PROMISE Winter School 2013
Bridging between
Information Retrieval and Databases

Maristella Agosti, Nicola Ferro, Gianmaria Silvello
University of Padua, Italy
{*agosti, ferro, silvello*}*@dei.unipd.it*

1 Introduction

The main mission of the PROMISE EU FP7 network of excellence is to advance the evaluation and benchmarking of multimedia and multilingual information access systems. Together with the ELIAS research network, funded by the European Science Foundation, on information access system evaluation, PROMISE has organized a winter school on "Bridging between Information Retrieval and Databases"[1] as a week long event in Bressanone, Italy, from 4th to 8th February 2013.

The aim of the school was to give participants a grounding in the core topics that constitute the multidisciplinary area of information access and retrieval to unstructured, semi-structured, and structured information. The idea of the school stemmed from the observation that, nowadays, databases are more and more getting into techniques that have traditionally been typical of information retrieval and, viceversa, information retrieval is using more and more database-oriented techniques.

17 high quality lecturers from academia and industry were invited to speak on a large variety of topics from introductory talks on databases, information retrieval, experimental evaluation, metrics and statistics to advanced topics such as semantic search, keyword search in databases, semi-structured search, and evaluation both in information retrieval and databases. Focused lectures have been devoted to bridging between information retrieval and databases and to the management and sharing of research data via evaluation infrastructures. Finally, hot topics concerned evaluation with respect to usefulness, crowdsourcing, evaluation on social media, and moving from evaluation to applications.

52 participants from 16 countries attended the courses (17% MsC students, 63% PhD students, 10% post-docs, 10% academic) with a background mostly on databases (32%), information retrieval (40%), both (15%), natural language processing (9%), and other topics. 15 scholarships (supported by ELIAS) have been granted to students to attend the school. The multidisciplinarity of the participants and lectures helped to create many lively discussions and a friendly atmosphere

[1] `http://www.promise-noe.eu/events/winter-school-2013/`

Figure 1: Different moments of the PROMISE Winter School 2013.

with many questions. Also most of the speakers stayed for the entire week and enriched the discussions as well. Interestingly enough, the school turned out to be a brainstorming and discussion opportunity also for the lecturers, since they had the occasion of meeting colleagues from a different field with their own perspectives on a ground of shared topics and issues.

2 Lectures

A total of 17 lecturers presented a lecture of 90 minutes on a specific topic. The goal was to have every day aspects of information retrieval and databases, so as to mix the topics and interests of the participants as much as possible.

Introduction to Information Retrieval – Fabio Crestani, University of Lugano, Switzerland

Prof. Crestani offered an introductory lecture to Information Retrieval (IR), reviewing some of the main concepts and providing an understanding of the architecture and functional specification on an IR system. The lecture defined what is IR, why it is hard and why it differs from DataBases (DB). Then, the whole IR process has been introduced, covering the different steps of indexing and querying.

Introduction to Databases – Maurizio Lenzerini, Sapienza University of Rome, Italy

Prof. Lenzerini introduced the notion of DB and DataBase Management System (DBMS) covering their distinguishing features, among those: the data model, the data language, a mechanism for specifying and checking integrity constraints, a transaction management, concurrency control and recovery mechanisms, and access control. He then went into the details of the relation model, relation algebra, and provided a brief overview of relational calculus.

Influence of Information Retrieval on Structured Data – Surajit Chaudhuri, Microsoft Research, USA

Dr. Chauduri started reviewing the core technologies used in the context of structured data (online transaction processing, online analytical processing, decision support systems, and data mining) in order to set the stage for the different flavors of search over structured data with particular reference to searching data in enterprises, via business objects, and consumer products over the Web. He then moved beyond searching static structured data by discussing several alternatives for middleware-based architectures.

Semantic Search – Kalina Bontcheva, University of Sheffield, UK

Dr. Bontcheva explained what is semantic search and why is it useful in order to be able to search for "thing" and not just tokens. She then provided details about the semantic annotation and semantic search processes, and the use of ontologies for semantic search. Finally, she discussed several examples of applications of semantic search such as faceted entity search and natural language queries.

The Keyword Search on Databases – Sonia Bergamaschi, University of Modena-Reggio Emilia, Italy

Prof. Bergamaschi presented the problem of extending relational databases with the capability of processing keyword queries by modeling a database as a data graph. She provided a general overview of the keyword search process and a conceptual system architecture for it. Then she reviewed the different keyword search techniques comparing them in terms of expressiveness and easiness of use. Then, she covered the alternatives for creating full-text indexes in databases and the schema-based and graph-based approaches to query them.

The Keymantic and Keyry Approaches for Querying Relational Databases – Francesco Guerra, University of Modena-Reggio Emilia, Italy

Dr. Guerra introduced the Keymantic approach for querying relational databases able to move from keywords to database queries by extending the Hungarian algorithm. Then, he presented the evaluation of the proposed approaches using two real datasets and keyword queries provided by real users. Finally, he presented an alternative approach for mapping from keywords to database queries based on hidden Markov models and its evaluation.

Semistructured Data Search – Krisztian Balog, University of Stavanger, Norway

Prof. Balog introduced the problem of searching and querying semi-structured data, such as XML (eXtensible Markup Data), in order to support users who cannot express their need in structured query languages and to deal with heterogeneity. He discussed how to exploit the structure available in the data for retrieval purposes, the different types of structure we can rely on (document, query, context), and how to use language models for semistructured data search.

Bridging Information Retrieval and Databases – Norbert Fuhr, University of Duisburg-Essen, Germany

Prof. Fuhr discussed how logic can be seen as a bridge between information retrieval and databases where databases look for objects that imply a query while information retrieval adopt a probabilistic approach looking for documents with a high probability of implying a query. The lecture introduced the probabilistic relational model and the possibility of expressing vague predicates both in database and information retrieval systems. Finally, he outlooked at the possibility of designing and developing general-purpose Information Retrieval Management Systems as it has been done in the past decades for databases.

Information Retrieval Evaluation – Ian Soboroff, National Institute of Standards and Technology (NIST), USA

Dr. Soboroff presented the topic of experimental evaluation in the IR field motivating the need for evaluation in order to assess and improve systems and reviewing its long history in the field from its inception with Cranfield experiments. Then, he discussed as this has evolved in the current paradigm based on test collections and large-scale evaluation initiatives, such as the Text REtrieval Conference (TREC) and went into the details of several issues to be considered when apply this methodology, such as robustness of the experimental collections.

Metrics, Statistics, Tests – Tetsuya Sakai, Microsoft Research Asia, China

Dr. Sakai started introducing traditional IR metrics – both set-based and rank-based – and compared them across several dimensions such as underlying user model, capability of using graded relevance, possibility of normalizing them, discriminative power and so on. Then, advanced metrics have been presented, including those targeting diversity, sessions, summarization, and question answering. Finally, correlation among rankings and statistical testing have been discussed.

Semistructured Data Search Evaluation – Ralf Schenkel, Max-Planck-Institut für Informatik (former) and University of Passau (current), Germany

Prof. Schenkel explained the problem of evaluation information search and access to semistructured data using the experience of the Initiative for XML Retrieval (INEX) as a source of concrete cases and example of tasks, metrics, and statistical analyses.

Evaluation of Semantic Technologies – Peter Mika, Yahoo! Research, Spain

Dr. Mika provided motivations pushing the need for semantic search, providing several examples of applications needing it, and introduced a framework for the evaluation of semantic search technologies. He, then, discussed the case of related entity suggestion as an example of the different evaluation alternatives that can be put in place (evaluation based on usage data, evaluation using experts, side-by-side testing, and bucket testing). He finally concluded presenting the semantic search challenge run by Yahoo!.

Evaluation with Respect to Usefulness, Some perspectives from industry – Omar Alonso, Microsoft Bing, USA

Dr. Alonso briefly revised the traditional evaluation methodologies in order to introduce the deep discussion on crowd sourcing, its benefits, and it caveats, supported by several examples and concrete cases. Finally, he discussed whether social features are useful and how we can evaluate their utility and predict the relevance of a social annotation. Finally, he concluded with a demo of an existing infrastructure for human computation, crowd sourcing, and data analysis.

Sharing Scientific and Research Data – Peter Wittenburg, Max Planck Institute for Psycholinguistics, The Netherlands

Dr. Wittenburg presented how sharing data changed over the time and the various initiatives at European level to foster the creating of common infrastructures for data sharing. He then discussed of the various work flows that need to be put in place for sharing data and the different policies needed to adopt a proper data management plan. Finally, he introduced the problem of long term preservation and curation of shared data.

Evaluation Infrastructures – Nicola Ferro, University of Padua, Italy

Dr. Ferro discussed the meaning and motivation behind the development of an evaluation infrastructure for information retrieval experimental data and then presented a high level conceptual model of the different areas that need to be covered in the information space of information retrieval evaluation. He then presented a possible service-based architecture for such an infrastructure and discussed several alternative applications it is possible to build upon it.

From Evaluation To Applications, beyond evaluating retrieval effectiveness of IR systems – Martin Braschler, Zurich University of Applied Sciences, Switzerland

Prof. Braschler presented an approach to evaluation of real applications, seen as black-boxes that can be assessed only from the outside, which represents an alternative to the traditional Cranfield paradigm when it comes to evaluating information retrieval systems deployed into real applications. He explained the different criteria and tests to be adopted to conduct such kind of evaluation and then reported the outcome on an actual study conducted on the Web portals of several profit and no-profit organization in the medical, financial, intellectual property, and cultural heritage domains.

Going social for training, tuning and evaluation – Maarten de Rijke, University of Amsterdam, The Netherlands

Prof. de Rijke discussed how to use naturally occurring side-products of user interactions or user generated content creation for training, tuning, and testing purposes. In particular, he presented how to create pseudo test collections, how to exploit click models in order to try to infer the quality of the search results based on logs of user actions, and how to use interleaving in order to try to infer the relative quality of rankers based on examining interactions with combined result lists.

3 Poster Session

To favour discussion and reciprocal knowledge, participants were asked to bring a poster describing their own research activities and plans. A committee was setup to review the posters and the three best posters have been awarded with a small prize and inviting the winners to contribute a short paper on their activities to the volume on the school lectures, currently under preparation.

The following students have been awarded:

- Ke Tao, Technical University of Delft, The Netherlands, “Twinder - Enhancing Twitter Search”.
- Mihail Minev, University of Luxembourg, Luxembourg, “Feature Extraction and Representation for Economic Surveys – Dimensionality reduction of news texts using composite features”.
- Marc Franco Salvador, Polytechnic University of Valencia, Spain, “Cross-language plagiarism detection using a multilingual semantic network”.

4 Conclusions

The fact that participants were staying together during all five days of the winter school gave them many possibilities to meet with the other participants and the lecturers. This gave place to many discussions and to a stimulating environment for both the participants and the lecturers.

Altogether the PROMISE winter school can be seen as a great success in connecting two research domains and allowing a large number of participants to get in contact with high quality lecturers. Hopefully an important outcome is that the participants have now a better view of the DB and IR research domains and also on the ways they can evaluated their own research and profit from available tools of visualization. Most participants gave a very positive feedback and hopefully the proceedings of the winter school will also help to keep the main outcomes of the winter school available for the future and persons who could unfortunately not participate.

An analysis of the evaluation forms compiled after the school highlighted that most students very much enjoyed it (97% of the participants) and the atmosphere among participants and lecturers. Most presentations were liked (95% of the participants with 77% highly appreciating the lectures) and the students were generally interested in the different topics offered by the school (95% of the participants with 76% highly interested). We noted a positive correlation (0.78) between the interest in the topics and the perceived quality of the lectures, even if some lectures were initially not regarded as on especially interesting topics have been considered very high quality.

The proceedings of the lectures of the winter school are currently under preparation and will be published in the Springer Tutorials series.

Acknowledgments

We would like to thank all the lecturers and participants to the PROMISE Winter School 2013 who gave an extremely valuable contribution and made the school a success.

The PROMISE Winter School has been supported by: the PROMISE[2] network of excellence (contract n. 258191) and by the ELIAS[3] research networking programme of the European Science Foundation.

[2] http://www.promise-noe.eu/

[3] http://www.elias-network.eu/