Preface

Jorge Cardoso

SAP Research CEC, SAP AG,
Chemnitzer Strasse 48, 01187 Dresden, Germany
E-mail: jorge.cardoso@sap.com

Christoph Bussler

BEA Systems, Inc.,
475 Sansome Street, San Francisco, CA 94111, USA
E-mail: chbussler@aol.com

Francesco Guerra

Dipartimento di Economia Aziendale,
Universita' di Modena e Reggio Emilia,
via le Berengario 51, 41100 Modena, Italy
E-mail: guerra.francesco@unimore.it

Biographical notes: Jorge Cardoso (http://www.dme.uma.pt/jcardoso) joined SAP Research (Germany) in 2007. He previously gave lectures at the University of Madeira, the University of Georgia and at the Instituto Politécnico de Leiria. He received his PhD in Computer Science from the University of Georgia in 2002. He worked at the Boeing Company on enterprise application integration and at CCG, Zentrum fur Graphische Datenverarbeitung on computer supported cooperative work. He has organised several international conferences on semantic web and information systems, and has published several refereed papers and edited several books in the areas of workflow management systems, semantic web, and related fields.

Christoph Bussler (http://hometown.aol.com/chbussler) is Senior Staff Software Engineer at Merced Systems, Inc. His interests include workflow and process management, B2B and EAI integration, and semantic computing. He is author of several books and journal articles on integration and semantics. He is active in the professional community as keynote speaker, conference and workshop organiser as well as program committee member. He has a PhD in Computer Science from the University of Erlangen, Germany, and worked in several organisation, including BEA, Cisco Systems, Digital Enterprise Research Institute, Oracle, The Boeing Company and Digital Equipment Corporation.

Francesco Guerra (http://www.dbgroup.unimo.it/~guerra/) is an Assistant Professor in Computer Engineering at the Faculty of Economics of the University of Modena and Reggio Emilia, where he teaches enterprise information systems. His main research interests include integration of heterogeneous information sources, ontologies, and the semantic web. He has participated in several Italian and European projects. At present, he is involved in the Italian FIRB project NEP4B: Networked Peers for Business (years 2006–2008), and in the European FP6 STREP project STASIS: Software for Ambient Semantic Interoperable Services (2006–2008). He has a PhD in information engineering from the University of Modena and Reggio Emilia.

Traditional search techniques establish a direct connection between the information provided by users with the search engine. Users are only allowed to specify a set of keywords that will be syntactically matched against a database of keywords and references. This simple approach has several drawbacks since it gives rise to a low precision (the ratio of positive results with respect to the total number of false and positive results retrieved) and low recall (the ratio of positive results retrieved with respect to the total number of positive results in the reference base). Many factors influence this low precision and recall, namely polysemy and synonymy. In the first case, one word specified in a query might have several meanings and, in the second case, distinct words may designate the same concept. If appropriate strategies are used and included in a new generation of search engines, the number of false results can be drastically reduced. As a result, the impact of these two degrading factors can be reduced and even eliminated.

As the interconnection of research areas such as artificial intelligence, semantic web, and linguistics becomes stronger and more mature, it is reasonable to explore how better search engines can be developed to more adequately respond to users' needs. A new kind of search engine that has been explored for a few years now has been termed "semantic-based search engines" by many researchers. The underlying paradigm of these engines is to find resources based on similar concepts and logical relationships and not just similar words. These engines typically rely on the use of metadata, controlled vocabularies, thesauri, taxonomy, and ontologies to describe the searchable resources to ensure that the most relevant items of information are returned.

The intend of this special issue is to bring together a compilation of recent research and developments toward the creation of a new paradigm for search engines that relies on metadata, semantics and ontologies, by providing readers with a "broad spectrum vision" of the most important issues on semantic search engines.

One of the main problems concerns the recognition of items of interest in web documents. The first three papers face this issue following three different perspectives. In the first paper, Xu and Embey apply techniques from data extraction, information retrieval and machine learning for achieving this goal. Their approach is based on an extraction ontology describing the information of interest for users that is exploited for extracting data from web documents. The application of several heuristics on the extracted data provides some statistical measures that are exploited by means of machine-learned rules for determining the documents containing relevant information.

The second paper presents the CORE module, which enables incremental searching based on co-occurrence of entities – as well as ranking – tracking trends and popularity timelines. CORE extends the KIM platform for semantic annotation, indexing and retrieval, which provides an infrastructure for the automatic extraction of named entity reference and descriptions from text documents. CORE introduces the concept of context (block of content) and evaluates associative relations between entities on the basis of their frequency in a specific context. Moreover, CORE can track trends in time for a set of entities based on their occurrence and co-occurrence frequency enabling the popularity rank of these in specific period of times. The user interface giving access to the CORE functionality is described with details as well as some interesting types of user search.

In the third paper, Sindice is presented, a lookup index over resources crawled on the semantic web to locate semantic web data sources. Sindice collects RDF documents and indexes these on resource URIs, inverse functional properties and keywords. By means of a web front-end and a public API, it is possible to look up resources and to search full-text descriptions obtaining as a result the URL of sources where the resources occur.

In several approaches, thesaurus-based indexes improve the result of document retrieval, in particular by solving the problems of synonyms and allowing the disambiguation of homonyms. Owing to the large amount of documents, these approaches often rely on automatic techniques for annotating documents with terms from thesauri. Eckert et al. claim that the quality of the thesaurus used as a basis for indexing ensure the quality of the whole indexing process. In the forth paper, the authors present and evaluate a method combining the application of statistics and appropriate visualisation techniques supporting the detection of potential problem in a thesaurus.

The fifth paper addresses another interesting problem, i.e., the expressive power of query languages used with current search engine. Schellhase and Lukasiewicz face this issue by proposing a search query paradigm developed for literature searches. Their approach exploits the same metadata about research publications, authors, organisations and scientific events used by other scientific search engines, but it provides a query language based on description logics and variable-strength conditional preferences. The theoretical foundation of their language is domain-independent and thus it may be adapted to other areas.

Finally, in the last paper, Battré introduces an optimisation approach for RDF stores based on distributed hash tables that cache and reuse intermediate results created in previous queries in order to allow a quick processing of new queries. The paper includes an evaluation section where different caching strategies are compared.

Acknowledgement to reviewers

Dean Allemang; Francesco Bellomi; Abraham Bernstein; Omar Boucelma; Paolo Bouquet; Patrick Brezillon; Andrea Calì; Steve Cayzer; Oscar Corcho; Matteo Cristani; Emanuele Della Valle; Martin Dzbor; Alfio Ferrara; Doug Foxvog; Rosa Galli; Stefan Grimm; Mohand-Said Hacid; Hyoil Han; Andreas Harth; Jeff Heflin; Vipul Kashyap; Atanas Kiryakov; Claus-Peter Klas; Ruben Lara; Domenico Lembo; Frank Leymann; Pasquale Lops; Federica Mandreoli; Mihhail Matskin; Andrea Maurino; Brian McBride; Gregoris Mentzas; Gianluca Moro; Lyndon Nixon; Matteo Palmonari; Massimo Paolucci; Dimitris Plexousakis; Axel Polleres; Christoph Quix; Evangelos D. Sakkopoulos; Giovanni Semeraro; Amit Sheth; Pavel Shvaiko; Kiril Ivanov Simov; Sergej Sizov; Peter Spyns; Giorgos Stamou; Armando Stellato; Heiner Stuckenschmidt; Gerd Stumme; Ludger van Elst; Roland Wagner; Benjamin Yen; Ilya Zaihrayeu.