promise

PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Researchers Exchange Report
Report on the exchange ZHAW → SICS in the context of the PROMISE project

Technology Take-up, Use Cases, Best Practices & Evaluation in the Wild

ZHAW → SICS

August 15-19, 2011

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

Document Information

Report title:	Report on the exchange ZHAW → SICS in the context of the PROMISE project Technology Take-up, Use Cases, Best Practices & Evaluation in the Wild
Researcher Exchange date:	August 15-19, 2011
Visitor(s):	Martin Braschler, bram@zhaw.ch, Zurich University of Applied Sciences, Winterthur, Switzerland
Host(s):	Jussi Karlgren, jussi@sics.se, SICS, Stockholm, Sweden Anni Järvelin, anni@sics.se Gunnar Eriksson, guer@sics.se Preben Hansen, preben@sics.se
Preparation date:	13.09.11
Author(s):	Martin Braschler, Jussi Karlgren

Table of Contents

1 Introduction

The goals of the visit from 15.8.-19.8. in Stockholm revolved around the central aim of PROMISE to accelerate technology take-up in the field of Information Access and Retrieval. Workpackage 2 contains multiple tasks designed around this aim; chief among them the incorporation of a technology take-up group, the formulation of best practices, and the exploration of use cases. During the visit, it was explored how evaluation can be used as a tool for involvement of industrial stake-holders, by using black-box ("guerilla-style") evaluation as an attractor for other activities, and also by developing an application-centric evaluation methodology that should apply broadly to operational settings. Additional work during the visit focused on the role of uses cases in evaluation, and on linking WP2 activities to WP4 (evaluation in the wild).

2 Planned Work

The following points were fixed for the agenda of the meeting:

- collaboration SICS/ZHAW in WP2
- T2.2 use case design, paper "use cases as component of information access evaluation"
- T2.4 evaluation tasks: guerilla-style evaluation, evaluation campaigns in education
- T2.5 best practices, related workshop
- T2.6 technology take-up groups, attractors for stakeholders, synergies with T2.5
- input for D2.2
- discussion of D2.4
- plans for D2.5 technology transfer report
- WP4: evaluation in the wild
- T4.5 evaluation in the wild, associated tutorial. link to T2.4, 2.5, 2.6?
- discussion of D4.2
- new partners, new use cases

In general, all these points were discussed and worked on during the visit.

Martin Braschler travelled to and from Stockholm on 15.8 and 19.8. The work was conducted during three full days, on 16.8.-18.8.

3 Conducted Work

Main work during the visit focused on the following points:

1. application centric evaluation
2. uses cases as a tool for information access evaluation
3. technology take-up group, "attractors" for the technology transfer day
4. understanding of the interplay between best practices and use case domains

For 1), the discussion was based on a methodology partly developed in the context of two studies that had taken place in 2006 and 2007 that ZHAW had been part of [1]. The idea is to work on the level of "Information Access Applications", consisting of an IR system, a specific document collection, and the associated parametrization (configuration). This is a departure from the Cranfield paradigm of IR evaluation, where the system effectiveness is in focus, and the document collection is fixed as much as possible (test collection). The two studies mentioned were used essentially as promotional tools, and are thus lacking in fundamental scientific rigor. It was agreed that WP2 activities in PROMISE should exploit the direction given by these studies. To do so, it will be necessary to develop a new, more general and more rigorous methodology for application-centric evaluation. This work was started by developing an early idea of the evaluation criteria needed for application-centric evaluation (a hierarchy of criteria that can be evaluated individually and scored based on simple counting measures), as well as a better idea of the overall measure to be applied to the information access applications. The preliminary understanding from this work is that the resulting evaluation methodology will measure user perception of the system, based on a broad range of criteria that cover the indexing (data management), the efficiency, the user interface/experience, and the retrieval mechanism of the system. Continued work on these issues will take place in T2.2, T2.5, maybe T2.6, and also T4.3 and T4.5. Ultimately, it is foreseen that this work, tied with the work on best practices, may lead to the ability to establish a certification process for information access systems, thus directly exploiting the PROMISE results after the termination of the project. A document summarizing the early work on application evaluation conducted during the visit has been drawn up and posted to the PROMISE website. It is attached to this exchange report (Appendix A).

For 2), discussion was based on the draft of an extended paper eventually to be published authored by Karlgren et al. "Use cases as a component of information access evaluation" (based on [2]). One of the goals of SICS is to both promote more awareness among researchers and practitioners on how evaluation scenarios relate to the actual use of a system or application, and also to understand how the different use case domains impact the functioning of IR components and systems. SICS is also currently preparing a questionnaire to be used for assessing the connection between use case domains and past, present and future CLEF evaluation activities. Both the paper and the questionnaire were extensively discussed.

For 3), extensive discussion on the role of the techology stakeholders and the technology take-up group was conducted. Potentially interesting stakeholders were identified and listed. The role of these stakeholders, and their functions were discussed. SICS intends to concentrate on site visits in the next year to address the most interesting stakeholders and gain them for involvement in the project. The possibility of a workshop for industrial stakeholders was analyzed. Such a workshop brings the possibility for synergies with the best practice elaboration process. A possible date is in May 2012. The use of application-centric evaluation as an attractor tool for industrial stakeholders was also discussed (see also point 1). To this end, application-centric evaluation is conducted in a black-box ("guerilla") style, using the results to attract the attention of operators of information access applications. A similar idea has been used to considerable success in the conjunction with the earlier mentioned studies in 2006 and 2007. During the visit, an initial document summarizing the ideas on stakeholder involvement, site visits, and technology transfer day has been produced and posted to the PROMISE website. It is attached to this exchange report (Appendix B). Continued work will take place in WP2, WP4 and WP7.

For 4), initial ideas for Task 2.5 "Best Practices", due to start soon (Month 13) have been discussed. The interplay between use case domains and best practices has been analyzed. ZHAW has some earlier experience with the elaboration of best practices in the field of information access from an earlier EU FP7 project ("TrebleCLEF") [3]. However, the interlink to use case domains is entirely new and it is intended to fully exploit this link by qualifying all best practices by their applicability to use case domains where possible.

4 References

[[1] Braschler, M., Herget, J., Pfister, J., Schäuble, P., Steinbach, M., Stuker, J. (2006): Evaluation der Suchfunktion von Schweizer Unternehmens-Websites. Churer Schriften zur Informationswissenschaft.

[2] Karlgren, J., Järvelin, A., Eriksson, G., Hansen, P. (2011): Use cases as a component of information access evaluation. DESIRE 2011 workshop.

[3] Braschler, M., Gonzalo, J. (2009): Best Practices in System and User Oriented Multilingual Information Access. TrebleCLEF consortium.

5 Appendices

(see following pages)

Appendix A: Application-centric evaluation and certification of operational search services

Evaluation of IR systems versus evaluation of IR applications

The CLEF campaigns (and the similar TREC, NTCIR etc. campaigns) have been excellent drivers in developing metrics and frameworks for the evaluation of the effectiveness of IR systems. They are important events for the academic IR community. However, it has been shown to be difficult to disseminate the results outside the "core" community. Effectiveness of retrieval system most directly concern system developers. However, the number of these developers is rather small, especially when looking at enterprise retrieval systems. Some of these developers also have large research labs, which limits the interaction with the wider academic community. We argue that Promise, with its focus on knowledge transfer and uptake, should start to additionally directly address IR application implementors. For the purpose of this proposal, we define an IR application as a combination of an IR system, its document collection, and its configuration.

Technology take-up by IR application implementors

The pool of such application implementors is much greater than that of system developers, as any single IR system can be the basis for many installations. Thus, there is potential to carry evaluation to many more sites, directly addressing the promised growth in resource usage (forecasted at 300% in the description of work; figure 2). By adopting an application-centric focus, we also directly address the industrial stakeholders at companies that operate these IR applications, such as CTOs, product managers, IT managers etc. These people have to date not been reached by the output of the CLEF campaigns.

Black-box, "guerrilla" tests as a recruitment tool

Application-centric tests can be conducted as black-box or minimally invasive tests, which allows direct testing of operational search services (living retrieval laboratories T4.4, evaluation in the wild T4.5). We propose to additionally use application-centric testing as an "attractor" tool to recruit application implementors to the Promise Technology Day event (D7.11). To this end, we propose to complement the development of a methodology for application-centric evaluation (alternative retrieval scenarios and evaluation metrics T4.3) with a "guerrilla-style" evaluation of a number of public enterprise search applications. The companies responsible for the search applications picked for these evaluations are then directly addressed and invited to join the technology take-up group (T2.6) and attend the Technology Day event, where they are – among other things – presented with the results of this evaluation ("honey pot" to attract participants to the event).

Certification of IR applications

Application-centric tests are the basis for a certification of applications. Enterprise search applications are usually tied to one or many knowledge intensive business processes. The use cases proposed in Promise map closely to such knowledge intensive business processes. By selecting the right set of parameters pertaining to each use case, and testing the application through tests associated with these parameters, proper operation of the application can be validated and certified. The different tests are informed through the distillation of best practices (as part of Task 2.5).

Operation of IR application testing

Ultimately, the tests tied to the different parameters intend to measure the user's overall perception of the system, thus addressing the system in a more integral way. It is envisioned to test each parameter with a number of (small) tests, which can be carried out automatically, semi-automatically or by trained testers. The tests should be geared towards reproducibility and objectiveness, i.e. the testers will work according to a script. This means that in the simplest form, no special skills are needed for testers, ensuring low cost of the approach. Applications can be evaluated in isolation, without the need for campaigns, by normalizing the test scores and using appropriate thresholding for certification purposes.

Possible parameters for IR application evaluation:

- multilinguality (coverage of language specifics, morphology, diacritics, ..)
- cross-language functionality (translation etc.)
- coverage (index coverage, freshness, timeliness, ...)
- search quality (collection quality, retrieval effectiveness, tied to the core entities of the underlying business processes)
- performance
- usability (user interface)
- features (interaction modes)
- cost (of operation; not possible in "black box" evaluation?)
- persistence/personalization
- collaboration (user engagment, social media, hookup)
- geoposition (mobile applications,..)
- auditabiliy
- predictability

Exploitation

The certification process outlined has potential to be used in the exploitation phase, being spearheaded by a Promise successor legal entity, as is mentioned in the proposal and description of work.

Appendix B: Achieving impact and raising awareness for stakeholders

An integral part of the PROMISE work involves raising the awareness at stakeholder sites of our activities and infusing development of applicaitons and systems in our general technical area with recent research results, contributing to best practice among developers, designers, and information providers in the field.

To achieve these goals, among the tasks PROMISE has set out for itself in Work package 2 ("Stakeholders Involvement and Technology Transfer"), Work package 4 ("Evaluation Metrics and Methodology"), and Work package 7 ("Dissemination, IPR and language resources") we have tasks for formulating use cases (T2.2), validating use cases with respect to stakeholders (T2.3), formulating evaluation tasks (T2.4), formulating best practices for developers in the field (T2.5), forming a technology take-up group to facilitate the incorporation of stakeholders in academic activites (T3.6), developing new usage scenarios and attendant evaluation metrics (T4.3), for evaluation outside the laboratory benches (T4.5), and a task for organising a Technology Transfer Day towards the end of the project lifetime (T7.4). Deliverables related to this task are D2.3 ("Best practices report" in M24), D7.11 ("Technology Transfer Day" in M33), and D2.5 ("Technology transfer report" in M36). An associated event planned for D2.3 is a workshop on Best Practices in M21 in May 2012.

Proposal: Site visits

As an activity cutting across these tasks, we have planned a series of site visits to gather information about usage, refine our picture of how use cases can be formulated to conform with practical usage among stakeholders, promote best practices and sensible evaluation methodologies.

Proposal: Interview day

In conjunction with the Best Practices workshop planned for M21 we propose to interview participants about their use case models.

Coverage

To gather enough information to span the application space we want to have breadth of coverage. Firstly we want to cover the three use case domains of PROMISE:

- Unlocking Culture: museums, libraries, archives and other memory institutions
- Search for Innovation: patent offices, patent lawyers, innovation brokers, IP officers at large corporations
- Visual Clinical Decision Support: medical clinicians, health information sites, pharma companies, medical image analysis companies, public information sites

and we will want to find stakeholders in each of these domains. We will also explore those current CLEF labs (Music retrieval, Plagiarism and Authorship analysis, Question Answering) not covered by the PROMISE use case domains to see if we can find stakeholders through them. But we also want to go further afield. Current suggestions are given in the (obviously incomplete) list below. We need a systematic choice of partners through the list.

Methodology

The procedure at a site visit should be somewhat systematic. The questionnaire being developed at present for the purpose of discussing use cases within labs will provide us with a backbone, and to ensure continuity, an overlapping group of people should be responsible for conducting the visits, with each visit being attended by a mix of previous and new participants.

Timeline

The timeline for these visits is generous, but coming up soon, beginning immediately after the Amsterdam CLEF, and continuing through the Best Practices workshop in M21 (May 2012) where some of the sites can be invited to participate and discuss their use cases - obviously closely related to best practice formulation.

Stakeholder list

- Unlocking Culture: museums, libraries, archives and other memory institutions (Planned by SICS with UBER)
- Search for Innovation: patent offices, patent lawyers, innovation brokers, IP officers at large corporations (Planned by SICS with TU-Wien)
- Visual Clinical Decision Support: medical clinicians, health information sites, pharma companies, medical image analysis companies, public information sites (SöS? Linköping?) Picsearch? (Planned by SICS with HES-SO)
- e-government: municipal, government and EU agencies, with legal requirements for information provision and content of high editorial quality, great pertinence, and high redundancy
- hardware manufacturers: tools with the capability to serve material in ways which content and network providers might not have envisioned (or the converse)
- media and publishing houses, broadcasters: mostly local in scope but with (mostly) a willingness to find a global audience, monitoring for plagiarism and abuse, vectored towards paying customers and associative advertising
- network providers
- educational institutions
- social media sites
- activist sites

- consumer goods and service purveyours
- legal practicioners