Evaluating Interactive Information Retrieval Systems

Pertti Vakkari

School of Information Sciences
33014 University of Tampere
Finland
pertti.vakkari@uta.fi

Resumo

Neste artigo é proposta uma extensa metodologia para avaliar a recuperação da informação interativa. A proposta baseia-se em princípios fundamentais apresentados na literatura de avaliação para definir os objetivos do sistema ou ferramenta a ser avaliado, e inferir as medidas e os critérios de sucesso na consecução dos objetivos. Propõe-se que, ao avaliar uma ferramenta de pesquisa, seja analisado em que medida ela beneficia os utilizadores, aumentando a sua capacidade de pesquisa e, consequentemente, contribuindo para a qualidade da lista de resultados. Além da qualidade da lista de resultados, é importante avaliar até que ponto o processo de busca e as ferramentas que o suportam atingem os seus objetivos.

Palavras-chave:

Recuperação da informação, avaliação, busca, metodologia.

Abstract

An extended methodology for evaluating interactive information retrieval is proposed. It is based on principles presented in core evaluation literature to define the goals of the system or tool to be assessed, and infer from that measures and criteria of success in attaining the goals. It is proposed that in assessing a search tool it is analyzed to what extent it benefits users by increasing their ability to search, and consequently, contributes to the quality of the result list. In addition to the quality of the result list, it is important to assess to what extent the search process and tools supporting it met their goals.

Keywords:

Information retrieval, evaluation, searching, methodology.

1. Introduction

The methodological rule given in literature is to begin an evaluation by analyzing what is the objective of the system, process or service to be evaluated. It is assessed to what extent the object of evaluation attains the goals defined. Therefore, it is necessary to identify the goals of the system, and measures of goal attainment and criteria for assessing goal attainment. Goals are typically defined in terms of what the system aims at achieving (Rossi & al. 2004; Rubin 2006).

The goal of the system-oriented IR research is to develop retrieval models, techniques and algorithms to identify and rank topically relevant documents, given a topical query. A retrieval model consists of the specification of the document representation and query representation and the definition of the matching method. This laboratory model of IR evaluation aims at measuring the effectiveness of retrieval models and algorithms, expressing their ability to identify topically relevant documents. The evaluation framework consists of a test environment containing a document collection, a set of test requests i.e. topics representing information needs and a set of relevance assessments indicating the documents that are relevant to each search topic, i.e. which should be retrieved as answers to requests (Tamine-Lechani &al. 2010).

The methodology used in information retrieval experiments for evaluating the functioning of particular system features like indexing methods or algorithms corresponds to the rule mentioned above. The goal of the system is typically defined as to retrieve all and only those documents pertaining to a topic. This goal originate from the time of compiling subject bibliographies aiming at covering all and only those documents, which belong to a subject field (Belkin 2010). For measuring the goal attainment in retrieval experiments the output indicators recall and precision were derived. Recall is the proportion of topically relevant (pertinent) items retrieved of all relevant items (a/a+c). Precision is the proportion of relevant items retrieved of all items retrieved (a/a+b) (figure 1).

Figure 1 – ***Results of a search***

System Relevance Prediction	User Relevance Decisions	
	Relevant	Not Relevant
Relevant	Hits (a)	Noise (b)
Not relevant	Misses (c)	Correctly rejected (d)

Relevance assessments are typically made by the external judges who have also designed the search topics. These judges assess, which documents are topically relevant to the given request. They are expert opinions, what is considered as topical document.

The following step in evaluation is to define criteria for success in attaining the goal given. If the goal of the system is to retrieve all and only those documents pertaining to a topic, should we set as the criterium of success a 100 % precision and recall? However, it is not typical to define absolute criteria for success in retrieval evaluation experiments. A typical standard of success in retrieval experiments is the performance of a system similar to the object system not including the feature under evaluation. If the object of evaluation produces a significantly higher recall or precision compared to the baseline system, then it is considered as being successful. Thus, the goal of the system evaluated is to produce a higher performance rate compared to the baseline system.

As stated above, this paradigm for assessing and comparing systems and tools for searching has been productive for developing more effective techniques for finding relevant documents. However, due to excluding human involvement in the search process the paradigm includes evident limitations like static information needs, orientation towards output of the system instead the whole search process or outcomes of the search. There is evidence that human performance in searching cannot be further improved by further improving traditional retrieval effectiveness (Järvelin 2011).

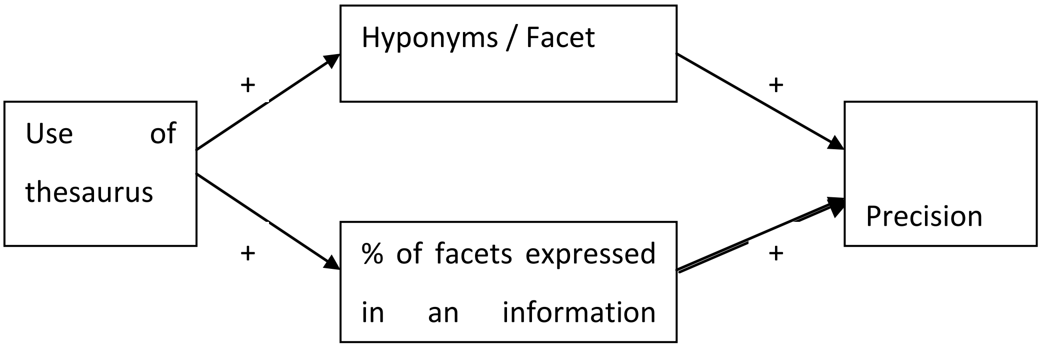
In this laboratory evaluation paradigm human involvement with system is excluded (Kekäläinen & Järvelin 2002). Development of interactive systems and increasing end-user searching has brought human searchers into the evaluation settings (Robertson & Hancock-Beaulieu 1992). To what extent has this changed the evaluation methodology used? For answering this question, typical research designs of interactive retrieval experiments are analyzed using the methodological rules proposed in evaluation literature (Rossi & al. 2004; Rubin 2006). Based on the results of this analysis, ideas for developing evaluation methodology for interactive retrieval experiments are proposed.

2. Goals, Outputs and Outcomes

The methodological rule in evaluation requires defining the goals of the object of evaluation, developing measures of goal attainment and criteria of success in attaining the goal (Rossi & al. 2004; Rubin 2006). In addition to explicit conceptualization of a system's objectives, it is necessary to explicate how they are supposed to be achieved (Rossi & al. 2004). This implies some kind of pre-understanding of the mechanisms connecting the use of the search tools evaluated to the explicated goals. How the use of the tools is associated to reaching the goals. Explication of the relations between the tools used and the intended goal attainment is called program theory or program model (Rossi & al. 2004). It includes a goal indicator as a dependent variable, the use of a tool as an independent variable (a treatment in an experiment), and possible additional, e.g. mediating or moderating variables. The program theory indicates how the system features and user interaction with them are supposed to achieve the intended objectives. If this cannot be plausibly explicated, it is difficult to infer what actually produced the level of goal attainment observed. Poorly specified program theory limits the ability to identify and measure the intervening variables on which outcomes may depend and correspondingly, the ability to explain what went right or wrong in producing the expected outcomes (cf. Rossi & al. 2004).

The following illustrates a program theory. The goal of a search is to maximize the precision of search results, i.e. to produce on the top of search results as many relevant items as possible (or some good hits, depending on the goal). It is known that other factors given there are two major means for increasing precision. First, increase in the number of specific terms in the query (more specifically in the facets of the query), will lead to improved precision, and second, comprehensive expression of query facets (concepts) will enhance precision. A known tool for supporting term selection is a thesaurus (or ontology). We may construe the following program theory for predicting precision by using a thesaurus (figure 2).

Figure 2 – ***A program model for increasing precision by the use of thesaurus***



In interactive retrieval evaluation experiments it is not common to explicate how the search support tool under evaluation is supposed to increase searchers' ability to produce better search results. It is mostly expected that the tool functions as a black box causing an improved value in search goal indicator. There are typically no hypotheses concerning which factors mediate the effect of the search tool to search results. Naturally, the research designs include variables, which could be though to act as mediating factors between the dependent (goal indicator like precision) and independent variables (treatment like use of a search tool). These variables include e.g. the number of search terms, the number of search iterations, i.e. queries, and sometimes the type of search terms used. Like stated above, without knowing, how the object of evaluation is expected to produce the expected search goal, i.e. which factors mediate the effect of the tool for achieving the search goal, it is not possible to improve the tool in a reliable way.

It is typical also to distinguish between the outputs and outcomes of a system or service to be evaluated. Outputs are the products delivered by a system, whereas outcomes are the benefits the system produces to its users (Rossi & al. 2004). In information retrieval experiments, outputs are relevant documents retrieved by the system. Outcomes are the benefits the system or system feature produces to searchers, or conceptualizing differently, the benefits searchers derive by using the feature observed. The benefits are usually changes in user's knowledge, skills, behavior, attitude, or condition that may not have happened without the system's support (cf. Rubin 2006). These benefits can be divided into two groups. It is possible to focus either on the benefits the system produces to users during the search process or on the benefits the information items retrieved produce to the searchers' task performance (Vakkari 2010). In the following we focus first on the benefits of the search process, and after that on the benefits of the search to searchers' task performance.

3. Criteria of Success in Evaluation

The aim of interactive retrieval evaluation experiments is often to find out to what extent the use of a search support tool contributed to successful searching. Success is assessed by indicating the degree to which this tool reached its goal. In interactive evaluation experiments it is untypical to give a clear outcome definition unless the system output, i.e. the number of relevant documents retrieved is not taken into account. It is common to reduce the goal achievement to high precision or recall or similar measures. Thus, the evaluation focuses on the quality of the result list.

It is very rare to find in experiments definitions of success, which would conceptualize success in terms of using the tool evaluated. What are the benefits the users derive from using that tool if the number of relevant items retrieved is excluded? What would be the intended outcomes of the tool? We may imagine a tool supporting term selection. The aim of the tool is to help users to identify major types of terms, hyponyms, hyperonyms and synonyms for query formulation. Let's suppose, that the search task requires the searchers to express the topic by very specific terms, i.e. by hyponyms. The outcome definition of the tool could be as follows: The use of the tool increases the number of hyponyms used in queries. The intended change in users' search behavior produced by the tool is that users are able to identify and use hyponyms in their queries. This definition allows us to infer measures for the success in using the tool. These measures could be the average number of hyponyms used per query or the average number of hyponyms used in a search session or the proportion of hyponyms of all query terms.

Although we are able to define the outcome of the tool and infer measures of success in using the tool, the criteria of success in using it are open. It is difficult, and likely impossible to establish absolute criteria in the case of term selection, and perhaps in general concerning searching. In our example, there are naturally ways to overcome this problem by using relative criteria. In given search tasks (topics), the facets, i.e. exclusive aspects of a request (Lancaster & Warner 1993), can be identified. Facets are typically expressed by search terms. If the aim of the tool is to support the use of hyponyms, then the criteria of success can be defined e.g. by establishing the number of hyponyms per facet, which is considered as success. We could consider expressing each facet in the request at least by one hyponym as success. This could be justified by the fact that searchers typically use short queries and that they do not use hyponyms.

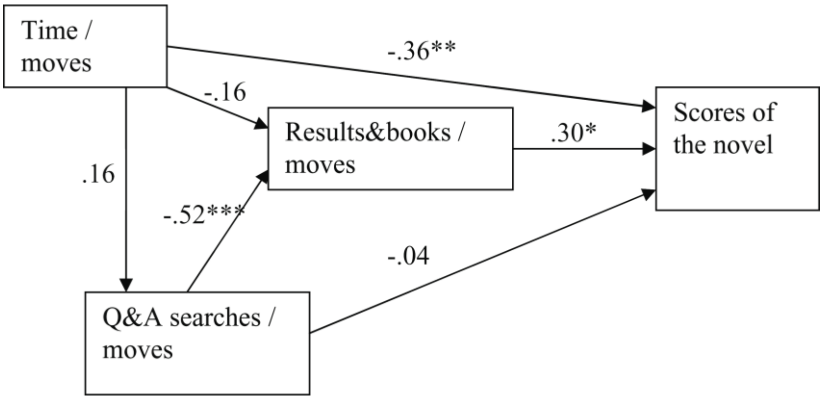
In interactive retrieval experiments the second option for the criteria of success would be to compare the average number of hyponyms in test and control systems. If the number of hyponyms is significantly larger in the test system, then it was more successful in reaching the goal observed.

Naturally, the use of search tools is a sub-goal serving the major goal of search tasks, which is to retrieve an optimal number of relevant items, or ultimately, obtain needed information. Thus, increase in the number of hyponyms in the query is not an end as such, but a sub-goal of the search task, which in the case of using hyponyms is to improve the precision of the search (cf. Lancaster & Warner 1993). However, although producing a high quality result list is a necessary condition for successful searching, the search process – including the use of search tools – is in its turn a necessary condition for a good retrieval result. Therefore, in evaluating search systems, it is important also to assess to what extent the search process variables reach their objectives, and through those objectives contribute to retrieval effectiveness (Vakkari 2010).

The test collection oriented model of evaluation has emphasized the output of the system as the indicator of search success. The following example indicates that it is not only the output as such that matters in evaluating search success but the whole search process and relations between the elements in that process.

We studied how an enriched public library catalogue supported fiction readers to find interesting novels. The enrichment included e.g. index terms in novels from a fiction thesaurus and tags by users and librarians. We asked the readers to search for interesting novels to read in a simulated situation when they did not have a clear idea what they wished to read. This is a typical situation in a public library. The results showed that the search process variables like free text or key word searches and the effort put on querying had no bearing on the search success, i.e. finding an interesting novel (figure 3). Deviating from the previous, the more effort the searchers invested in examining the search result list and book metadata, the more interesting novel they found (Oksanen & Vakkari 2012). Thus, our results indicate, that the querying process did not influenced the search success, but the effort in examining the search results. This hints, that in this case search success depends on the quality of the search results presentation. It is likely that enriching search result presentation so that it would help readers to decide about the value of the books would enhance search success. This means that focusing on a sub-goal of search, in our case search result representation, and not only search output, would improve search success. Thus, improving the achievement of a particular sub-goal in searching would increase search effectiveness.

Figure 3 – ***A path model for predicting the interest scores of the novel retrieved (n=58)***



Note: * = p<.05; ** = p<.01; ***=p<.001; Q&A = Quick and Advanced

4. Modes of Analysis

The notion of interactive information retrieval refers to interaction between human searcher and the features of a system. The system responds to human activities, and humans react to those responses. The interplay between humans and the system leads to the realization of the search goals.

Most of the interactive information retrieval experiments do not analyze how the interaction proceeds during the search sessions (Rieh & Xie 2006). A typical way to represent the results is to average all variables reflecting interaction over the whole search session. It is not analyzed how these variables vary and are associated within and between the queries within a session, and how they jointly contribute to the search success. However, there are some exceptions seeking to unfold the interaction process (e.g. Rieh & Xie 2006; Ruthven & al.

2003; Vakkari & al. 2004). It is typical to count e.g. the average number of queries, search terms and relevant documents within a search session. This kind of analysis reveals how interactions on average were associated to search output indicators, but it does not inform what was the process, that led to search results. It leaves open what kind of combination of queries and terms, i.e. search tactics produced the search results obtained. It seems that a heavy focus on search output has limited the interest in process evaluation. As a consequence, the options to understand the search as a process and improve system design enhancing search process are limited. If we can reliably recognize different types of query formulation behavior, then it would be useful to see if we can predict query reformulations that support this behavior and make these suggestions clearer at the interface level (Ruthven 2008).

An additional limitation in interactive retrieval experiments is variable by variable analysis of the results. The association between a dependent and each independent variable is given separately. E.g. it is shown that the average number of queries or the average number of search terms is associated to the indicators of the search output. It is untypical to use multivariate analysis for presenting how the dependent variable is associated to several independent variables, e.g. how the number of terms and the number of queries may interact. It is not possible to observe how searchers change their behavior during the process, and how this is related to the output of the search. Variable by variable analysis conceptualizes the search process as consisting of disconnected entities, which contribute separately to search output indicators. However, search process is an interconnected whole, where search process variables interact and influence conjointly on search output.

It could be argued, that the small number of cases in typical interactive experiments is not in favor using multivariate techniques. However, there are techniques like regression analysis, which does not loose degrees of freedom, and supports model building. Thus, the smallish number of cases does not necessarily restrict the use of multivariate analysis.

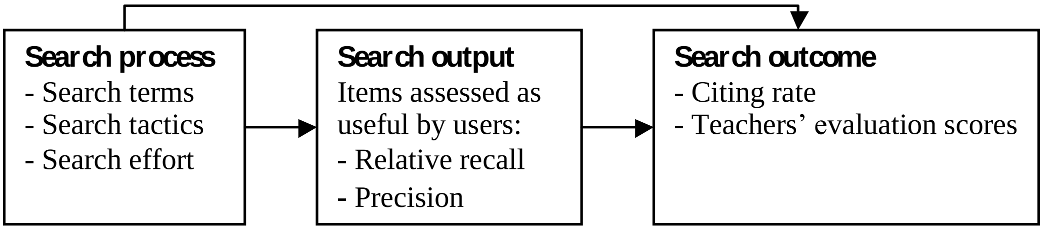
Variable by variable analysis restricts also the possibility to build models for predicting either search behavior or search output. Models are needed for giving a more accurate account of these factors for systems design. It is naturally possible to build a model using only one independent variable, but it is likely that a combination of independent variables produces a more accurate and powerful prediction with a greater proportion explained in the variance of dependent variable.

5. Evaluating information searching as a part of larger task

Information search systems produce benefits to users during the search process and also by providing information for the task that generated searching. Robertson (2008) has proposed, that "from the point of view of a user engaged in a larger task, the retrieval of items of information must at best be a sub-goal. Our understanding of the validity of this as a sub-goal, and how it relates to the achievement of wider goals, is limited and deserves more attention." Robertson is not the only one who has suggested evaluating information searching from the point of view of its contribution to task performance, the ultimate goal of information searching. Information is typically sought for proceeding in a task or for an interest. Unfortunately, there has been a handful of studies exploring how information searching benefits larger tasks like work tasks or learning tasks. They have typically analyzed how information retrieval influenced answering factual questions. The results have been inconclusive.

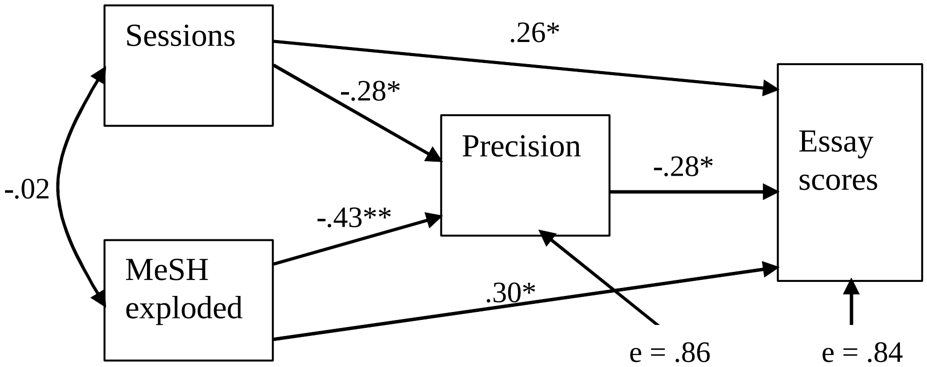
It is not only the quantity of relevant documents that matter in task performance. It is suggested that recall and precision can be placed among the many factors that may be associated with the ability to complete a task successfully (Hersh 2003). Does a higher precision or recall, i.e. search effectiveness, predict high quality task performance like a superior investment decision or a successful operation of a patient? This lack of information speaks for analyzing more closely how search effort is associated both with search output and outcome. Next I will present results of my study exploring how search effort is related to search output and outcome (Vakkari & Huuskonen 2012). In a field study we examined how students' search effort for an assigned learning task was associated with precision and relative recall, and how this, in its turn, was associated with the quality of learning outcome. We explored how information retrieval benefits students in writing an essay for a class in medicine. Figure 4 present our research setting.

*Figure 4 – **Research setting***



We studied how medical students searched for information for their class assignment by using Medline database. The assignment included answering a clinical problem by writing a 5-10 page essay. Searching information in Medline was part of their assignment. They were expected to use the information found in the assignment. The students were asked to attach printed search logs in their essays. Logs included information about search process variables and search results. The students were also asked to assess on the printouts the relevance of the references retrieved for their essays. We also elicited information about various background factors like familiarity with the topic or information concerning the searching by a questionnaire. The teachers of the course assessed the essays according to five criteria measuring the quality of the essay. The marks received reflect the students' ability to utilize the information in the items retrieved in the essays. It is expected that searching contributes to the quality of the essay. We also counted the citing rate of the references retrieved, i.e. the proportion of references retrieved cited in the essays. These two measures indicated the outcome of the searches.

*Figure 5 – **A path model for predicting essay scores (n=41)***



Note: * = p<0.05, ** = p<0.01

The major findings were surprising like figure 5 indicates. The results show that search effort was not associated to relative recall, while increasing search effort decreased precision, but led to better essays. Thus, the higher the quality of result list, the poorer the quality of task outcome, i.e. the essay. The more sessions the search consisted of and the more exploded MeSH terms the students used the poorer the precision, but the higher the quality of the essays. Students' efforts between the sessions to familiarize themselves with items retrieved improved their understanding of the task. This tended to increase their selectivity in accepting documents from the result list (Vakkari 2001), which lowered precision, but improved essays. The results hint, that students dwelling on querying in Medline instead of putting more effort into exploring the documents retrieved for the essay achieved higher precision, but poorer essay scores. This emphasizes the importance of assessing and exploring items retrieved for successful work task performance, whatever precision or recall achieved. Actually, this finding resembles our finding from the study on fiction retrieval that emphasis on examining search results contributes to search success.

Our findings confirm the suggestion that indeed recall and precision can be placed among the many factors, which may influence on the ability to accomplish a task successfully (Hesh 2003). Our model explained about 28 % of the variance of essay scores. Thus, other factors accounted for 72 % of the variance of essay quality. As the model includes also other factors (the number of sessions and the number of exploded MeSH terms), precision (and recall) cover much less than 28 % of the variance in essay scores. Sarcastically we may state that this is fortunate, because a greater account of variance by precision and recall would have led to a still worse task outcome.

Our study is one of the first steps to understand more in detail how search behavior and output are associated with the outcome of the task generating searching. We need more effort to explore the relationships between information searching and task performance in various professional and leisure settings and in various types of tasks. Only after an abundance of studies like ours we will be able to conclude more validly what is the role of information searching, precision and recall, in particular in task performance.

We may differentiate between four evaluation contexts: information retrieval contexts – lab context and interactive IR context – information seeking contexts, work task context, and socio-organizational contexts (Ingwersen & Järvelin 2005). The point of departure of the framing is the ultimate goal of information searching to enhance human task performance by providing information. We may infer for these various contexts varying evaluation criteria. We need evaluation results from all of these contexts. However, in information retrieval (evaluation) research the proportion of resources invested and consequently, results obtained between the contexts decreases radically from laboratory evaluation context to socio-organizational context. In order to evaluate more many-sided and accurately the success of information retrieval, the emphasis should be directed more to broader contexts of evaluation.

6. Conclusions

I have discussed various limitations in evaluating interactive information retrieval systems and suggested an extended methodology for overcoming these limitations. It is based on the methodological rules represented in the core evaluation literature like in Rossi & al. (2004). The point of departure is to explicate the goals of the object of evaluation, and derive explicit measures and criteria of success in attaining these goals. Also a distinction between outputs and outcomes of a system was made. Outputs are the products delivered by a system, whereas outcomes are the benefits users derive by using the system. In evaluating information retrieval it has been typical to reduce users' benefits into output indicators, i.e. indicators of the quality of the result list. However, it is important to assess how the specific search support tools benefit users and consequently, contribute to the quality of the result list. It is also necessary to extend the evaluation to include tasks, and assess to what extent the search process and search results benefit task performance.

A major implication of the proposed methodology is that in interactive retrieval evaluation it is critical to define the goals of the tools assessed in improving human performance in information searching. Without reflecting and defining the objectives of the system it is difficult to infer appropriate evaluation criteria (cf. Salton & McGill 1983).

The methodology recommended enhances the validity of experiments evaluating interactive information retrieval by analyzing the search process as a whole, by explicating in detail the goals of the system or search tool to be evaluated, and by facilitating the inference of the indicators and criteria for meeting the aims of the tool. The methodology encourages proactive building of logical models representing the mechanisms, which connect the use of the tool to its intended goal, and finally its contribution to the quality of the result list, and ultimately how the search process contributes to task performance. This, in its turn, is an essential condition for building models predicting users' search behavior for improving system design.

7. References

Belkin, N.J. (2010). On the evaluation of interactive information retrieval systems. In The Janus faced scholar. A festschrift in honor of Peter Ingwersen. ISSI, 13-22.

Borlund, P. (2000). Evaluation of interactive information retrieval systems, Åbo: Åbo Akademi University Press.

Hersh, W.R. (2003). Information retrieval: A health and biomedical perspective, 2nd ed. New York: Springer.

Ingwersen, P. and Järvelin, K. 2005. The Turn: Integration of Information Seeking and Retrieval in Context. Dordrecht: Springer.

Kekäläinen, J. & Järvelin, K. (2002). Evaluating information retrieval systems under the challenges of interaction and multi-dimensional dynamic relevance. In Proceedings of the 4th CoLIS Conference. Greenwood Village, CO: Libraries Unlimited, 253-270.

Lancaster, W., & Warner, A. (1993). Information retrieval today. Arlington, VA: Information Resources Press.

Oksanen, S. & Vakkari, P. 2012. Emphasis on Examining Results in Fiction Searches Contributes to Finding Good Novels. In Proceedings of Joint Conference on Digital Libraries 2012. (Washington DC, USA, June 10-13, 2011). JCDL 2012. ACM, New York, NY, 199-202. Doi 10.1145/2232817.2232855

Pharo, N., Nordlie, R., Fuhr, N., Beckers, T., & Khairun, N. (2009). Overview of the INEX interactive track. In: Lecture Notes in Computer Science 6203, 303-311.

Rieh, R. & Xie, H. (2006). Analysis of multiple query reformulations of the Web: The interactive information retrieval context. Information Processing and Management 42: 751-768.

Robertson, S., & Hancock-Beaulieau, M. (1992). On the evaluation of IR systems. Information Processing & Management 28: 457-466.

Rossi, P., Lipsey, M., & Freeman, H. (2004). Evaluation. A systematic approach. 7th ed. Thousand Oaks: Sage.

Rubin R.J. (2006). Demonstrating results using outcome measurement in your library. Chigaco: ALA.

Ruthven, I. (2008). Interactive Information Retrieval. In Annual Review of Information Science and Technology (ARIST). vol. 42, 43-92.

Ruthven, I., Lalmas, M., & van Rijsbergen, K. (2003). Incorporating user search behavior into relevance feedback. Journal of the American Society for Information Science and Technology 54(6): 529-549.

Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Vakkari, P. (2001). A theory of the task based information retrieval process: a summary and generalisation of a longitudinal study J.Doc. 57(1): 44-60.

Vakkari, P. (2010). Exploratory searching as conceptual exploration. In Proceedings of 4 th workshop on Human-computer interaction and information retrieval, 24-27. (http://research.microsoft.com/en-us/um/people/ryenw/hcir2010/docs/papers/Vakkari_fp10.pdf).

Vakkari, P. & Huuskonen, S. (2012) Search effort degrades search output but improves task outcome. Journal of the American Society for Information Science and Technology 63(4): 657-670. DOI: 10.1002/asi.21683

Vakkari, P., Jones, S., MacFarlane, A., & Sormunen, E. (2004). Query exhaustivity, relevance feedback and search success in automatic and interactive query expansion. Journal of Documentation 60(2): 109-127.

Wildemuth, B. & Freund, L. (2009). Search tasks and their role in studies of search behaviors. In Proceedings of 3rd workshop on Human-computer interaction and information retrieval, 17-21.