIST, Gaithersburg, MD 20879, USA, Ellen.Voorhees@nist.gov

Rebecca Watson iLexIR Ltd, Cambridge, UK, Bec.Watson@gmail.com

Peter Willett Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK

Cynthia Barcelon Yang Patent Information Users Group (PIUG), 505 Amberleigh Drive, Pennington, NJ, USA

Veronika Zenz max.recall information systems, Vienna, Austria, v.zenz@max-recall.com

Jianhan Zhu True Knowledge Ltd., Cambridge, UK, jianhanzhu@gmail.com