CHEVAL: Chur Evaluation Laboratory

Thomas Weinhold, Lydia Bauer, Josef Herget, Sonja Hierl, Joachim Pfister
Swiss Institute for Information Research (SII), Hochschule für Technik und Wirtschaft (HTW), Chur, Switzerland
thomas.weinhold@fh-htwchur.ch
lydia.bauer@fh-htwchur.ch
josef.herget@fh-htwchur.ch
sonja.hierl@fh-htwchur.ch
joachim.pfister@fh-htwchur.ch

Abstract: Incorporating novel approaches like visual components, semantic web ideas or Web 2.0 concepts into information retrieval systems poses new challenges for their systematic evaluation. Currently, the development of valid evaluation settings cannot keep up with the development of new search engines and innovative information retrieval concepts. Therefore, the Swiss Institute for Information Research (SII), is currently developing a testbed called CHEVAL (Chur Evaluation Laboratory) to tackle this problem. The vision of CHEVAL is to design an integrated, multi-level and multi-methodological web-based system and framework to support different kinds of evaluation types (e.g. usability tests, IR efficiency measurement, benchmarking studies etc.) of several types of information retrieval systems. In the context of CHEVAL, an evaluation can have multiple dimensions regarding the type of the evaluation (long-term or short-term test phase, comparative or non-comparative evaluation, field or laboratory test environment) and the methods used for the evaluation, which can either be from IR efficiency measurement or usability testing as well as a combination of both. The paper will give an overview of some well-known and widely accepted evaluation initiatives. This includes as well background information about the history of these initiatives. Furthermore the strengths and weaknesses of the described evaluation initiatives will be presented and discussed. Based on the deficiencies of current approaches for evaluating information retrieval systems with visual or semantic components the vision and the goals of the Chur Evaluation Laboratory will be explained. Following, the architecture of the testbed will be introduced. An example will illustrate how the system is intended to be used and what advantages CHEVAL will give to evaluators of information retrieval systems. Finally, the paper will present the success factors and a short roadmap for the further development of the Chur Evaluation Laboratory.

Keywords: evaluation, information retrieval systems, system-supported evaluation, usability testing

1 Evaluation: new search concepts require new methods and frameworks for evaluation

Nowadays, internet users are confronted with rapidly growing information resources. This applies as well for organizational environments (intranets). Therefore, it is not surprising that search engines are regularly among the most frequently used websites and software applications. During the last years, a growing number of search engines and other information retrieval systems (IRS) that incorporate novel approaches like visual components, semantic web ideas or Web 2.0 concepts could be noticed. One example for such a new approach is the convincingly visualization of search terms or search results using elaborate graphical components, like applied in Quintura (www.quintura.com/), Grokker (http://live.grokker.com.) and various other systems. The question is whether these new approaches really are suited to improve the access to information for the users – this question can only be answered by solid and comprehend evaluations.

While the evaluation of common IRS and search engines is based on standard procedures and measurements, no such standards have yet been established for IRS incorporating novel approaches. Because of the fact that the causal relation between semantic and visual components and the search results of the user interaction with the information system is quite complex, such systems need to be examined through various approaches and methods (Bauer et al. 2007). Currently most authors either use standard information retrieval evaluation measurements and settings or apply usability measurements for evaluating their visual IRS (Koshman 2005; Reiterer et al. 2005; Zwol & Oostendorp 2004; Reiterer 2004; Sebrechts et al. 1999; Veeresamy & Belkin 1996). However only in a few cases do the authors motivate their choice and apply a combination of methods from those two areas. Therefore, Vaughan (2004) states that the design of valid evaluation settings and techniques is not keeping up with the rapid development of visual IRS.

One of the main problems that can be identified in the evaluation of IRS using novel concepts is the impact of customer habits, which have been established over several years when using standard IRS like Google, Yahoo or MSN (Hierl 2007). A number of evaluations comparing search engines with visual components to traditional systems come to the conclusion that their results have to be put into perspective, since the long-time experience was not considered in the evaluation setting (Arnold & Wolff 2005). Thus, an evaluation framework for systems integrating novel approaches urgently needs to take into account user habits, e.g. by applying long-term studies or field studies (Shneiderman & Plaisant 2006).

2 Established Evaluation Initiatives

The evaluation of information retrieval systems already has a long tradition, starting in the 1960s with the Cranfield tests. Cranfield I and II created a major interest in the automatic indexing and searching of texts. Also the tests emphasized the importance of creating test collections, which are applicable for comparative evaluation (Harman 1993).

In 1992, a large test collection was made available to the research community by the formation of the Text Retrieval Conference (TREC). TREC came up with the idea that providing an open testing event, with common tasks and a standard evaluation scenario, would lead to the acceleration of research on a realistic scale (Voorhees & Harman 2005).

Based on the idea of TREC, a couple of other evaluation initiatives were established over the last years. The Cross Language Evaluation Forum (CLEF), which was established as an independent project in the year 2000, focuses on cross-language information retrieval. The declared objective of CLEF is to create an R&D community in the cross-language information retrieval sector (CLIR) and to provide an infrastructure for the testing and evaluation of IRS operating on European languages in both monolingual and cross-language contexts (Peters 2002).

Another important project is the INitiative for the Evaluation of XML Retrieval (INEX) which was launched in 2002. INEX faces the challenge of how to measure an XML information retrieval systems' effectiveness (Kazai & Lalmas 2005).

2.1 Strengths

The mentioned initiatives and projects have contributed to the standardization of the evaluation methodology of IRS and are widely accepted through the field of information science. Because of the fact that these initiatives offer the possibility to perform the same tasks on the same data several times on different systems, they can be used for the purpose of benchmarking as well as for repetitive studies, which analyse the improvement of the tested system after further development on the system has been conducted.

Additionally, the initiatives provide a forum where a large community of researchers can openly discuss their retrieval techniques. The conferences offer researchers the possibility to efficiently learn from one another and thus facilitate technology transfer. Finally, TREC and the other evaluation initiatives have served as an incubator for new technologies for innovative retrieval applications (Vorhees & Harman 2005).

2.2 Weaknesses

Although the mentioned evaluation initiatives have contributed a lot to the improvement of IRS, they are aligned with several problems. First of all, the evaluation methodologies suggested by TREC and other evaluation initiatives primarily focus on the measurement of recall and precision and not on the measurement of the target achievement of a real user. Thus, only text-based IRS (or with regard to INEX XML based IRS) can be evaluated within these initiatives.

Another major problem is the fact that the user interface and the user-system interaction cannot be analysed within the suggested frameworks. In their approach, controlled set of documents and tasks are used, which make them comparable to laboratory studies. Therefore, they are suited for comparing IRS in a controlled environment, but they are not applicable for analysing real user interaction with an IRS based on authentic information needs, for example in the setting of a field study, like postulated by Shneiderman & Plaisant (2006).

2.3 Consequences

As mentioned above, the presented evaluation initiatives focus mainly on the measurement of recall and precision. Thus, they are not suitable for evaluating IRS with visual or semantic components, because these measurements can be calculated for systems without a textual display of the retrieved documents. The efficiency of such IRS can only be analysed with regard to the user and their interactions. Therefore, new evaluation frameworks are needed which allow to analyse IRS in real scenarios within real users and which allow a combination of methods both from IR efficiency measurement and usability testing. Currently, there is a lot of ongoing research in this field but no accepted standards to solve these problems have been developed yet.

3 Description of CHEVAL

3.1 Vision and goals

On basis of the above described weaknesses of current evaluation initiatives and the lack of suitable evaluation concepts along with the absence of an integrated, system-supported approach led to the conceptualisation and development of CHEVAL.

The vision of CHEVAL is to design an integrated, multi-level and multi-methodological web-based system and framework to support different kinds of evaluation methodologies (e.g. usability tests, IR efficiency measurement, benchmarking studies etc.) for multiple types of information retrieval systems with integrated visual, multi media, Semantic Web or Web 2.0 components. One evaluation concept implemented in CHEVAL can have multiple dimensions regarding the type of the evaluation (long-term or short-term test phase, comparative or non-comparative evaluation, field or laboratory test environment) and the methods used for the evaluation.

CHEVAL aims to provide a general infrastructure for performing evaluation tests, selecting the suitable methods, administrating test collections, supervising the whole evaluation process, storing results and generating reports. Because of the fact that an evaluation of an IRS consists of various steps which often need to be iterated for several evaluation test runs, the repetitive character of such evaluations is one of the main inspirations for CHEVAL.

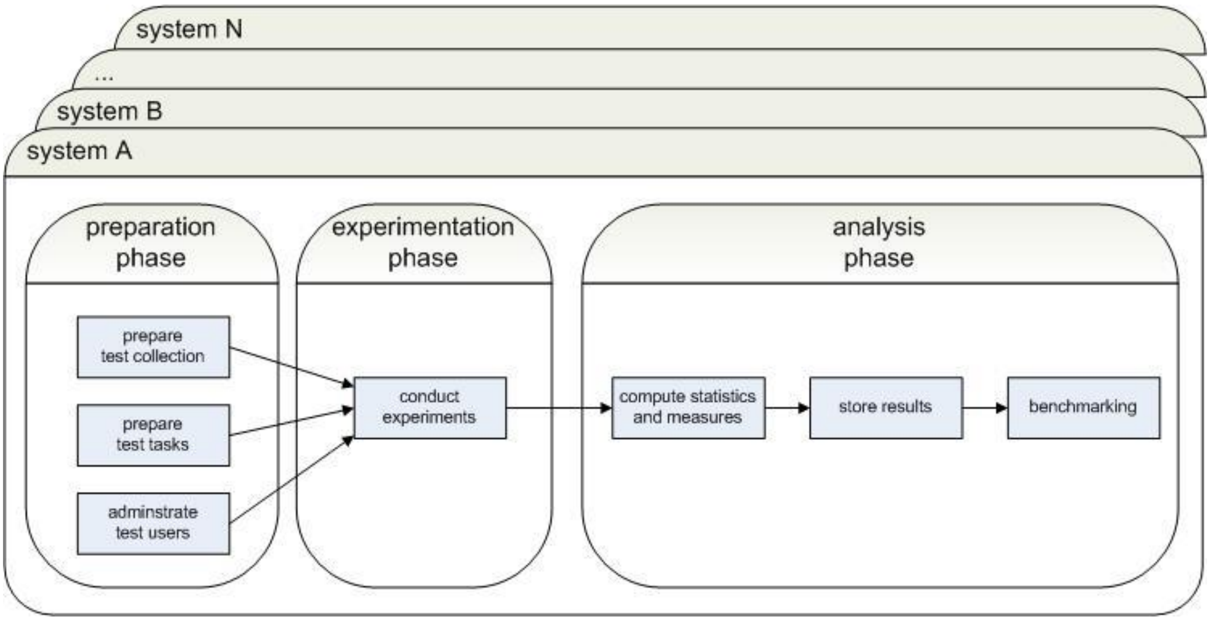

Figure 1: Evaluation phases

The framework will support the entire evaluation process in all its different steps, from planning to the actual experiment phase and the final analysis phase (figure 1). By the help of a decision tree the evaluator will be guided through all steps to set up the evaluation with CHEVAL and system support is provided throughout the whole process. The tree will help to make decisions in questioning whether a long- or short-term evaluation should be conducted, if a comparative evaluation should be set up and what kind of IRS are going to be tested. According to these answers given, the testbed offers a set of suitable methods to choose from. Finally, the IRS to evaluate will be connected to CHEVAL, so that the experiment phase with test users can be conducted. All results are stored and saved by CHEVAL for further processing. The reporting module will offer different means of data analysis with various kinds of methods to extract all results from the tests. If some results require intellectual or semi-automatic reporting and analysis, the CHEVAL expert team will offer their expertise to interpret the collected data material.

Because of the fact, that all results will be stored, over time a broad basis for a benchmarking of all kinds of IRS will be available. Due to the questions from the decision tree, all different kinds of IRS can be identified and can be compared to each other's results.

In case further explanation on any topic is needed, the CHEVAL knowledge base, which is build up successively in form of a wiki, will provide further detail information. Furthermore, a CHEVAL evaluation community will be established that allows the discussion of different cases in detail. The members of the community will consist of all interested parties, test participants and system users as well as team members of the CHEVAL expert team.

3.2 Architecture

The idea of CHEVAL is to build a software framework supporting activities throughout the whole evaluation process. The testbed will be implemented as a web application with a server component the clients can access via the internet or the intranet of an organization. Its architecture is structured in several modules (figure 2).

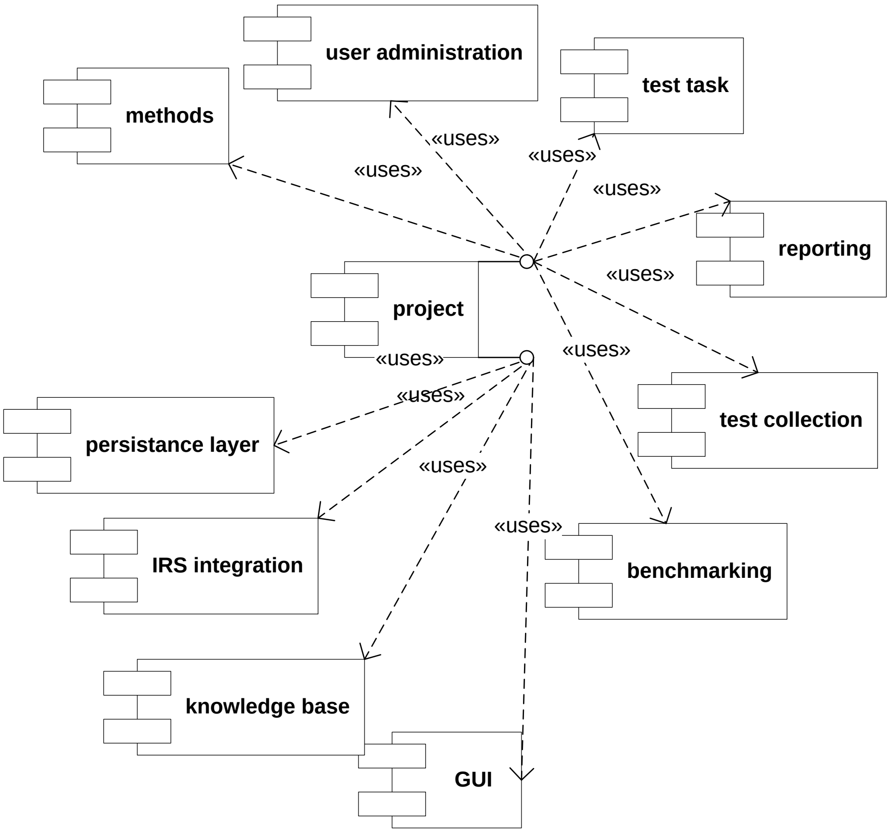

Figure 2: Architecture of CHEVAL

The component "user administration" contains all functionalities related to the authentication and authorization of users working with CHEVAL. These functions are needed, since the testbed environment implements a role concept model. The following roles are implemented:

- A system administrator takes care of the whole CHEVAL setup. He authorizes other users or modifies test collections.
- The test administrator is responsible for setting up a test project. He is the one deciding which test methods and test collections will be used. He also specifies which test users are requested to take part in the experimentation phase.
- Finally, the test user is the person, who actually carries out a test during the experimentation phase.

This top down implementation is necessary for ensuring that comprehensive evaluations with a large number of test persons, test cases, different test collections or multiple evaluation focuses can be conducted in a structured and systematic way. For less complex evaluations, CHEVAL will also provide support for the ad-hoc use of several instruments, such as questionnaires or check lists.

The details of a test are defined in the component “test task” the test administrator will work with. Parameters determine which test collection and which test methods are defined there. The different test methods, their description and initial configuration can be modified by the system administrator by using the “method” component. All information regarding a test project is integrated in the component “project”.

When a test run has been performed, the component “reporting” provides functionality for generating statistics and reports, whereas the component “benchmarking” is used to compare test runs with each other. Data produced within the experimentation phase and during the set up of an evaluation project is stored using the persistence layer component. The component “knowledge base” encompasses all the methods needed to store, access, create and modify contents of the knowledge base. For instance, you will find general information there on how to carry out an evaluation and the (dis-) advantages of a test method. This can help test administrators to decide which method to choose. The interaction with the users of CHEVAL is managed by a graphical user interface in the “GUI” component.

3.3 An illustrative example

As described above, CHEVAL will be applicable for different kinds of evaluations. One possible domain is the evaluation of the search functionalities of corporate websites. The quality of these search functions is of great relevance for companies. According to a study conducted by Forrester Research, about 50 % of the users apply the search functions to find the information they are looking for on unfamiliar websites and about 13 % of the users are likely to visit another site if they can not find what they are looking for immediately (Cremers 2006). Therefore, the SII, in cooperation with two Swiss companies (namics and Eurospider) and another university of applied sciences (Zürcher Hochschule Winterthur), has developed a methodology for assessing the quality of the search functionality of a corporate website (Braschler et al. 2006). The methodology for the assessment consists of 74 single tests which are structured as shown below:

1. Search Index
    - 1.1 Completeness
    - 1.2 Actuality
    - 1.3 Query- and document representation
2. Query / Document matching
    - 2.1 Query execution
    - 2.2 Performance of the query language
    - 2.3 Quality of the meta data
3. User Interaction
    - 3.1 Presentation of the results list
    - 3.2 User guidance
    - 3.3 Performance
4. Search Results
    - 4.1 Navigational queries
    - 4.2 Informational queries
    - 4.3 Factual queries

Examples for tests in the category “query and document matching” are, if it is possible to search for phrases which include stop words or if very long queries can be executed by the search engine. One major advantage of this assessment scheme is the fact, that not only aspects concerning the completeness and accuracy of the search results are considered, but also aspects which apply the user interaction with the systems.

In autumn 2006 the assessment scheme was used to evaluate the website search functionalities of 54 Swiss companies in cooperation with all the partners mentioned above. Since CHEVAL is in the conceptual development phase, no concrete implementation was available for the assessment.

Therefore the test setting was set up as follows: The testers had to follow a manual which described how to carry out each test, step by step. This manual was available in print as well as electronically in form of a word file. Most of the tests required the testers to note down the query terms they used for a specific step and to document the results for each step. For this purpose, Excel spreadsheets were used. In this setting, the users had to work with three different sources to complete a test: the manual to obtain the instructions, the spreadsheet to document the results and finally a browser to actually perform the test. Sometimes the test tasks contained initial checks if an item to be testes, really exists. If not, the testers had to proceed to another question thus leading to a non-linear way of task execution. To aggregate the individual test results and build a report, excel was used.

With the help of CHEVAL, the test setting would look like the following: A test user logs in into the CHEVAL system using a standard web browser. Instead of having separate places for looking up the manual and documenting the results, the instructions for a task as well as form elements to enter e.g. query terms or test results is presented to the user in one screen. This increases the usability since it is avoided that a user has to switch between different media or applications. Checks for completeness of the data entered are also implemented to guarantee a higher data quality. If a user decides to quit the test execution, he simply can log out and after logging in again, he is transferred to the question he worked on at last. Test administrators can also profit from a test execution using CHEVAL. If they assigned a test to several testers, they can use the statistical functions to obtain progress information which is useful for project controlling.

Due to the fact that the results are stored according to a predefined scheme, the system will offer the possibility to summarise the results automatically and to generate a report. Based on these results the organisation performing the evaluation can either carry out a benchmark analysis or derive recommendations for improving the quality of the search function. On the one hand, the data for the benchmarking can be created by the organisations themselves by performing several evaluations of different systems. On the other hand, the CHEVAL benchmarking database can be accessed where organisations can voluntarily store their evaluation results.

The main advantages of cheval are twofold: by using CHEVAL, no time has to be spent on developing an evaluation framework by the organisation itself which considerably shortens the evaluation period and thus, lowers the costs of the evaluation. Second, the benchmarking possibilities offer organisations the opportunity to compare the performance of the tested systems and to derive recommendations for improvement.

4 CHEVAL – a viable approach?

4.1 Success factors

CHEVAL is designed to offer support for different kinds of evaluations such as usability tests, information retrieval efficiency measurements or customer satisfaction analysis. One important factor for the success of CHEVAL is the question whether there are enough requests for the provided services. A major plus for CHEVAL is the fact that currently no systems or frameworks are available, which provide an inclusive approach with all the services planed for the Chur Evaluation Laboratory. In a comprehensive market study we have found, that no provider exists offering usability evaluation of IRS in a combined framework with retrieval efficiancy testing. Furthermore, the study described in chapter 3.3 indicates, that there is a market for such a service due to the fact that today the website search functions of many companies show significant deficiencies (Braschler et al. 2006). Another advantage of CHEVAL will be the possibility to perform benchmarking studies of different sorts of IRS, either in relation to a certain branch or in relation to all existing IRS of the same type. Thus, CHEVAL meets the demand to act as the first full-service framework concerning evaluations in the field of information science.

To ensure the success of CHEVAL, it is also planned to provide the community with the software as an open source product. This will help to increase the awareness level of the framework and offers the possibility that interested researchers can contribute to the further development of CHEVAL.

4.2 Road map

The implementation of CHEVAL is planned to be conducted in several iterations. Thus, in the beginning the framework will not support all of the planned services. Rather the services will be implemented consecutively and gradual in connection to their complexity. It is planned to start with services which only require methods and resources that can be easily adapted such as questionnaires or interview guidelines. Step by step CHEVAL will be enhanced with more complex services. The following figure illustrates this approach:

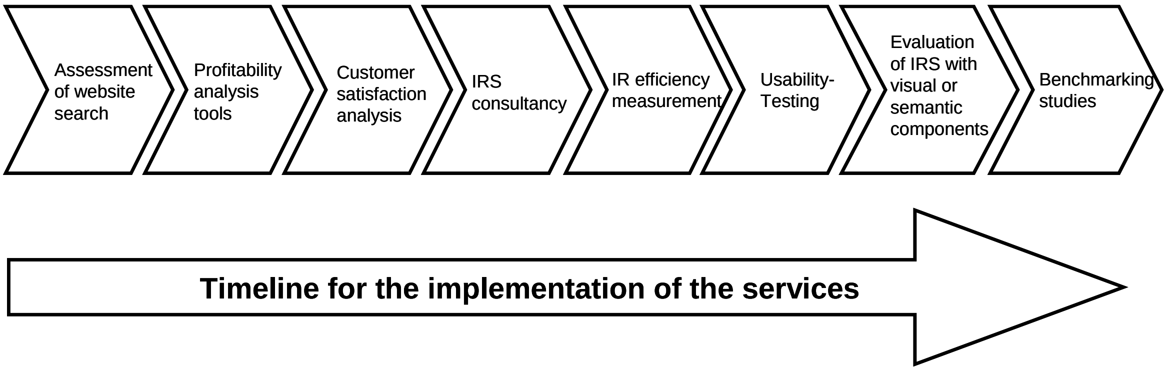
Assessment of website search
Profitability analysis tools
Customer satisfaction analysis
IRS consultancy
IR efficiency measurement
Usability-Testing
Evaluation of IRS with visual or semantic components
Benchmarking studies
Timeline for the implementation of the services

Figure 4: Roadmap for the implementation of CHEVAL

Currently, the project is in the analysis phase. The implementation of the first services is planned for the first quarter of 2008.

4.3 Conclusion

By offering a wide range of services, CHEVAL will give support to all the people, who have to deal with evaluations of information systems or services. The system will help to conduct evaluations faster and more efficiently, which allows reducing the costs for the evaluations and at the same time increases the quality of the evaluations.

CHEVAL aims especially to provide a framework to evaluate IRS which incorporates semantic or visual components in a systematic, standardized and methodologically sound way. Since many research projects are dealing with the development of new search engines and search concepts, CHEVAL will be a tremendous help for fellow researchers who will altogether have to evaluate their newly developed systems and who up to now seemingly have had big trouble in doing so due to nonexistent support tools (Arnold & Wolff 2005, Hierl 2007). Due to the modular structure of the testbed and its gradual composition, new developments in the area of IRS like social search on the basis of Web 2.0 concepts and other upcoming future trends will be able to be taken into account for the further development of CHEVAL. Furthermore the Chur Evaluation Laboratory will establish a forum for researchers and developers to discuss new evaluation methods and techniques for IRS with semantic or visual components, which are not in the focus of current evaluation initiatives. It is our belief that CHEVAL is an approach that will provide a basis for widely serving the IRS community in evaluation issues in the future.

References

Arnold, C. and Wolff, C. (2005) "Evaluierung von Visualisierungsformaten bei der webbasierten Suche.", *Knowledge eXtended. Die Zusammenarbeit von Wissenschaftlern, Bibliothekaren und IT-Spezialisten*, Schriften des Forschungszentrums Jülich, Reihe Bibliothek, Band 14, 2005, 275-286.
Bauer, L., Herget, J. and Hierl, S. (2007) "Conceptual challenges for the evaluation of digital repositories with multiple access options", *Conference on Semantic Web & Digital Libraries (ICSD)*, Bangalore (India), Feb. 2007.
Braschler, M., Herget, J., Pfister, J., Schäuble, P., Steinbach, M. and Stuker, J. (2006) "Evaluation der Suchfunktion von Schweizer Unternehmens-Websites", *Churer Schriften zur Informationswissenschaft,* Schrift 12, Chur, Dezember 2006.
Cremers, Iris et al. (2006). It's Time To Update Site Search Functionality. Best Practices.
Forrester Research, Inc.
Cross-Language Evaluation Forum (CLEF) [Online], http://www.clef-campaign.org/.
Google [Online], http://www.google.de.
Grokker [Online], http://live.grokker.com.
Harman, D. (1993) "Overview of TREC-1", Proceedings *of the workshop on Human Language Technology*, Princeton, New Jersey, 61-65.
Hierl, S. (2007) "Bezugsrahmen für die Evaluation von Information Retrieval Systemen mit Visualisierungskomponenten", *B.I.T. Online*, 2/2007.

INitiative for the Evaluation of XML Retrieval (INEX) [Online], http://inex.is.informatik.uni-duisburg.de/.

Kazai, G. and Lalmas, M. (2005) “Notes on what to measure in INEX”, *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, Glasgow, July 2005.

Koshman, S. (2005) “Testing user interaction with a prototype visualization-based information retrieval system”, *Journal of the American Society for Information Science and Technology*, 56(8) 2005, 824-833.

MSN [Online], http://de.msn.com/.

Peters, C. (2002) “The contribution of evaluation: the CLEF experience”, *Proceedings of SIGIR 2002: a Research Roadmap*. Workshop held at the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, August 15, 2002, 68-71.

Pirolli, P. and Card, S. (1999) “Information Foraging”, In *Psychological Review* 106(4): 643-675.

Plaisant, C. (2004) “The Challenge of Information Visualization Evaluation”, *Proceedings of Conference on Advanced Visual Interfaces AVI'04*.

Quintura [Online], http://www.quintura.com/.

Reiterer, H. (2004) “Visuelle Recherchesysteme zur Unterstützung der Wissensverarbeitung“, Hammwöhner, R., Rittberger, M., Semar, W. (Eds.): *Wissen in Aktion. Der Primat der Pragmatik als Motto der Konstanzer Informationswissenschaft. Festschrift für Rainer Kuhlen.* Constance, 1–21.

Reiterer, H., Tullius, G., and Mann, T. M. (2005) „INSYDER: a content-based visual-information-seeking system for the Web”, *International Journal on Digital Libraries*, Volume 5, Issue 1, Mar. 2005, 25 - 41.

Sebrechts, M., Vasilakis, J., Miller, M., Cugini, J. and Laskowski, S. (1999) “Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces” In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3-10.

Shneiderman, B. and Plaisant, C. (2006) “Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies”, *Proceedings of the BELIV'06 Workshop*, Venice 61–77.

Text REtrieval Conference (TREC) [Online], http://trec.nist.gov/.

Ujiko [Online], http://www.ujiko.com/.

Vaughan, L. (2004) “New measurements for search engine evaluation proposed and tested”, *Information Processing and Management* 40 (2004) 677–691.

Veerasamy, A. and Belkin, N. J. (1996) “Evaluation of a tool for visualization of information retrieval results”, *Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '96*, ACM Press, New York, NY, 85-92.

Voorhees, E. and Harman, D. (2005) “The Text Retrieval Conference”, Voorhees, E and Harman, D. (Eds.): *TREC - Experiment and Evaluation in Information Retrieval*, The MIT Press, Cambridge.

Yahoo [Online], http://de.search.yahoo.com.

Zwol, R. Van and Oostendorp, H. Van (2004) “Google's ‘I'm feeling lucky’, Truly a Gamble?”, Zhou, X. et al. (Eds.): *Web Information Systems - WISE 2004, Proceedings of the 5th International Conference on Web Information Systems Engineering*, Brisbane, Australia, 378-390.