DESIRE 2011
Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation

Maristella Agosti
University of Padua, Italy
agosti@dei.unipd.it

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Costantino Thanos
ISTI-CNR, Pisa, Italy
costantino.thanos@isti.cnr.it

1 Introduction

This paper reports on the *Data infrastructurEs for Supporting Information Retrieval Evaluation* - DESIRE 2011 - Workshop[1] [3, 4] held on 28 October 2011 in conjunction with the 20th ACM International Conference on Information and Knowledge Management (CIKM), Glasgow, UK.

Information Retrieval has a strong and long tradition dating back to the 1960s in producing and processing scientific data resulting from the experimental evaluation of search algorithms and search systems [8]. This attitude towards evaluation has led to fast and continuous progress in the evolution of information retrieval systems and search engines.

However, in order to make the data test collections, that are used in the context of the evaluation activities, understandable and usable they must be endowed with some auxiliary information, i.e., provenance, quality, context. Therefore, there is a need for metadata models able to describe the main characteristics of evaluation data. In addition, in order to make distributed data collections accessible, sharable, and interoperable, there is a need for advanced data infrastructures.

In contrast, the information retrieval area has barely explored and exploited the possibilities for managing, storing, and effectively accessing the scientific data produced during the evaluation studies by making use of the methods typical of the database and knowledge management areas. Over the years, the information retrieval area has produced a vast set of large test collections which have become the main benchmark tools of the area and contribute to reproducible and comparable experiments [2]. However, these same collections have not been organised into coherent and integrated infrastructures which make them accessible, searchable, citable, exploitable, and re-usable to all possibly interested researchers, developers, and user communities [1].

[1] `http://www.promise-noe.eu/events/desire-2011/`

It is thus time for these three communities – information retrieval, databases, and knowledge management – to join efforts, meet, and cooperate to envisage and design useful infrastructures able to coherently manage pertinent data collections and sources of information, and so take concrete steps towards developing them. The information retrieval experts have to recognise this need, while the database and knowledge management experts have to understand the problem and work together to solve it by using the methods and techniques specific to information management.

To reach the objectives of the workshop, the following have been considered pertinent topics to address: conceptual and logical data models for representing information retrieval evaluation scientific data; metadata formats for describing scientific data produced during information retrieval evaluation; knowledge management for information retrieval experimental evaluation; data quality, provenance, adaptability and reusability in the information retrieval evaluation; data pre- and post-processing, metrics, and analyses in the information retrieval evaluation; data exchange, integration, evolution and migration for information retrieval evaluation infrastructures; workflow, Web services and Web service composition for information retrieval evaluation infrastructures; metadata formats for describing scientific data produced during information retrieval evaluation; information extraction and text mining for linking scientific literature and experimental data; data citation; evaluation, test collections, crowdsourcing for information retrieval evaluation; visualization of scientific data coming from experimental evaluation.

To address the workshop issues, experts have been invited to give keynote addresses and relevant papers have been accepted, among the submitted ones, for presentation. The two following sections are reporting the addressed specific topics.

2 Keynote Addresses

The keynote address by Norbert Fuhr of the University of Duisburg-Essen in Germany[2], entitled *An Infrastructure for Supporting the Evaluation of Interactive Information Retrieval* [7], addressed the presentation of a testbed for the evaluation of interactive information access. Starting with the INEX[3] interactive track in 2004, the group lead by professor Fuhr developed the Daffodil (now ezDL) framework, providing an experimental framework for interactive retrieval, that allows for easy exchange or extension of the system components. Moreover, this framework also contains tools for organizing laboratory experiments. Besides extensive logging (including the possibility to exploit eye tracking data), the system allows for presenting questionnaires at all stages of a search session (pre-/post- task/session), as well as the scheduling of search tasks and monitoring task time.

Due to a last minute problem, Maurizio Lenzerini of the Sapienza University of Rome, Italy[4] was unable to give his keynote address on *Ontology-based data management* [10].

[2] http://www.is.informatik.uni-duisburg.de/staff/fuhr.html.en

[3] https://inex.mmci.uni-saarland.de/

[4] http://www.dis.uniroma1.it/~lenzerin/index.html/

3 Position and Communications Papers

Among the papers that have been submitted, six have been accepted for presentation. The main topics addressed by each contribution are reported in the following.

The paper *Principles for Robust Evaluation Infrastructure* [13] by Justin Zobel, William Webber, Mark Sanderson, and Alistair Moffat makes reference to the standard “Cranfield” approach to the evaluation of information retrieval systems that has been used and refined for nearly fifty years. Over the last few years, investigation of the strengths and limitations of this approach have led to identification of serious flaws in some experiments. Since the knowledge of these flaws can prevent their perpetuation into future work and informs the design of new experiments and infrastructures, the authors review relevant aspects of evaluation and, based on their research and observations over the last decade, outline principles on which new infrastructures should rest, among those principles they emphasize that the evaluation work is only of value if the gains it describes can be verified and incorporated by others, to achieve this goal public infrastructures and shared standards are needed.

The paper *A Lightweight Framework for Reproducible Parameter Sweeping in Information Retrieval* [5] by Richard Eckart de Castilho and Iryna Gurevych introduces a lightweight framework for parameter sweep experiments geared towards evolution, efficiency and reproducibility of experiments running on a single machine. To reduce the computational effort of running an experiment with many different parameter settings, the framework uses the tasks and the dataflow dependency information to maintain and reuse intermediate results whenever possible. And for the future the authors plan to extend the framework towards support for tasks that require more processing power by allowing to run experiment work not only locally, but also on a computing cluster.

The paper *Evaluation with the VIRTUOSO platform* [6] by Gérard Dupont, Gaël de Chalendar, Khaled Kheliff, Dmitri Voitsekhovitchy, Géraud Canet, and Stéphan Brunessaux describes a software architecture for providing an open technical framework for the integration of tools for collection, processing, analysis and communication of open source information. The integration of heterogeneous components is implemented in a way that also permit the comparison of capabilities of multiple tools. The platform that supports the evaluation framework has been named VIRTUOSO. It supports an evaluation framework that allows to deploy and run evaluation kits for different use-cases.

The paper *Use Cases as a Component of Information Access Evaluation* [9] by Jussi Karlgren, Anni Järvelin, Preben Hansen, and Gunnar Eriksson argue that use cases for information access can be written to give explicit pointers towards benchmarking mechanisms and that if use cases and hypotheses about user preferences, goals, expectation and satisfaction are made explicit in the design of research systems, they can more conveniently be validated or disproved - which in turn makes the results emanating from research efforts more relevant for industrial partners, more sustainable for future research and more portable across projects and studies.

The paper *PatOlympics - An Infrastructure for Interactive Evaluation of Patent Retrieval Tools* [11] by Mihai Lupu presents the infrastructure behind the PatOlympics interactive evaluation campaign. This infrastructure, consisting of a relational database back-end, a Java processing core and a JavaScript interface, makes it possible for real users and researchers to interact in a competitive environment, while maintaining, to the extent possible, the evaluation procedures of standard information retrieval campaigns.

The paper *Infrastructure and Workflow for the Formal Evaluation of Semantic Search Technologies* [12] by Stuart N. Wrigley, Raùl García-Castro, and Càssia Trojahn describes an infrastructure for the automated evaluation of semantic technologies and, in particular, semantic search technologies. For this purpose, an evaluation framework is introduced which follows a service-oriented approach for evaluating semantic technologies and uses the Business Process Execution Language (BPEL) to define evaluation workflows that can be executed by process engines.

4 Conclusions and Future Work

The discussion among the participants has been active and productive giving insights that are going to influence the future convergence of the three communities that meet at CIKM.

In particular, the necessity for open and public benchmarks and infrastructures has been stressed since they represents the foundations of the scientific method adopted in the IR community. Indeed, algorithms and solutions tested and evaluated on private data not publicly accessible make it difficult for researchers and developers to reproduce them, verify their performances, and compare with the state-of-the-art or with own solutions.

Another important point that has been highlighted is the need for a proper and shared modeling of the experimental data produced by IR evaluation, in terms of conceptual model, descriptive metadata, and their semantic enrichment, in order to facilitate their management, access, interpretation, and re-use over the time.

Finally, it has been pointed out how much important is the community involvement to ensure that a consensus is reached in order to share approaches for open and public benchmarks as well as for modeling and describing the experimental data. Indeed, as it also emerges from the papers discussed and presented during the workshop, there are many similar and partially overlapping solutions which, in a sense, contribute to fragmentation while it would be much more beneficial to join the efforts and design and develop general-purpose and widely adopted data infrastructures for experimental evaluation.

The infrastructure that is envisaged and developed in the context of the PROMISE network of excellence[5] is going to take into consideration the feedback and discussions raised during the workshop as requirements in order to extend its support to information retrieval evaluation activities.

Acknowledgments

We would like to thank all the participants to the DESIRE workshop who gave an extremely valuable contribution and made the workshop a success.

We would like to thank those institutions and individuals who have made this workshop possible: the ACM CIKM 2011 Conference organisation, the Program Committee members, and the Department of Information Engineering of the University of Padua, Italy.

The DESIRE 2011 workshop has been supported by the PROMISE network of excellence (contract n. 258191), as part of the 7th Framework Program of the European Commission.

[5] `http://www.promise-noe.eu/`

References

[1] M. Agosti, G. M. Di Nunzio, and N. Ferro. A Data Curation Approach to Support In-depth Evaluation Studies. In F. C. Gey, N. Kando, C. Peters, and C.-Y. Lin, editors, *Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006)*, pages 65–68. `http://ucdata.berkeley.edu:7101/projects/sigir2006/papers/pdf-final/MLIA-2.pdf`, 2006.

[2] M. Agosti, G. M. Di Nunzio, N. Ferro, and C. Peters. CLEF: Ongoing Activities and Plans for the Future. In N. Kando and D. K. Evans, editors, *Proc. 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 493–504. National Institute of Informatics, Tokyo, Japan, 2007.

[3] M. Agosti, N. Ferro, and C. Thanos. DESIRE 2011: First International Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation. In I. Ounis, I. Ruthven, B. Berendt, A. P. de Vries, and F. Wenfei, editors, *Proc. 20th International Conference on Information and Knowledge Management (CIKM 2009)*, pages 2631–2632. ACM Press, New York, USA, 2009.

[4] M. Agosti, N. Ferro, and C. Thanos, editors. *Proc. Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation (DESIRE 2011)*. ACM Press, New York, USA, 2011.

[5] R. E. de Castilho and I. Gurevych. A Lightweight Framework for Reproducible Parameter Sweeping in Information Retrieval. In Agosti et al. [4], pages 7–10.

[6] G. Dupont, G. de Chalendar, K. Khelif, D. Voitsekhovitch, G. Canet, and S. Brunessaux. Evaluation With the VIRTUOSO Platform. In Agosti et al. [4], pages 13–17.

[7] N. Fuhr. An Infrastructure for Supporting the Evaluation of Interactive Information Retrieval. In Agosti et al. [4], page 1.

[8] D. Harman. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, 2011.

[9] J. Karlgren, A. Järvelin, G. Eriksson, and P. Hansen. Use Cases as a Component of Information Access Evaluation. In Agosti et al. [4], pages 19–24.

[10] M. Lenzerini. Ontology-based Data Management. In Agosti et al. [4], page 11.

[11] M. Lupu. PatOlympics - An Infrastructure for Interactive Evaluation of Patent Retrieval Tools. In Agosti et al. [4], pages 25–28.

[12] S. N. Wrigley, R. García-Castro, and C. Trojahn. Infrastructure and Workflow for the Formal Evaluation of Semantic Search Technologies. In Agosti et al. [4], pages 29–34.

[13] J. Zobel, W. Webber, M. Sanderson, and A. Moffat. Principles for Robust Evaluation Infrastructure. In Agosti et al. [4], pages 3–6.