

*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*

Tiziana Catarci Pamela Forner  
Djoerd Hiemstra Anselmo Peñas  
Giuseppe Santucci (Eds.)

# Information Access Evaluation

Multilinguality, Multimodality,  
and Visual Analytics

Third International Conference  
of the CLEF Initiative, CLEF 2012  
Rome, Italy, September 17-20, 2012  
Proceedings



Springer

## Volume Editors

Tiziana Catarci

Giuseppe Santucci

Sapienza University of Rome, Dept. of Computer, Control and Management

Engineering Antonio Ruberti

Via Ariosto 25, 00185 Rome, Italy

E-mail: {catarci, santucci}@dis.uniroma1.it

Pamela Forner

Center for the Evaluation of Language and Communication Technologies (CELCT)

Via alla Cascata 56/c, 38123 Povo, TN, Italy

E-mail: forner@celct.it

Djoerd Hiemstra

University of Twente, Dept. of Computer Science, Database Group

PO Box 217, 7500 AE Enschede, The Netherlands

E-mail: hiemstra@cs.utwente.nl

Anselmo Peñas

UNED Natural Language Processing and Information Retrieval Research Group

E.T.S.I. Informática de la UNED

c/ Juan del Rosal 16, 28040 Madrid, Spain

E-mail: anselmo@lsi.uned.es

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-33246-3

e-ISBN 978-3-642-33247-0

DOI 10.1007/978-3-642-33247-0

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012946265

CR Subject Classification (1998): I.7, I.2.7, H.3.1, H.3.3, H.3.7, H.4.1, H.5.3, H.2.8, I.1.3

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

Since 2000 the Cross-Language Evaluation Forum (CLEF) has played a leading role in stimulating research and innovation in a wide range of key areas in the domain of multimodal and multilingual information access. Through the years, CLEF has promoted the study and implementation of evaluation methodologies for diverse tasks, resulting in the creation of a broad, strong, and multidisciplinary research community.

Until 2010, the outcomes of experiments carried out under the CLEF umbrella were presented and discussed at annual workshops in conjunction with the European Conference for Digital Libraries. CLEF 2010 represented a radical departure from this “classic” CLEF format. While preserving CLEF’s traditional core goals, namely, benchmarking activities carried in various tracks, we complemented these activities with a peer-reviewed conference component aimed at advancing research in the evaluation of complex information access systems in different languages and modalities.

CLEF 2011 and CLEF 2012 continued to implement this format, with keynotes, contributed papers, lab sessions, poster sessions, reporting of other benchmarking initiatives and, for the first time this year, an evaluation clinic session where people with retrieval evaluation problems of some kind would be able to talk to evaluation experts and get methodological advice, new ideas, pointers to related problems, available solutions, etc.

This year, the papers accepted for the conference included research on information access and evaluation initiatives, methodologies, and infrastructures. Two keynote speakers highlighted important issues related to our field. Peter Clark (Vulcan Inc.) presented a case of innovation turned into a company product that allows users not only to read and browse a textbook, but also to ask questions and get reasoned or retrieved answers back, explore the material through semantic connections, and receive suggestions of useful questions to ask. Tobias Schreck (University of Konstanz), on the other hand, showed current approaches, applications, and challenges for the application of visual analytics in document repositories.

CLEF 2012 featured seven benchmarking activities: RepLab, INEX, QA4MRE, CLEF-IP, ImageCLEF, PAN, and CHiC. In parallel, the CLEFeHealth workshop was hosted, dealing with cross-language evaluation of methods, applications, and resources for eHealth document analysis with a focus on written and spoken natural-language processing.

All the experiments carried out by systems during the evaluation campaigns are described in a separate publication, namely, the Working Notes, distributed during CLEF 2012 and available on-line.

The success of CLEF 2012 would not have been possible without the invaluable contributions of all the members of the Program Committee, Organizing Committee, students and volunteers that supported the conference in its various stages. We would like to express also our gratitude to the sponsoring organizations for their significant and timely support. These proceedings were prepared with the assistance of the Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy.

July 2012

Tiziana Catarci  
Pamela Forner  
Djoerd Hiemstra  
Anselmo Peñas  
Giuseppe Santucci

# Organization

CLEF 2012 was organized by Sapienza University of Rome, Italy.

## General Chairs

Tiziana Catarci  
Djoerd Hiemstra

Sapienza University of Rome, Italy  
University of Twente, The Netherlands

## Program Chairs

Anselmo Peñas  
Giuseppe Santucci

National Distance Learning University, Spain  
Sapienza University of Rome, Italy

## Evaluation Labs Chairs

Jussi Karlgren  
Christa Womser-Hacker

Swedish Institute of Computer Science, Sweden  
University of Hildesheim, Germany

## Resource Chair

Khalid Choukri

Evaluations and Language Resources Distribution  
Agency (ELDA), France

## Organization Chair

Emanuele Pianta

Center for the Evaluation of Language and  
Communication Technologies (CELCT), Italy

## Organizing Committee

*Sapienza University of Rome, Italy:*

Carola Aiello  
Giuseppe Santucci

*Consulta Umbria Congressi, Perugia, Italy*

*Center for the Evaluation of Language and Communication Technologies  
(CELCT), Italy:*

Pamela Forner  
Giovanni Moretti

## Program Committee

Alexandra Balahur	Joint Research Centre - JRC - European Commission, Italy
Yassine Benajiba	Philips, USA
Khalid Choukri	Evaluations and Language Resources Distribution Agency (ELDA), France
Walter Daelemans	University of Antwerp, Belgium
Nicola Ferro	University of Padua, Italy
Norbert Fuhr	University of Duisburg, Germany
Julio Gonzalo	National Distance Learning University, Spain
Donna Harman	National Institute of Standard and Technology, USA
Gareth Jones	Dublin City University, Ireland
Noriko Kando	National Institute of Informatics, Japan
Evangelos Kanoulas	Google, Switzerland
Bernardo Magnini	Fondazione Bruno Kessler, Italy
Prasenjit Majumder	Dhirubhai Ambani Institute of Information and Communication Technology, India
Thomas Mandl	University of Hildesheim, Germany
Paul McNamee	Johns Hopkins University, USA
Manuel Montes-y-Gómez	National Institute of Astrophysics, Optics and Electronics, Mexico
Henning Müller	University of Applied Sciences Western Switzerland, Switzerland
Jian-Yun Nie	University of Montreal, Canada
Carol Peters	ISTI CNR Pisa, Italy
Vivien Petras	Humboldt University, Germany
Álvaro Rodrigo	National Distance Learning University, Spain
Paolo Rosso	Universitat Politècnica de València, Spain
Tobias Schreck	University of Konstanz, Germany
José Luis Vicedo	University of Alicante, Spain
Christa Womser-Hacker	University of Hildesheim, Germany

## Sponsoring Institutions

CLEF 2012 benefited from the support of the following organizations:

### Gold Sponsors



ELIAS



PROMISE Network of Excellence



Sapienza University of Rome

### Silver Sponsors



European Science Foundation



Quaero



# From Information Retrieval to Knowledgeable Machines

Peter Clark

Vulcan Inc.,  
505 Fifth Ave South, Suite 900,  
Seattle, WA, 98104  
[peterc@vulcan.com](mailto:peterc@vulcan.com)

**Abstract.** Ultimately we would like our machines to not only search and retrieve information, but also have some “understanding” of the material that they are manipulating so that they can better meet the user’s needs. In this talk, I will present our work in Project Halo to create an (iPad hosted) “knowledgeable biology textbook”, called Inquire. Inquire includes a formal, hand-crafted knowledge base encoding some of the book’s content, being augmented (this year) with capabilities for textual entailment and question-answering directly from the book text itself. Inquire allows the user to not only read and browse the textbook, but also to ask questions and get reasoned or retrieved answers back, explore the material through semantic connections, and receive suggestions of useful questions to ask. In this talk I will describe the project, in particular the textual question-answering component and its use of natural language processing, paraphrasing, textual entailment, and its exploitation of the formal knowledge base. I will also discuss the interplay being developed between the hand-built knowledge and automatic text-extracted knowledge, how each offers complementary strengths, and how each can leverage the other. Finally I will discuss the value of this approach, and argue for the importance of creating a deeper understanding of textual material, and ultimately more knowledgeable machines.

# Visual Search and Analysis in Textual and Non-textual Document Repositories-Approaches, Applications, and Research Challenges

Tobias Schreck

University of Konstanz,  
Computer and Information Science,  
Universitaetsstrasse 10, Box 78,  
D-78457 Konstanz, Germany  
[Tobias.Schreck@uni-konstanz.de](mailto:Tobias.Schreck@uni-konstanz.de)

**Abstract.** Information retrieval and analysis are key tasks in dealing with the information overload problem characteristic for today's networked digital environments. Advances in data acquisition, transmission and storage, and emergence of social media, lead to an abundance of textual and non-textual information items available to everyone at any time. Advances in visual-interactive data analysis can provide for effective visual interfaces for query formulation, navigation, and result exploration in complex information spaces. In this presentation, we will discuss selected approaches for visual analysis in large textual and non-textual document collections. First, recent techniques for visual analysis of readability, sentiment and opinion properties in large amounts of textual documents, including promising application possibilities, will be discussed. Then, we will focus on visual analysis support for information retrieval in non-textual documents, in particular multimedia and time-oriented research data. We argue that new visual-interactive approaches can provide for effective user access to large document corpora, including discovering of interesting relationships between data items, and understanding the space of similarity notions for a given document repository. We will conclude the presentation by discussing research opportunities at the intersection of visual data analysis, information retrieval, and evaluation.

# Table of Contents

## Benchmarking and Evaluation Initiatives

Analysis and Refinement of Cross-Lingual Entity Linking . . . . .	1
<i>Taylor Cassidy, Heng Ji, Hongbo Deng, Jing Zheng, and Jiawei Han</i>	
Seven Years of INEX Interactive Retrieval Experiments – Lessons and Challenges . . . . .	13
<i>Ragnar Nordlie and Nils Pharo</i>	
Bringing the Algorithms to the Data: Cloud-Based Benchmarking for Medical Image Analysis . . . . .	24
<i>Allan Hanbury, Henning Müller, Georg Langs, Marc André Weber, Bjoern H. Menze, and Tomas Salas Fernandez</i>	
Going beyond CLEF-IP: The ‘Reality’ for Patent Searchers? . . . . .	30
<i>Julia J. Jürgens, Preben Hansen, and Christa Womser-Hacker</i>	
MusiClef: Multimodal Music Tagging Task . . . . .	36
<i>Nicola Orio, Cynthia C.S. Liem, Geoffroy Peeters, and Markus Schedl</i>	

## Information Access

Generating Pseudo Test Collections for Learning to Rank Scientific Articles . . . . .	42
<i>Richard Berendsen, Manos Tsagkias, Maarten de Rijke, and Edgar Meij</i>	
Effects of Language and Topic Size in Patent IR: An Empirical Study . . . . .	54
<i>Florina Piroi, Mihai Lupu, and Allan Hanbury</i>	
Cross-Language High Similarity Search Using a Conceptual Thesaurus . . . . .	67
<i>Parth Gupta, Alberto Barrón-Cedeño, and Paolo Rosso</i>	
The Appearance of the Giant Component in Descriptor Graphs and Its Application for Descriptor Selection . . . . .	76
<i>Anita Keszler, Levente Kovács, and Tamás Szirányi</i>	
Hidden Markov Model for Term Weighting in Verbose Queries . . . . .	82
<i>Xueliang Yan, Guanglai Gao, Xiangdong Su, Hongxi Wei, Xueliang Zhang, and Qianqian Lu</i>	

## Evaluation Methodologies and Infrastructure

DIRECTIONS: Design and Specification of an IR Evaluation Infrastructure .....	88
<i>Maristella Agosti, Emanuele Di Buccio, Nicola Ferro, Ivano Masiero, Simone Peruzzo, and Gianmaria Silvello</i>	
Penalty Functions for Evaluation Measures of Unsegmented Speech Retrieval.....	100
<i>Petra Galuščáková, Pavel Pecina, and Jan Hajič</i>	
Cumulated Relative Position: A Metric for Ranking Evaluation .....	112
<i>Marco Angelini, Nicola Ferro, Kalervo Järvelin, Heikki Keskustalo, Ari Pirkola, Giuseppe Santucci, and Gianmaria Silvello</i>	
Better than Their Reputation? On the Reliability of Relevance Assessments with Students .....	124
<i>Philipp Schaer</i>	

## Posters

Comparing IR System Components Using Beanplots.....	136
<i>Jens Kürsten and Maximilian Eibl</i>	
Language Independent Query Focused Snippet Generation .....	138
<i>Pinaki Bhaskar and Sivaji Bandyopadhyay</i>	
A Test Collection to Evaluate Plagiarism by Missing or Incorrect References .....	141
<i>Solange de L. Pertile and Viviane P. Moreira</i>	
<b>Author Index</b> .....	145