

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Michail Salampassis Birger Larsen (Eds.)

Multidisciplinary Information Retrieval

5th Information Retrieval Facility Conference
IRFC 2012
Vienna, Austria, July 2-3, 2012
Proceedings

Volume Editors

Michail Salampasis
Vienna University of Technology
Institute of Software Technology
and Interactive Systems
Favoritenstr. 9-11/188
1040 Vienna, Austria
E-mail: salampasis@ifs.tuwien.ac.at

Birger Larsen
Royal School of Library
and Information Science
2300 Copenhagen, Denmark
E-mail: blar@iva.dk

The conference is co-organised by
the Multilingual and Multifaceted Interactive Information Access
COST Action and the Vienna University of Technology,
Information Management and Preservation Group

ISSN 0302-9743
ISBN 978-3-642-31273-1
DOI 10.1007/978-3-642-31274-8

e-ISSN 1611-3349
e-ISBN 978-3-642-31274-8

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012939842

CR Subject Classification (1998): H.3.3, H.3, H.2.8, I.2.7, H.4, H.5

LNCS Sublibrary: SL 3 – Information Systems and Application,
incl. Internet/Web and HCI

© Springer-Verlag Berlin Heidelberg 2012

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Preface

These proceedings contain the refereed papers presented at the 5th Information Retrieval Facility Conference (IRFC) for Science and Industry, which was held in Vienna during July 2–3, 2012. The conference was organized by the IC1002 COST Action Multilingual and Multifaceted Interactive Information Access (MUMIA) and the Vienna University of Technology.

COST is an intergovernmental framework for European Cooperation in Science and Technology, allowing the coordination of nationally funded research on a European level. The IC1002 COST Action on Multilingual and Multifaceted Interactive Information Access (MUMIA) coordinates collaboration between researchers and professionals working on three key disciplines for the future of search: machine translation (MT), information retrieval (IR) and multifaceted interactive information access (MIIA). To form a common basis for collaboration the domain of patent retrieval has been selected as a use case, as it provides highly sophisticated and information intensive search tasks. The MUMIA action aims to empower the synergies from the disparate research fields of MT/IR/MIIA within the specific context of patent search and other next-generation Web applications. More than 60 experienced scientists and young researchers from 26 countries participate in MUMIA's research network.

The 5th IRFC aimed to tackle three complementary research areas: information retrieval, machine translation for search solutions and interactive information access.

The IRF Conference provides a multi-disciplinary, scientific forum that aims to bring young researchers into contact with industry at an early stage. Another emphasis is put on the applicability of IR solutions to real industry cases and the respective challenges. A panel of experts from both academia and industry reviewed the relevance of the submitted papers. IRFC 2012 received 17 high-quality submissions, of which 11 were accepted for presentation at the conference and appear here. Each submission was reviewed by at least three Program Committee members. Five of these are IR papers in the patent search domain, two deal with the machine translation and natural language processing for the IR area, while the remaining four deal with Web IR and associated tasks.

In addition to these papers we also invited experts from industry, technology providers and implementers of novel technology to present challenges, real problems and tools in search and search-related areas, delivered as talks in the Industry and Demo Sessions. We would like to thank all of them for presenting novel and innovative search tools and UIs and sharing with researchers other ideas related to the process of information finding.

Our sincere thanks go out to: the Local Organizing Committee, Andreas Rauber, Allan Hanbury, Mihai Lupu, Marie-Pierre Garnier, Christiane Troniger and Christine Haas; Norbert Fuhr for his support as General Chair of the

conference; the IRF scientific board for agreeing that we organize the conference; the keynote speaker; the members of the Program Committee and the additional reviewers for their thorough reviews; the Municipality of Vienna for their support and for providing us with the City Hall to host the conference dinner; the MUMIA COST Action, the Vienna University of Technology, and the Austrian Computer Society for their organizational assistance, and, to Microsoft Research for providing sponsorship to the conference.

We hope that you enjoy the conference proceedings.

April 2012

Michail Salampasis
Birger Larsen

Organization

Program Committee

Galia Angelova	Institute for Parallel Processing, Bulgarian Academy of Sciences, Bulgaria
Pavel Braslavski	Yandex
Paul Buitelaar	DERI - National University of Ireland, Galway, Ireland
Pablo Castells	Universidad Autónoma de Madrid, Spain
Ivan Chorbev	FEIT
Bruce Croft	University of Massachusetts Amherst, USA
Danco Davcev	UIST Ohrid
Michael Dittenbach	Max Recall
Gideon Dror	The Academic College of Tel-Aviv-Yaffo, Israel
Anna Esposito	Second University of Naples, Italy
Sebastien Ferre	Université de Rennes 1, France
Wilfried Gansterer	University of Vienna, Austria
Gregory Grefenstette	Exalead
Allan Hanbury	Vienna University of Technology, Austria
David Hawking	Funnelback
Ivan Koichev	University of Sofia “St. Kliment Ohridski”, Bulgaria
Yiannis Kompatsiaris	CERTH - ITI
Udo Kruschwitz	University of Essex, UK
David Lamas	Tallinn University, Estonia
Birger Larsen	Royal School of Library and Information Science
Christina Lioma	University of Copenhagen, Denmark
David Losada	University of Santiago de Compostela, Spain
Mihai Lupu	Vienna University of Technology, Austria
Bente Maegaard	University of Copenhagen, Denmark
Walid Magdy	Dublin City University, Ireland
Marie-Francine Moens	Katholieke Universiteit Leuven, Belgium
Igor Mozetic	Jozef Stefan Institute, Slovenia
Henning Müller	HES-SO
Hidetsugu Nanba	Hiroshima City University, Japan
Ragnar Nordlie	Oslo University College, Norway
Andreas Nuernberger	University of Magdeburg, Germany
Georgios Paltoglou	University of Wolverhampton, UK
Andreas Rauber	Vienna University of Technology, Austria

VIII Organization

Patrick Ruch	University Hospital of Geneva, Switzerland
Tony Russell-Rose	UXLabs
Ian Ruthven	University of Strathclyde, UK
Michail Salampanis	Vienna University of Technology, Austria
Jan Snajder	University of Zagreb, FER, Croatia
Oren Somekh	Yahoo! Labs
Hannu Toivonen	University of Helsinki, Finland
Elaine Toms	University of Sheffield, UK
Dolf Trieschnigg	University of Twente, The Netherlands
Howard Turtle	Syracuse University, USA
Ioannis Tzitzikas	ICS-FORTH, University of Crete, Greece
David Vallet	Universidad Autónoma de Madrid, Spain
Suzan Verberne	Centre of Language and Speech Technology, Radboud University Nijmegen, The Netherlands
Thijs Westerveld	Teezir Search Solutions

Additional Reviewers

Arcan, Mihael
Nitsche, Marcus
Sébillot, Pascale

Table of Contents

Patent Search

Learning-Based Pseudo-Relevance Feedback for Patent Retrieval	1
<i>Parvaz Mahdabi and Fabio Crestani</i>	
Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus	12
<i>Katharina Wäschle and Stefan Riezler</i>	
Query Enhancement for Patent Prior-Art-Search Based on Keyterm Dependency Relations and Semantic Tags	28
<i>Khanh-Ly Nguyen and Sung-Hyon Myaeng</i>	

Web Search

Analysis and Detection of Web Spam by Means of Web Content	43
<i>Víctor M. Prieto, Manuel Álvarez, Rafael López-García, and Fidel Cacheda</i>	
Discovery of Environmental Nodes in the Web	58
<i>Anastasia Moumtzidou, Stefanos Vrochidis, Sara Tonelli, Ioannis Kompatsiaris, and Emanuele Pianta</i>	
Web Searching with Entity Mining at Query Time	73
<i>Pavlos Fafalios, Ioannis Kitsos, Yannis Marketakis, Claudio Baldassarre, Michail Salampasis, and Yannis Tzitzikas</i>	

Applications

Patent and Norm Exploration with the m2n Knowledge Discovery Suite	89
<i>Tobias Schleser, Doris Ipsmiller, and Roland Mörzinger</i>	
Hierarchical Classification of Web Documents by Stratified Discriminant Analysis	94
<i>Juan Carlos Gomez and Marie-Francine Moens</i>	
PatMedia: Augmenting Patent Search with Content-Based Image Retrieval	109
<i>Stefanos Vrochidis, Anastasia Moumtzidou, Gerard Ypma, and Ioannis Kompatsiaris</i>	

Query Formulation and Analysis

Generating Variant Keyword Forms for a Morphologically Complex
Language Leads to Successful Information Retrieval with Finnish 113
Kimmo Kettunen and Paavo Arvola

Analyzing Query Logs of USPTO Examiners to Identify Useful Query
Terms in Patent Documents for Query Expansion in Patent Searching:
A Preliminary Study 127
Wolfgang Tannebaum and Andreas Rauber

Discovering Relevant Features for Effective Query Formulation..... 137
Luepol Pipanmaekaporn and Yuefeng Li

Author Index 153