Lecture Notes in Computer Science 6585

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

Stefano Ceri Marco Brambilla (Eds.)

Search Computing

Trends and Developments

Springer

Volume Editors

Stefano Ceri
Marco Brambilla
Politecnico di Milano
Dipartimento di Elettronica e Informazione
P.za L. Da Vinci, 32, 20133 Milano, Italy
E-mail: {stefano.ceri, marco.brambilla}@polimi.it

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-19667-6 e-ISBN 978-3-642-19668-3
DOI 10.1007/978-3-642-19668-3
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011922270

CR Subject Classification (1998): H.4, H.3, D.4, C.2.4, F.2, D.1.3

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

Searching for information is perhaps the most important application of today's computing systems. In the new century, all the World's citizens have become accustomed to thinking of the Web as the source for answering their information needs, and search engines as their Web interface. Websites reporting a movie's plot, tomorrow's weather in our next destination, the risks of a surgical procedure, the fastest route to a friend's house, the video of the last opera at La Scala can all be found as result of a keyword search. If any Web page in the world stores the answer to our information need, then we expect the search engine to link that page and describe it through a snippet appearing in the first page of the search results.

A few search engine companies are able to meet such expectations, and completely cover the search engine market. However, offering a link to a Web page does not cover all information needs. Many problems cannot be solved by simple keyword-based queries. The notion of "best page" for solving a given problem is typically inadequate when the problem requires solutions spanning over multiple pages. We are indeed accustomed to using a variety of Web resources to solve our problems: while the search engine hints to useful information, the user's brain is the fundamental platform for information integration.

Problems such as "who is the best doctor to cure insomnia in a nearby hospital" can be solved by using the Web multiple times, searching for partial results. Once a hospital's website is located and the listing of its doctors is extracted, one can find out that there is one doctor in the list who has published recent papers on insomnia. While doing so, the user is performing information integration in her brain; specifically, she is applying ranking while extracting hospitals based on proximity and doctors based on their publications on insomnia, then matching on the basis of doctor names. Of course, the best way to build the matching is to use the search engine itself, by entering doctors' names as keywords, extracted from either the hospital search or the literature search, but then the search is less focused and result interpretation is more difficult.

Complex queries are supported in certain domains, such as travels, hotel booking, and book purchasing, by specialized, domain-specific search systems or search engine integrators. It is important to assess how travel assistants solve the problem: they offer a few predefined queries to build the itinerary, then offer additional services (e.g., car rentals, hotels, local events, insurance) so as to complete the plan around the itinerary. Thus, they perform specialized steps of integration by substituting the user's brain, and then let the user enrich the solution incrementally and interactively, with customized interfaces. In other words, they solve complex queries in the context of given domains, which are

supported by a substantial business; such specialized search systems dominate over general purpose ones in their domain of expertise, and therefore attract users.

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. The project is now entering the third of a five-year lifespan (November 2008 – November 2013); it proposes a new paradigm for solving complex queries based on combining data extraction from distinct sources and data integration by means of specialized integration engines. Data extraction retrieves data from different sources, ordered based on local rankings, and data integration merges such results into result combinations, with an associated global ranking, such that combinations with the highest ranking are produced as fast as possible; a result combination represents the solution of a complex search problem. Thus, the search computing project has the ambitious goal of lowering the technological barrier required for building complex search applications, thereby enabling the development of many new applications which will cover relevant search needs.

Search computing covers many research directions, which are all required in order to provide an overall solution to a complex search. The core of the project is the technology for search service integration, which requires both theoretical investigation and engineering of efficient technological solutions. The core theory concerns the development of result integration methods that not only denote "top-k optimality," but also the need of dealing with proximity, approximation, and uncertainty. Such a theory is supported by an open, extensible and scalable architecture for computing queries over data services, designed so as to incorporate the project's results by adding new operations, by encoding new join methods, and by injecting new features dealing with incremental evaluation and adaptivity.

A number of further research dimensions complement such core. Formulation of a complex query and browsing over solutions is a complex cognitive task, whose intrinsic difficulty has to be lowered as much as possible so as to meet usability requirements. Therefore, we are investing a consistent effort in the development of user-friendly interfaces which are targeted at assisting users in expressing their needs and then browsing on results. Solving a complex problem requires supporting users in the interactive and incremental design of their queries, thereby assisting search as a long-term process for exploring the solution space; result differences can be better appreciated by visualizing results (e.g., through maps or timelines). The project success also depends on the ability of registering new sources and making them available for solving complex problems; therefore, we have designed abstractions, architectural solutions, and model-driven design tools for service registration and for application development, aiming at assisting service publishing, application design, and query execution tuning. While the current description of Web resources is very simple, so as to enable an equally simple description of Web interactions, we aim at linking the service description

to ontological sources, so as to enable high-level expressive interfaces covering the gap from high-level interactions to query expression.

While focusing on technological dimensions, we are also investigating crucial aspects to the project success, such as the business models and user involvement in the design process through user-centered design. We are additionally investigating the use of search computing for scientific applications, such as supporting bio-informatics research by enabling the access to genetic and proteomic data sources.

This book reports the proceedings of the workshop "New Trends in Search Computing," held in Como and Milan during May 25–31, 2010, as the follow-up of the workshop "Search Computing Challenges and Directions," also published by Springer in 2010 (LNCS 5950).

The workshop was divided into eight independent sessions, reflecting the many research directions of the project. It was held during five consecutive days, in Milan and Como, with about 60 participants equally divided between SeCo researchers and international experts. Each workshop session had editors, chosen within the SeCo research team, and external experts, who provided their tangible contribution to the project with feedback, advice, and contributed chapters. Session editors helped us in organizing the book's design, by interacting with the session experts and by shaping up each session and the corresponding book part.

Each part of the book reports the result of a workshop session; it includes one chapter describing the search computing approach to the problem, and one or more additional chapters reflecting the contribution and viewpoints of experts that participated in the workshop, broadening the spectrum of investigations which are currently ongoing in search computing. In some cases, the part is closed by a short chapter reporting different opinions that the workshop participants discussed at panels which closed the corresponding session.

The book is the result of the collective effort of all the project participants and has been reviewed by several experts. We would like to thank all of them for their efforts.

January 2011

Marco Brambilla
Stefano Ceri

Organization

Reviewers

Carlo Batini	Università degli Studi di Milano-Bicocca
Jordi Cabot	INRIA - École des Mines de Nantes
Andrea Calì	University of Oxford
Alex Komoroske	Google, Inc.
Rodrigo Lopez	European Bioinformatics Institute
Ioana Manolescu	INRIA - Université de Paris Sud
Massimo Paolucci	DOCOMO Euro-Labs
Alfonso Valencia	Spanish National Cancer Research Centre
Roberto Verganti	Politecnico di Milano
Gerhard Weikum	Max Planck Institute for Informatics

Part Editors

Part 1: The Search Process	Marco Brambilla and Stefano Ceri
Part 2: Interaction Design	Tiziana Catarci and Maristella Matera
Part 3: Semantic Description	Alessandro Campi and Davide Eynard
Part 4: Rank-Join	Davide Martinenghi and Marco Tagliasacchi
Part 5: Query Processing	Daniele Braga and Michael Grossniklaus
Part 6: Tools and Mashups	Marco Brambilla and Alessandro Bozzon
Part 7: BioSeco	Marco Masseroli
Part 8: Sustainable Exploitation	Emanuele Della Valle

Sponsoring Institutions

The Search Computing (Seco) Project is funded by the European Research Council (ERC), responding to the 2008 Call for "IDEAS Advanced Grants," a program dedicated to the support of investigation-driven frontier research. SeCo started on November 1, 2008 and will last until October 31, 2013.

Table of Contents

Part 1: The Search Process

Part 2: Interaction Design

Part 3: Semantic Description

Part 4: Rank Join

Part 5: Query Processing

Part 6: Tools and Mashups

Part 7: Bio-SeCo

Part 8: Towards a Sustainable Exploitation