Studies in Computational Intelligence 439

For further volumes:
http://www.springer.com/series/7092

Cristian Lai, Giovanni Semeraro,
and Eloisa Vargiu (Eds.)

New Challenges in Distributed Information Filtering and Retrieval

DART 2011: Revised and Invited Papers

Springer

Editors

Cristian Lai
CRS4, Center of Advanced Studies
Research and Development in Sardinia
Parco Scientifico e Tecnologico
della Sardegna
Pula
Italy

Eloisa Vargiu
Department of Electrical and
Electronic Engineering
University of Cagliari
Cagliari
Italy

Giovanni Semeraro
Department of Informatics
University of Bari “Aldo Moro"
Bari
Italy

ISSN 1860-949X
ISBN 978-3-642-31545-9
DOI 10.1007/978-3-642-31546-6
Springer Heidelberg New York Dordrecht London

e-ISSN 1860-9503
e-ISBN 978-3-642-31546-6

Library of Congress Control Number: 2012940861

Preface

Information filtering (IF) has drastically changed the way information seekers find relevant results. Information seekers effectively prune large information spaces and help users in selecting items that best meet their needs, interests, preferences and tastes. These systems rely strongly on the use of various machine learning tools and algorithms for learning how to rank items and predict user evaluation. Information Retrieval (IR), on the other hand, attempts to address similar filtering and ranking problems for pieces of information such as links, pages, and documents. IR systems generally focus on the development of global retrieval techniques, often neglecting individual user needs and preferences. This book focuses on new challenges in distributed Information Filtering and Retrieval. It collects invited chapters and research contributions from the DART 2011 Workshop, held in Palermo (Italy) and co-located with the *XII International Conference of the Italian Association on Artificial Intelligence*. DART aimed to investigate novel systems and tools to distributed scenarios and environments. Therefore, DART contributed to discuss and compare suitable novel solutions based on intelligent techniques and applied in real-world applications.

Chapter 1, *A Brief Account on the Recent Advances in the Use of Quantum Mechanics for Information Retrieval*, is an extended contribution by Massimo Melucci, invited speaker at DART. The chapter focuses on the attention and research that have been paid to the exploitation of Quantum Mechanics in IR. It introduces the basic notions of Quantum IR and explains how the retrieval decision problem can be formulated within a quantum probability framework in terms of vector subspaces rather than in terms of subsets as it is customary to state in classical probabilistic IR. Hence it shows that ranking by quantum probability of relevance in principle yields higher expected recall than ranking by classical probability at every level of expected fallout and when the parameters are estimated as accurately as possible on the basis of the available data.

In Chapter 2, *Cold Start Problem: a Lightweight Approach*, the SWAPTeam1 participation at the ECML/PKDD 2011 - Discovery Challenge for the task on the cold start problem focused on making recommendations for new video lectures.

The Challenge organizers encouraged solutions that can actually affect VideoLecture.net. The main contribution concerns about the compromise between recommendation accuracy and scalability performance of proposed approach. For facing the cold start problem, authors developed a solution that uses a content-based approach. Such an approach is less sensitive to the cold start problem that is commonly associated with pure collaborative filtering recommenders. The surrounding idea for the proposed solution is that providing recommendations about cold items remains a chancy task, thus a computational resource curtailment for such task is a reasonable strategy to control performance trade-off of a day-to-day running system.

Chapter 3, *Content-based Keywords Extraction and Automatic Advertisement Associations to Multimodal News Aggregations*, focuses on multimodal news aggregation retrieval and fusion. In particular, authors tackle two main issues: extracting relevant keywords to news and news aggregations, and automatically associating suitable advertisements to aggregated data. To achieve the first goal, authors propose a solution based on the adoption of extraction-based text summarization techniques; whereas to achieve the second goal, they developed a contextual advertising system that works on multimodal aggregated data. The proposed solutions have been assessed on Italian news aggregations and tested with suitable baseline solutions.

In Chapter 4, *ImageHunter: a novel tool for Relevance Feedback in Content Based Image Retrieval*, a Content-Based Image Retrieval (CBIR) engine, called ImageHunter, is thoroughly described. ImageHunter aims at providing users, especially unskilled ones, with an effective tool for image search, classification and retrieval within digital archives, photo-sharing web sites and social networks. The novelty of the approach, with respect to canonical techniques based on metadata associated to images, lies in the combination of content-based analysis with feedbacks from the users (Relevance Feedback). In addition, the modular structure permits the integration of ImageHunter in web-based applications as well as in standalone ones.

Chapter 5, *Temporal Characterization of the Requests to Wikipedia*, presents an empirical study about the temporal patterns characterizing the requests submitted by users to Wikipedia. The study is based on the analysis of the log lines registered by the Wikimedia Foundation Squid servers after having sent the appropriate content in response to users requests. The analysis has been conducted regarding the ten most visited editions of Wikipedia and has involved more than 14,000 million log lines corresponding to the traffic of the entire year 2009. The conducted methodology has mainly consisted in the parsing and filtering of users requests according to the study directives. As a result, relevant information fields have been finally stored in a database for persistence and further characterization.

In Chapter 6, *Interaction Mining: the new frontier of Customer Interaction Analytics*, authors present a solution for argumentative analysis of call center conversations. Their challenge is to provide useful insights for enhancing customer interaction analytics to a level that will enable more qualitative metrics and key performance indicators beyond the standard approach currently used. These metrics rely on understanding the dynamics of conversations by highlighting the way participants discuss about topics. In so doing, relevant situations (such as social behaviors,

controversial topics and customer oriented behaviors) could be detected. Moreover, customer satisfaction may be predicted.

Chapter 7, *A Linguistic Approach to Opinion Mining*, proposes an automatic linguistic approach to opinion mining. The proposed solution relies on a semantic analysis of textual resources and based on FreeWordNet, a new developed linguistic resource. FreeWordNet has been defined by the enrichment of the meanings expressed by adjectives and adverbs in WordNet with a set of properties and the polarity orientation. Reviews are used every day by common people or by companies who need to make decisions. Such amount of social data can be used to analyze the present and to predict the near future needs or the probable changes. Mining the opinions and the comments is a way to extract knowledge by previous experiences and by the feedback received.

Chapter 8, *Sentiment Analysis in the Planet Art: a Case Study in the Social Semantic Web*, set in a Social Semantic Web framework, explores the possibility of extracting rich, emotional semantic information from the tags freely associated to digitalized visual artworks, identifying the prevalent emotions that are captured by the tags. To this end, authors rely on ArsEmotica, an application software that combines an ontology of emotional concepts with available computational and sentiment lexicons. The Chapter reports and comments also results of a user study, aimed at validating the outcomes of ArsEmotica. Those results were obtained by involving the users of the same community which tagged the artworks.

In Chapter 9, *OntoTimeFL - A Formalism for Temporal Annotation and Reasoning for Natural Language Text*, an ontological formalism for annotating complex events expressed in natural language is defined. Compared to TimeFL, OntoTimeFL introduces new constructs in form of concepts for the annotation of three types of complex events: narrative, intentional, and causal events. In addition, the methodological choice of defining OntoTimeFL as a conceptualization of TimeFL makes easier the processes of automated annotation and of reuse and application of several types of axioms and rules for temporal reasoning to the annotated items.

Chapter 10, *Representing Non Classical Concepts in Formal Ontologies: Prototypes and Exemplars*, focuses on concept representation, an open problem in the field of ontology engineering and knowledge representation. Authors review empirical evidence from cognitive psychology, which suggests that concept representation is not an unitary phenomenon. In particular, they found that human beings employ both prototype and exemplar based representations in order to represent non classical concepts. Thus, authors suggest that a similar, hybrid prototype-exemplar based approach could be useful also in the field of formal ontology technology.

Chapter 11, *From Logical Forms to SPARQL Query with GETARUNS*, presents a Question Answering (QA) tool which integrates a full-fledged NLP system for text understanding, called GETARUNS. This system deals with different levels of syntactic and semantic ambiguity and generates some structures, called Logical Forms, by accessing computational lexical equipped with sub-categorization frames and appropriate selectional restrictions applied to the attachment of complements and adjuncts. Logical Forms are exploited by the QA system to compute a prospective

answer and to extract the semantic elements needed to produce a SPARQL expression that is then used to query Linked Open Data endpoints.

Chapter 12, *A DHT-based Multi-Agent System for Semantic Information Sharing*, presents AOIS (Agents and Ontology based Information Sharing), a multi-agent system that supports the sharing of information among a dynamic community of users connected through BitTorrent, the well-known peer-to-peer platform. Compared to web scale search engines, AOIS enhances the search through domain ontologies, avoids the burden of publishing the information on the Web and guarantees a controlled and dynamic access to the information. Agent technologies are exploited to filter information coming from different users on the ground of the user previous experience, to propose new information that can be potentially interesting for a user in a push modality as well as to delegate access capabilities on basis of a reputation network built by the agents of the system on the user community.

Chapter 13, *A Decisional Multi-Agent Framework for Automatic Supply Chain Arrangement*, proposes a multi-agent system for supply chain dynamic configuration. Agent brain is composed of a Bayesian decision network, thus allowing agent to take the best decisions by estimating benefits and potential risks of different strategies, analyzing and managing uncertain information about the collaborating companies. In so doing, each agent collects information about customers orders and current market prices and is able to analyze previous experiences of collaborations with trading partners. Therefore, the agent performs a probabilistic inferential reasoning to filter information modeled in its knowledge base in order to achieve the best performance in the supply chain organization.

We would like to thanks all the authors for their excellent contributions and reviewers for their careful revision and suggestions for improving the proposals. We are grateful to the Springer-Verlag Team for their assistance during the preparation of the manuscript. We are also indebted to all participants and scientific committee members of the fifth edition of the DART workshop, for their continuous encouragement, support and suggestions.

May 2012

Cristian Lai
Giovanni Semeraro
Eloisa Vargiu

Reviewers

Marie-Helene Abel	University of Compiègne, France
Andrea Addis	University of Cagliari, Italy
Giuliano Armano	University of Cagliari, Italy
Agnese Augello	University of Palermo, Italy
Pierpaolo Basile	University of Bari, Italy
Roberto Basili	University of Rome Tor Vergata, Italy
Ludovico Boratto	University of Cagliari, Italy
Claudio Carpineto	Fondazione Ugo Bordoni, Italy
Annalina Caputo	University of Bari, Italy
Antonio Corradi	University of Bologna, Italy
Josè Cunha	Universitade Nova de Lisboa, Portugal
Marco de Gemmis	University of Bari, Italy
Emanuele Di Buccio	University of Padua, Italy
Alessandro Giuliani	University of Cagliari, Italy
Maria Francesca Lisi	University of Bari, Italy
Pasquale Lops	University of Bari, Italy
Massimo Melucci	University of Padua, Italy
Claude Moulin	University of Compiègne, France
Vincenzo Pallotta	Univ. of Business and Int. Studies - Geneva, Switzerland
Marcin Paprzycki	Polish Academy of Sciences, Poland
Raffaele Perego	CNR, Italy
Agostino Poggi	University of Parma, Italy
Sebastian Rodriguez	University Tecnologica Nactional, Argentina
Paolo Rosso	Polytechnic University Valencia, Spain
Fabrizio Silvestri	CNR, Italy
Alessandro Soro	CRS4, Italy
Haibin Zhu, Nipissing	University, Canada

Contents