PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

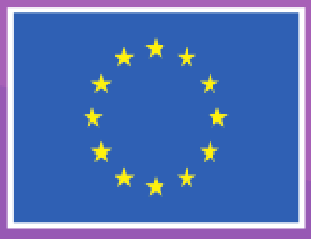
FP7 ICT 2009.4.3, Inteligent Information Management

Issues on organizing and formalizing system evaluation campaigns

SICS – HES-SO

June 14-15, 2011

Version 1.0, 15 May 2011



Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

Document Information

Report title: **Researcher Exchange date:**	Research exchange: Issues on organizing and formalizing IR evaluation campaigns June 14-15, 2011
Visitor(s):	Gunnar Eriksson, Anni Järvelin, SICS
Host(s):	Henning Müller, Theodora Tsikrika, HES-SO
Preparation date:	16/06/2011
Author(s):	Anni Järvelin, Gunnar Eriksson, Jussi Karlgren

Table of Contents

1 Introduction

Evaluation of multimedia and multilingual information access systems needs to be performed from a usage oriented perspective. Numerous factors related to usage, context and situation will influence the usefulness of a system for users and these factors should therefore also be taken into account when designing the benchmarking of such systems. More realism and context in the test collections will make the evaluation results more accessible and useful for the less initiated parties also. Even standard benchmarks make assumptions concerning the users and their information needs, even though often implicit. To make it easier to discuss and validate these assumptions, they need to be made explicit. Formulating use cases, where the system, the user and the goals of the user are defined, is a way of doing this.

SICS is responsible for the Work package 2 “Stakeholders Involvement and Technology Transfer”, where one of the goals is to formulate a use case space that can guide designing use cases for future evaluation efforts as well as make their results more useful and usable. Part of this work is identifying the features that should be included in the use case space. As a first step, some features are extracted by analyzing the PROMISE use case domains. The goal of the SICS’ research exchange to HES-SO was to acquire a better understanding of the medical image retrieval domain and the medical image retrieval evaluation campaigns that HES-SO has been organizing for several years. It is also necessary to understand the practical work of organizing an evaluation campaign, to get insight into the practical problems, and to highlight the compromises that are made between the realism of the task and the feasibility and controllability of the test setting. The researches at HES-SO have years of experience of

organizing and developing different evaluation campaigns and were therefore the perfect partners for discussing these issues.

2 Planned Work

There were several goals for this research exchange. The two-day visit was planned to include an introduction to the domain of medical image retrieval and to the practical work of organizing an evaluation campaign and a tutorial on HES-SOs previous work on use case modelling. We also wanted to discuss how the use cases should be reflected in the design of the test collections: in topics, databases, relevance judgments and in evaluation metrics.

3 Conducted Work

The visit began with an introduction to the domains of medical image retrieval and the more general purpose image search that is targeted in the Wikipedia retrieval task. We then got a thorough introduction to how the Wikipedia task has evolved through the years, how the scientific interests and the connection to authentic user needs have been balanced with the practical issues of developing, maintaining and distributing a test collection. We then continued discussing the use cases, modelling the hypothetical users and their context and how the use case should be reflected in each part of a test collection. Especially the creation of realistic topics was considered important and the different approaches previously adopted in the medical and Wikipedia tasks of ImageCLEF were discussed. Realism in databases, relevance judgments and evaluation metrics was met with slightly less enthusiasm, though many good points were made. HES-SO's previous work on modelling a use case for the medical image retrieval domain was discussed. Even if the scope of HES-SO's work here is slightly more specific than our work in PROMISE, the presentation was very interesting and useful. We also got an introduction to the ImageCLEF registration and user management system and discussed shortly its relation to the DIRECT system.

Finally, to wrap-up the visit we discussed the qualities of a good evaluation task or a CLEF lab and the issues that an organizer should consider when designing an evaluation task. Three central issues were recognized: Motivation, definition and balance. A task needs to be well-motivated and realistic to engage people and to attract participants. This also makes the results more accessible and usable. A task also needs to be well defined. It needs to be clear what is evaluated and why. The scope of the system and the assumptions made concerning the users and the usage need to be made explicit. Finally, it is important to find a balance between the novelty and continuity and between the realism and feasibility: the tasks need to evolve through the years and there should be long term goals stated for a track. At the same time, the tasks should not be too difficult, but adapted to the current level technology so that they will be challenging but not impossible for the participants. While new collections and challenging tasks should be introduced, it is important to also build on the advances from the previous campaigns and to try to engage participants in the tasks for several years. Some compromises in realism will need to be made due to practical problems with copyrights, distributing large collections and the manpower needed for relevance assessments, but there should be an awareness of these compromises and how they affect the underlying use case and the conclusion that can be drawn based on the evaluation.

All in all, we learned a lot and had a wonderful time. We are thankful for the great reception from the HES-SO crew: Thank you Henning, Theodora, Dimitrios, Manfredo, Ivan and Alex!