Lecture Notes in Computer Science 10190

More information about this series at http://www.springer.com/series/8851

Ngoc Thanh Nguyen · Ryszard Kowalczyk
Alexandre Miguel Pinto · Jorge Cardoso (Eds.)

Transactions on Computational Collective Intelligence XXVI

Springer

Editors-in-Chief

Ngoc Thanh Nguyen
Institute of Informatics
Wroclaw University of Technology
Wroclaw
Poland

Ryszard Kowalczyk
Swinburne University of Technology
Hawthorn, SA
Australia

Guest Editors

Alexandre Miguel Pinto
University of Lisbon
Lisbon
Portugal

Jorge Cardoso
Huawei German Research Center
Munich
Germany

and

University of Coimbra
Coimbra
Portugal

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISSN 2190-9288 ISSN 2511-6053 (electronic)
Transactions on Computational Collective Intelligence
ISBN 978-3-319-59267-1 ISBN 978-3-319-59268-8 (eBook)
DOI 10.1007/978-3-319-59268-8

Library of Congress Control Number: 2017942988

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Transactions on Computational Collective Intelligence XXVI

Preface

It is our pleasure to present the XXVI volume of the LNCS *Transactions on Computational Collective Intelligence*. This Special Issue is the compilation of selected papers of the First International KEYSTONE Conference 2015 (IKC 2015), part of the Keystone COST Action IC1302 (www.keystone-cost.eu). COST (European Cooperation in Science and Technology – www.cost.eu) is a pan-European intergovernmental framework. Its mission is to enable breakthrough scientific and technological developments leading to new concepts and products and thereby contribute to strengthening Europe's research and innovation capacities. It allows researchers, engineers, and scholars to jointly develop their own ideas and take new initiatives across all fields of science and technology, while promoting multi- and interdisciplinary approaches. COST aims at fostering a better integration of countries that are less research-intensive to the knowledge hubs of the European research area. The COST Association, an international not-for-profit association under the Belgian law, integrates all management, governing, and administrative functions necessary for the operation of the framework. The COST Association currently has 36 member countries.

This volume collects and analyzes the main results achieved by the research areas covered by KEYSTONE (the Action: *S*emantic *Key*word-Based Search on S*t*ructured data S*o*urc*e*s). For Action members, the conference was also the place to discuss the results obtained during the first two years of activities. The research theme of IKC 2015 was "Keyword-Search on Massive Datasets." It is an emerging and challenging theme. In particular, since large-scale data sources usually comprise very large schemas and billions of instances, keyword search over such datasets face several challenges related to scalability and interpretation of the keyword query intended meaning. Whereas state-of-the-art keyword search techniques work well for small or medium-size databases in a particular domain, many of them fail to scale on heterogeneous databases that are composed of thousands of instances. The discovery of semantically related data sources is another critical issue, hindered by the lack of sufficient information on available datasets and endpoints. Browsing and searching for data at this scale is not an easy task for users. Semantic search can support the process aiming at leveraging semantics to improve the accuracy and recall of search mechanisms.

This volume inaugurates the year 2017, the seventh year of TCCI activities. In the past 25 issues, we have published 253 high-quality papers. This issue contains 10 papers.

In the first paper "Professional Collaborative Information Seeking: Towards Traceable Search and Creative Sensemaking," Andreas Nuernberger et al. propose an adapted model for professional collaborative information seeking. The authors also introduce a system that has been specifically developed to support collaborative technology search.

The second paper entitled “Exploiting Linguistic Analysis on URLs for Recommending Web Pages: A Comparative Study” by Sara Cadegnani et al. analyzes and compares three different approaches to leverage information embedded in the structure of websites and the logs of their web servers to improve the effectiveness of web page recommendation. Their proposals exploit the context of users’ navigations, i.e., their current sessions when surfing a specific website. These approaches do not require either information about the personal preferences of the users to be stored and processed or complex structures to be created and maintained.

In the third paper, “Large-Scale Knowledge Matching with Balanced Efficiency-Effectiveness Using LSH Forest” by Michael Cochez et al., the authors investigate the use of LSH Forest (a self-tuning indexing schema based on locality-sensitive hashing) for solving the problem of placing new knowledge tokens in the right contexts of the environment. They argue and show experimentally that LSH Forest possesses the required properties and could be used for large distributed set-ups. Further, they show experimentally that for their type of data minhashing works better than random hyperplane hashing.

The fourth paper, “Keyword-Based Search of Workflow Fragments and Their Composition” by Khalid Belhajjame et al., presents a method for identifying fragments that are frequently used across workflows in existing repositories, and therefore are likely to incarnate patterns that can be reused in new workflows. They present a keyword-based search method for identifying the fragments that are relevant for the needs of a given workflow designer. They go on to present an algorithm for composing the retrieved fragments with the initial (incomplete) workflow that the user designed based on compatibility rules that they identified, and showcase how the algorithm operates using an example from eScience.

The fifth paper, entitled “Scientific Footprints in Digital Libraries” by Claudia Ifrim et al., analyzes citation lists to not only quantify but also understand impact by tracing the “footprints” that authors have left, i.e., the specific areas in which they have made an impact. They use the publication medium (specific journal or conference) to identify the thematic scope of each paper and feed from existing digital libraries that index scientific activity, namely, Google Scholar and DBLP. This allows them to design and develop a system, the Footprint Analyzer, which can be used to successfully identify the most prominent works and authors for each scientific field, regardless of whether their own research is limited to or even focused on the specific field. Various real-life examples demonstrate the proposed concepts, and results from the developed system’s operation prove the applicability and validity.

In the sixth paper titled “Mining and Using Key-words and Key-phrases to Identify the Era of an Anonymous Text,” Dror Mughaz et al. determine the time frame in which the author of a given document lived. The documents are rabbinic documents written in Hebrew-Aramaic languages. The documents are undated and do not contain a bibliographic section, which constitutes a substantial challenge. The authors define a set of key phrases and formulate various types of rules – “Iron-clad,” Heuristic, and Greedy – to define the time frame. These rules were tested on two corpora containing response documents, and the results are promising. They are better for larger corpora than for smaller corpora.

The next paper, “Toward Optimized Multimodal Concept Indexing” by Navid Rekabsaz et al., presents an approach for semantic-based keyword search and focuses especially on its optimization to scale to real-world-sized collections in the social media domain. Furthermore, the paper presents a faceted indexing framework and architecture that relates content to semantic concepts to be indexed and searched semantically. The authors study the use of textual concepts in a social media domain and observe a significant improvement from using a concept-based solution for keyword searching.

In the eighth paper, entitled “Improving Document Retrieval in Large-Domain Specific Textual Databases Using Lexical Resources,” Ranka Stanković et al. propose the use of document indexing as a possible solution to document representation. They use metadata for generating a bag of words for each document with the aid of morphological dictionaries and transducers. A combination of several tf-idf-based measures was applied for selecting and ranking of retrieval results of indexed documents for a specific query and the results were compared with the initial retrieval system that was already in place. In general, a significant improvement has been achieved according to the standard information retrieval performance measures, where the InQuery method performed the best.

In the ninth paper, “Domain-Specific Modeling: A Food and Drink Gazetteer,” Andrey Tagarev et al. build a food and drink (FD) gazetteer for classification of general, FD-related concepts, efficient faceted search or automated semantic enrichment. For general domains (such as the FD domain), re-using encyclopedic knowledge bases like Wikipedia may be a good idea. The authors propose a semi-supervised approach that uses a restricted Wikipedia as a base for the modeling, achieved by selecting a domain-relevant Wikipedia category as root for the model and all its subcategories, combined with expert and data-driven pruning of irrelevant categories.

The last paper, “What’s New? Analyzing Language-Specific Wikipedia Entity Contexts to Support Entity-Centric News Retrieval” authored by Yiwei Zhou et al., focuses on the problem of creating language-specific entity contexts to support entity-centric, language-specific information retrieval applications. First, they discuss alternative ways such contexts can be built, including graph-based and article-based approaches. Second, they analyze the similarities and the differences in these contexts in a case study including 220 entities and five Wikipedia language editions. Third, they propose a context-based entity-centric information retrieval model that maps documents to aspect space, and apply language-specific entity contexts to perform query expansion. Last, they perform a case study to demonstrate the impact of this model in a news retrieval application. The study illustrates that the proposed model can effectively improve the recall of entity-centric information retrieval while keeping high precision and can provide language-specific results.

We would like to thank all the authors for their valuable contributions to this issue and all the reviewers for their opinions, which contributed greatly to the high quality of the papers. Our special thanks go to the team at Springer, who have helped to publish the many TCCI issues in due time and in good order.

February 2017

Alexandre Miguel Pinto
Jorge Cardoso

Transactions on Computational Collective Intelligence

This Springer journal focuses on research in computer-based methods of computational collective intelligence (CCI) and their applications in a wide range of fields such as the Semantic Web, social networks, and multi-agent systems. It aims to provide a forum for the presentation of scientific research and technological achievements accomplished by the international community.

The topics addressed by this journal include all solutions to real-life problems for which it is necessary to use computational collective intelligence technologies to achieve effective results. The emphasis of the papers published is on novel and original research and technological advancements. Special features on specific topics are welcome.

Contents