Lecture Notes in Computer Science 7538

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Stefano Ceri Marco Brambilla (Eds.)

Search Computing

Broadening Web Search

Springer

Volume Editors

Stefano Ceri
Marco Brambilla
Politecnico di Milano
Dipartimento di Elettronica e Informazione
Via Ponzio, 34/5, 20133 Milan, Italy
E-mail: {ceri, mbrambil}@elet.polimi.it

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-34212-7 e-ISBN 978-3-642-34213-4
DOI 10.1007/978-3-642-34213-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012951180

CR Subject Classification (1998): H.3, H.4, H.5, C.2.4, F.2.2, D.1.3, J.1

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

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

The Search Computing project (SeCo), funded by the European Research Council as an advanced IDEAS grant, aims at building concepts, algorithms, tools, and technologies to support complex Web queries—whose answers cannot be gathered through conventional "page-based" search. Indeed, while the Web search arena is dominated by a few players offering gigantic systems capable of worldwide crawling and indexing Web pages, challenging research problems stem from the need of integrating data sources forming the so-called "deep Web"; this information is available through data collections which are often hardly accessible, requiring data extraction and integration both for understanding their semantics and for mastering their interplay. By studying issues and challenges involved with structured integration of Web data sources, the SeCo project has the ambitious goal of lowering the technological barrier required for building complex search applications, thereby enabling the development of many new applications, covering relevant search needs.

The project is now in the fourth of a five-year lifespan (Nov. 2008-Nov. 2013); during its third year, the project has broadened the spectrum of considered sources, by including structured tables and by including humans and social networks as a new, unconventional, but increasingly relevant source of information. We have also focused on new interaction paradigms, by adding natural language queries to exploratory queries, and by developing techniques for clustering and diversification of results, so as to broaden the spectrum of user experiences. Owing to the high cost of accessing data sources (and to other difficulties in managing them, such as access limitations and lack of stability), we have also considered the need of integrating search with a data materialization system that can produce local copies of most frequently used data.

This is the third book in the Search Computing Series; while the first two books reported the results of SeCo Workshops held in Como in 2009 and 2010, in this book we collect 16 articles which in most cases were contributed to several workshops during 2011, organized by members of the Search Computing project in the context of major international conferences: ExploreWeb at ICWE 2011 (co-chaired by Alessandro Bozzon and Marco Brambilla), Very Large Data Search at VLDB 2011 (co-chaired by Marco Brambilla and Stefano Ceri), DBRank also at VLDB 2011 (co-chaired by Davide Martinenghi), DATAVIEW at ECOWS 2011 (co-chaired by Alessandro Bozzon and Maristella Matera), and OrdRing at ISWC 2011 (co-chaired by Emanuele Della Valle and Alessandro Bozzon).

Articles were selected, extended, and revised in the first semester of 2012, so as to build a rather cohesive set of contributions; they are clustered in four parts according to their thematic similarity.

- Part 1: Extraction and Integration
- Part 2: Query and Visualization Paradigms

- Part 3: Exploring Linked Data
- Part 4: Games, Social Search, and Economics

The first part collects articles dealing with the problem of extracting and integrating data from heterogeneous sources. The first paper, by Blanco et al., addresses the issue of extracting the best estimate of data values which have several replicas on the Web based on extracting all the existing values and then applying a Bayesian model to them, thereby overcoming the uncertainty of accessing a given copy. The second paper, by Mulwad et al., describes the problem of extracting semantics from structured sources in very general terms, by contributing a classification of current tools for data extraction and then describing a specific tool for data extractions which turns relational data into the RDF format. The next two papers are focused on the extraction of semantic information from tables whose content is potentially very useful but whose schema and documentation are missing; in the third paper, by Unbehauen et al., the objective is to connect content to linked data; in the fourth paper, by Brambilla et al., the objective is to extract those tables which are similar both for their schema and for their content, so as to integrate them either through union or through joins. Finally, the last paper, by Bozzon et al., describes the features of a data materialization system – part of the SeCo framework that produces a local copy of content from a frequently queried remote source, thereby providing an active cash; specifically, the paper focuses on seeding the materialization system with data which can be progressively retrieved from the sources.

The second part of the book addresses new paradigms for expressing search queries or for managing query results, in a way that can be most expressive for final users. The first paper, by Guerrisi et al., describes a natural language interface – also part of the SeCo framework—which is capable of understanding complex queries upon multiple domains of interests, by employing rule-based and machine learning methods. The second paper, by Aral et al., discusses the use of mobile interfaces for exploratory Web searches. The third paper, by Brambilla et al., deals with how multi-dimensional data should be clustered together and labeled semantically so as to improve the user's capability of "reading" useful information from results. The fourth paper, by Morales-Chaparro et al., deals with visualization of search results under very different requirements and challenges offered by displays of different format and technology, by presenting a high-level model-driven approach to the development of visualization interfaces.

The third part of the book deals with exploration of semantic data sources presented as linked data. Linked data are clearly the most powerful solution that has been produced by the Semantic Web community for solving data integration and interoperability, hence the exploration of linked data is emerging as a new topic of research. The paper by Bozzon et al. presents how the Sparql query language for accessing RDF sources can be empowered by making ranking a first-class construct, thereby offering a query search paradigm for enabling the use of RDF sources. The paper by Castano et al. offers a method for the thematic exploration of linked data which aims at clustering them, thereby turning messy data into organized collections that expose strong internal cohesion. The paper

of Cohen et al. provides an example of a tool for exploring linked data helping users to explore linked data and also to reuse queries that were previously issued, thereby building a personalized collection of queries over linked data repositories.

Finally, the fourth part of the book deals with emerging paradigms for search which deal with social aspects. The first paper, by Cohen et al., studies proximity measures in social networks, which are at the basis for solving problems such as user-centric person search. The second paper, by Bozzon et al., dwells on social search, by describing an architecture that can span a search query of known format to people who are reachable through their social platform; the resulting Crowdsearcher system is an extension of the SeCo framework for alternating exploratory and crowd-based search. The third paper, by Hees et al., proposes a game-based method for associating linked data with popularity, as a measure of its relevance and strength that could then be used for directing exploratory search over linked data. Finally, the paper by Brambilla et al. addresses the problem of economic sustainability of complex search by studying how an ecosystem of application and data source providers could share a revenue system that satisfies the heterogeneity of players and revenue redistribution among them.

The book is the result of collective efforts of many participants of the SeCo project and of a variety of contributors we have met in the context of five workshops. All of them have provided very useful insights on search computing problems and issues. The chapters have been reviewed by several experts. We would like to thank them all for their efforts.

August 2012

Stefano Ceri
Marco Brambilla

Reviewers

Pierre Andrews	University of York, UK
Sören Auer	Universität Leipzig, Germany
Alessandro Bozzon	Politecnico di Milano, Italy
Emanuele Della Valle	Politecnico di Milano, Italy
Alfio Ferrara	Università degli Studi di Milano, Italy
Piero Fraternali	Politecnico di Milano, Italy
Paolo Papotti	Università Roma Tre, Italy
Silvia Quarteroni	Politecnico di Milano, Italy
Sebastian Stein	University of Southampton, UK
Carmen Vaca	Politecnico di Milano, Italy

Table of Contents

Part 1: Extraction and Integration

Part 2: Query and Visualization Paradigms

Part 3: Exploring Linked Data

Part 4: Games, Social Search and Economics