A Use Case Framework for Information Access Evaluation

Preben Hansen[1], Anni Järvelin[2], Gunnar Eriksson[1], and Jussi Karlgren[3]

[1] Department of Computer and Systems Sciences, Stockholm University
[2] School of Information Studies, University of Tampere
[3] Gavagai, Stockholm & School of Computer Science and Communication, KTH
{preben,gerik}@dsv.su.se,
anni.jarvelin@uta.fi,
jussi@gavagai.se

Abstract. Information access is no longer only a question of retrieving topical text documents in a work-task related context. Information search has become one of the most common uses of the personal computers; a daily task for millions of individual users searching for information motivated by information needs they experience for some reason, momentarily or continuously. Instead of professionally edited text documents, multilingual and multimedia content from a variety of sources of varying quality needs to be accessed. Even the scope of the research efforts in the field must therefore be broadened to better capture the mechanisms for the systems' impact, take-up and success in the marketplace. Much work has been carried out in this direction: graded relevance, and new evaluation metrics, more varied document collections used in evaluation and different search tasks evaluated. The research in the field is however fragmented. Despite that the need for a common evaluation framework is widely acknowledged, such framework is still not in place. IR system evaluation results are not regularly validated in Interactive IR or field studies; the infrastructure for generalizing Interactive IR results over tasks, users and collections is still missing. This chapter presents a use case-based framework for experimental design in the field of interactive information access. Use cases in general connect system design and evaluation to interaction and user goals, and help identifying test cases for different user groups of a system. We suggest that use cases can provide a useful link even between information access system usage and evaluation mechanisms and thus bring together research from the different related research fields. In this chapter we discuss how use cases can guide the developments of rich models of users, domains, environments, and interaction, and make explicit how the models are connected to benchmarking mechanisms. We give examples of the central features of the different models. The framework is highlighted by examples that sketch out how the framework can be productively used in experimental design and reporting with a minimal threshold for adoption.

Keywords: Evaluation, benchmarking, use cases, interaction.

G. Paltoglou et al. (Eds.): Professional Search in the Modern World, LNCS 8830, pp. 6–22, 2014.

1 Introduction

For decades, the Cranfield model [11, 31] has provided an effective backbone for information access research, offering a methodological vehicle for systematic and quantifiable evaluation and comparison of system components. This has contributed greatly to the success of the field, both in terms of research and in terms of practical application to task. However, the last two decades have seen a drastic broadening of information access system usage. Information access is no longer only a question of retrieving topical documents in a work-task related context. Document retrieval has become an embedded component in many systems which neither to their users nor their providers appear to be classic document retrieval systems: entertainment systems, communication platforms, time management systems, and the like.

This change in the information access landscape has rendered the classic Cranfield model insufficient as a framework for bringing together algorithm benchmarking with system and service validation: document retrieval performance is not necessarily what makes or breaks a service. Services may be popular, useful, and successful in spite of unimpressive retrieval components that are built to be satisfactory rather than optimal. Static test collections, viewed in a research context to be necessary for reproducibility of results, do not offer relevant data for testing fielded systems against a vast and vastly growing stream of human-generated data. Measurements of system quality based on classic benchmarking have thus become less reliable as a prediction mechanism for the systems' impact, user take-up, and eventual success in the marketplace. This is not news to the information retrieval field. Some of the very first discussions on the potential for interactive bibliographic retrieval pointed out the necessity of rich evaluation metrics [7, 8] and further contributions to that line of thought have continued by formulating ways to relate the usage at the interface to other human behaviour and the tasks users are concerned with to achieve a richer understanding of users, their intentions, sessions, and the evaluation thereof in formal, quantitative, or qualitative ways [5, 6, 18, 27, 34, 38] through more elaborate theoretical background models, better quantification or results, or the introduction of observational methodologies with a finer resolution better to model the task at hand [2, 15, 21, 28].

From this perspective, enriching the Cranfield-based approaches which abstract away from the user and usage situation, can be done using several contrasting approaches to evaluation. The different approaches form a continuum [14, 23]. At one end, we find laboratory based benchmarking evaluations, which seek to hold a maximal number of variables constant to be able to assess the effect of some variation as precisely as possible [31, 35, 37]. At the other end, naturalistic field studies using an ethno-methodological approach to understand the behaviour and preferences of real users with real information needs [25, 26, 33, 36]. In between a range of approaches: user studies with simulated information needs ranging from set queries to more comprehensive models of workplace tasks which users have been asked to emulate [9, 16]; and laboratory interaction simulations, which expand the user and interaction models of the traditional benchmarking evaluations [1, 3, 4, 22, 24, 32, 39].

However, performing more naturalistic user studies is not enough if they do not build up a successive body of knowledge which can be put into use for building practical systems. If we wish to see research efforts published in our field to remain relevant to commercial service providers, we must include the embedded information access components of various systems, in various contexts and domains, and cover more varied user communities, search tasks, and goals. What we still need is a framework to integrate all these components, to support richer and broader benchmarking and bring together benchmarking with system and service validation, including current research in human-computer interaction and support for industrial and commercial concerns. In this chapter, we propose such a framework based on use cases and user centered design principles.

2 Use Cases as a Model for Interaction

Use cases are a user-oriented software development methodology, first developed by Ivar Jacobson and colleagues [19, 20] for capturing interaction-based functional requirements in software development, and further developed by others, e.g. [12, 13]. Use cases are intended to capture a user's point of view; technical solutions or system implementation are not considered in a use case. The requirements are documented by describing how a user interacts with a system to carry out a task, or to reach a goal. The focus is on task modeling[1] or modeling one kind of use that a system can be put to, given a specific user role. Users may normally use a system in several ways and for different purposes. To be practical, use cases focus on a specific kind of system usage, instead of trying to cover all possible different interactions and goals. Numerous approaches to system development and software engineering, commercial and academic, consultancy-based and programmatic, take use cases as a point of departure; many leverage the information in use cases for testing protocols and quality assurance. Typically evaluation metrics in software engineering are closely tied to system effectiveness and are used as performance indicators. The aim is to verify and test functional behaviour when the system under consideration is scaled up from development operation to actual usage and to monitor system behaviour during subsequent versioning.

In the use case based framework presented in this chapter, observable patterns of human information access behavior are described through a selection of variables that can be linked to properties of the experimental design and to the system and interface features of the evaluated systems, as illustrated in Figure 1. The features of interest could be system performance variables such as those typically measured in software engineering, but in our framework, we reach further into the use case to allow for features which measure user, context, and task-related aspects of usage.

---

[1] "Task" in use cases differs from the "work tasks" often discussed in information access literature: use cases focus on users' immediate tasks when interacting with systems, the task the user expects the system to support and not the broader work tasks that the users are engaged in.

The evaluation framework presented in this chapter integrates the features affecting information access system usage, with the constraints presented by the system and interface design on one hand and experimental design on the other. This way the framework can indicate evaluation approaches for measuring the value of an information access system to its users given some real-world constraints of the system usage, and can describe what kind of real-world information access system usages the results of a specific experiment can apply to.

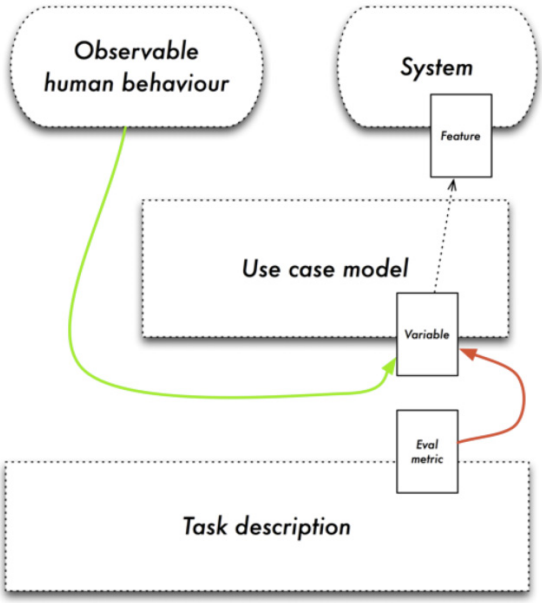

Fig. 1. Relating human behavior to system and evaluation features

The framework assists evaluation design by supporting explicit mapping between relevant features of information access system usage on the one hand, and experimental design decisions and benchmarking mechanisms on the other. This is done along a number of dimensions, held together in larger bundles of features: Interaction, Interface and System, Background, and Evaluation, cf. sections 3 and 4. The framework is called a "use case framework", as the use case, a model of system usage through the description of the user-system interaction, is at the very heart of the framework. It is in the interaction model, described in section 3.1 that the constraints and demands related to the users and usage of systems meet the evaluation mechanisms: the characteristics of the envisioned users, their tasks, contexts and environments all affect what interaction sequences are relevant to consider in evaluation. The background features, described in section 3.3, cover these aspects. In contrast with the original purpose of use cases, this is an evaluation framework and not a system design methodology. The interface and system features of the operational systems under evaluation, or of the experimental systems as defined in the experimental design, constrain the possible interaction patterns for a use case and thus limit the validity of the evaluation with

respect to the users, search tasks, domains and environments covered. They are therefore described separately in the interaction and system model (section 3.2).

For each of the three feature bundles, a corresponding checklist has been formulated to support thinking about, designing, and documenting a certain aspect of information access usage or evaluation, as well as noting dependencies to other aspects. In section 4, the checklists are put to use in two examples of evaluation design. In the following section, the components of the use case framework are discussed on a more general level.

3 Modelling Usage

3.1 Modeling Interaction

In the use case framework, a model of the interaction between a user and a system forms the interface between the background models and evaluation. Interaction is limited both by the background conditions and by the interface properties defined in an experiment, but the model of interaction also carries forward the requirements of the background and the interaction to the experimental design.

Correctly modeling the ways in which users interact with a system is essential for establishing the success criteria for an evaluation. An interaction model connects user goals to interaction sequences, and depicts the complexity of typical search sessions: search and result inspection strategies, result use, iterations of query reformulations, goal-orientation or randomness of the interaction. These aspects affect what results the users are likely to encounter and find relevant, given a certain time or effort of searching. They should therefore be reflected in both test collections and evaluation measures.

Use cases provide a useful framework for thinking about interaction in information access evaluation. There is no single established way of writing use cases, but use cases are typically organized around a main success scenario describing the simplest successful interaction sequence through the use case. The sequence is commonly presented as ordered steps, where each step describes one interaction between the user and the system. The main success scenario is complemented by a set of extensions that describe all the other possible interaction sequences through the use case, including any alternative user actions, exceptions and failures. A typical search use case may have a simple main success scenario (1. User types a query; 2. System shows results; 3. User clicks on a result; 4. System presents result), but very many paths through the use case are possible due to the high degree of freedom of user actions. Thus iterations of the different user actions in varying order need to be modeled through extensions.

The number of interaction sequences (main success scenarios and extensions) needed for describing most information access system usages is limited however: the number of identifiable user actions is not very high, and while the number of possible paths through the use cases might be overwhelming, the types of iterations of and switches between the actions are limited and thus possible to model through a limited number of interaction sequences and extensions.

The interaction sequences are here structured following [13, 40], by dividing the scenarios into user intentions and system responsibilities that show what the user aims to do in each step of the interaction and what system responsibilities relate to each user intention. Figure 2 depicts an example of a structured main success scenario for a use case for finding an illustrative image to insert in a blog post.

insertingIllustration

USER INTENTION	SYSTEM RESPONSIBILITY
request illustration	
	show appropriate images
select image	
	show preview
confirm	
	insert image
	close
EXTENSIONS	
browsingResults	
reformulatingRequest	

Fig. 2. Example main success scenario

A goal in use cases refers to the concrete, immediate goal of a user interacting with the system, such as inserting an illustration in the above example. It defines the expected outcome of interactions and thus introduces the immediate use of information as a factor affecting evaluation criteria. Goal categories with clear impacts on interaction patterns have been recognized in previous studies, mainly based on analysis of web search logs [10, 30]. These categories offer a solid starting point for considering goals, even if new categories to cover more varied usage and more specific goals may be needed. We separately define a second aspect of user goals following Ingwersen and Järvelin [18], i.e., the type and amount of information looked for: single items or several items; ready answers, facts or notifications, or for topical content from which information can be extracted by the user.

3.2 Modeling Interface and System

Interface design is closely tied to the interaction model, as even experiments where no users or interface designs are purposely included make assumptions concerning the user interface and system functionality: depending on how the experiments are set up, the functionality may be fixed to e.g. a certain type of request formulation, or a specific type of result presentation. Such assumptions have a major effect on the applicability of the evaluation results and should be carefully modeled.

From the use case example in Figure 2, three types of user actions and thus three groups of interface and system features may be identified: request formulation, result presentation (in two levels: a set of ranked results, and image preview), and result use

(inserting image). The interaction model then needs to be completed with a detailed (black-box) description of the interface features affecting the user's interaction with the system in these interaction points. The relevant aspects may include e.g.:

- Supported means for expressing requests: by querying or browsing; using different modalities; querying by examples or specifying queries by e.g. typing or humming.
- The granularity of the searchable information items: can queries target individual images, or (curated) collections or sets of images, or details in images, etc.
- Organization and presentation of the results: textual or visual results; thumbnails or full images, with context and copyright information, or without, etc.
- Result use such as manipulation, sharing, onsite consumption, exporting, ordering, etc.

3.3 Modeling Background

Individuals perceive their information needs subjectively and the way they interact with information access systems depends on their goals, personal characteristics, and attitudes. While some of the differences are genuinely individual, the users' group membership offers a strong signal of their possible needs and goals. User role models then define (abstract) user groups with respect to specific system usages. They are based on the tasks that users in specific roles are trying to accomplish while interacting with the system, but also describe the shared characteristics of those users, their interaction with the system and the information exchanged between the system and the users. The central user role model features include:

- User features, such as: user demographics (age, gender, education, social status); user knowledge and skills (with respect to the task, domain, system, language); physical characteristics ((dis)abilities); orientation and attitudes (towards the task, the system, co-searchers).
- Interaction features, related to the complexity, predictability, and frequency of the interaction; locus of control of the interaction, and information flow direction.
- Information features, related to the volume and complexity of the information exchanged between the user and the system, as well as the clarity of the users' information needs.
- Users' primary success criteria, including: efficiency and effectiveness, system reliability and comprehensibility, actionability (does results enable taking intended action?).

Information access interactions are constrained by the activities that trigger them. A domain model captures the different constraints that govern a domain of activity: how the search behavior and goals of users are constrained by the activity at large (e.g. the "work" task) and the topic of interest; by the professional, private, or social context of the activity (presence or absence of peers or collaborators while searching and sharing results with others); or by the characteristics of the data and repository accessed. A domain model may define e.g.:

- The cost of errors if search task is not duly completed (economic, social, societal, career, etc.).
- Time restrictions limiting the length of the interaction.
- Restrictions to accessing the contents of the repository (access rights, cost).
- Data and repository features such as media, genre, language quality, and dynamics of the information and repository.

Moreover, different surroundings trigger different information needs and different interactions. The physical surroundings in which a user interacts with a system affect the search goals and the preferred way of interaction. An operational environment model depicts factors related to the surroundings, mobility, and locality of the users, distractions from the search interaction, and issues related to devices and network connections. The factors include, e.g.:

- Mobility and geo-position of the users
- Device and network restrictions (small screens, limited input ergonomics, high cost, or low speed of data transfer)
- Distractions (interruptions, multiple parallel tasks, noise)

4 Evaluation

So, how do these models facilitate systematic construction of experiments based on rich models of users, domains, environments and interaction? The goal is a framework that can make explicit the functional requirements and success criteria of information access systems, and to connect them to benchmarking mechanisms, i.e., to the components of experimental settings and the criteria and metrics used for measuring system performance. Figure 3 depicts how the models are brought together.

The background models (user, domain, environment) collect the information needed for understanding the users' success criteria, and describe the preconditions of their interaction with the system: their abilities and preferences when it comes to formulating queries, inspecting results, and interpreting and processing information. This information is then used in the design of experimental settings: for defining relevant information need (e.g. topics) and query types; the test data, relevance criteria, and characteristics of the relevance assessors; interaction patterns that need to be modeled; and system interface features to cover.

The success criteria for the users under consideration together with the interaction and interface models are needed for defining reasonable evaluation criteria. Evaluation must also be based on what results are likely to be retrieved when interacting with the system: Even if high recall is a prioritized success criterion for users, there is no point to base evaluation on users ploughing through an entire result lists for one-shot queries if users typically search in sessions of several fast query reformulations and shallow result scans. The evaluation criteria as described through the interaction patterns are then operationalized in suitable metrics. Patience, time or cost parameters may be added into the standard metrics [e.g. 21, 28], but probably yet new metrics need to be developed for measuring the quality of systems, given the varied success

criteria of users. The models and the process of mapping their features into experimental design can quite easily be formulated as easy-to-use checklists, similar to those used for documenting software system requirements, as implied in Figure 3.

To give an example, a classic TREC-style batch experiment starts from topics which describe well formulated, clear, topical information needs[2]. It extracts verbose keyword queries from textual topic descriptions. These are tested against static test collections with relevance assessments made by human expert assessors based on static relevance criteria. System performance is evaluated over ranked lists of document pointers returned by the system. Users' interaction with the system is modeled as sequences of one-shot queries and perusing the result list. The main success criterion used is effectiveness, as measured by MAP.

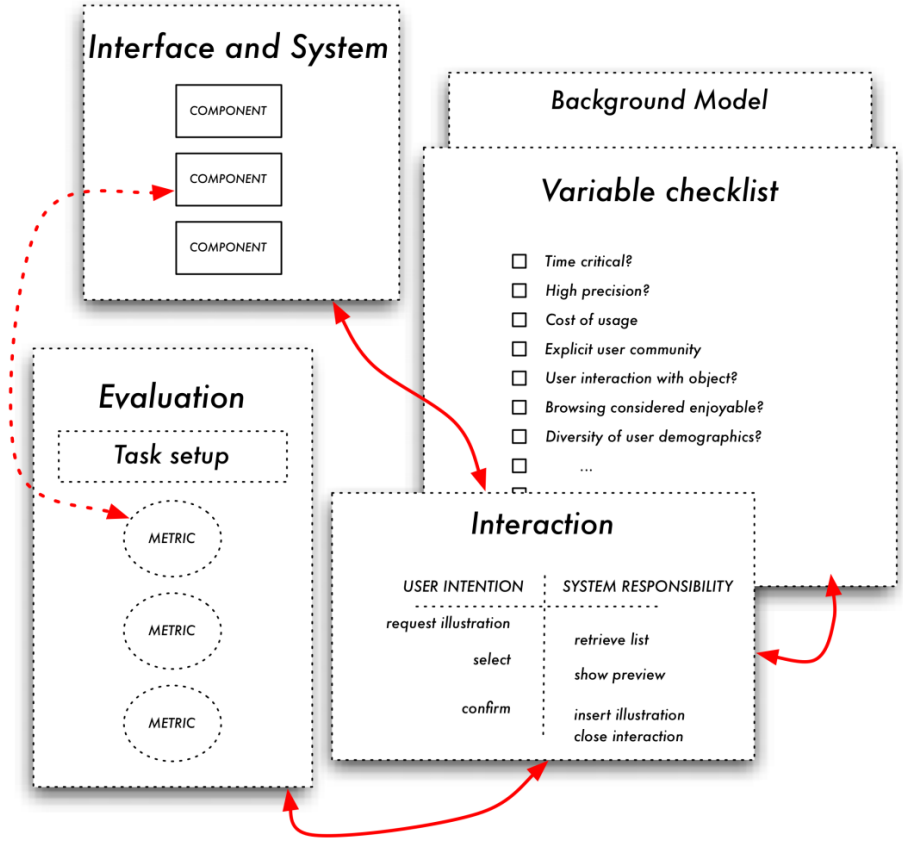

Fig. 3. Bringing it all together

[2] This example describes the classic experiment. Much more varied tasks, data, topics, and relevance criteria are covered in the present day evaluation campaigns in TREC, CLEF, NTCIR, and the like.

Table 1 summarizes the components of this kind of an experiment and lists some of the use case features that are (often implicitly) defined by the experimental setting. This is potentially a useful experiment for evaluation of the quality of a ranking component in a search system for a use case describing professional search tasks (e.g. on the patent domain), where the cost of missing relevant documents may be high and users are thus willing to spend considerable effort in formulating their queries and working down result lists.

Table 1. Evaluation task summary for a classic (stereotypical) TREC experiment. Depicts the components of the experimental setting and how they relate to an underlying use case.

Component	Use case features considered	Instantiation of the component
1. Test subjects	N/A	No test subjects. Minimal user model (not explicitly based on any specific users) reflected in topics, requests, relevance criteria, and metrics.
2. Topics	User role; clarity of information need; volume and complexity of information.	Topical, clear specifications of information needs and relevance criteria; created by experts.
3. Requests	User proficiency, domain know-ledge, language skills. Supported search strategies, query formulation means and modality.	Verbose, ad hoc, keyword queries.
4. Data	Repository: media, genre, language, technical quality, source dynamics. Data volume and complexity.	Static test collection of full text documents. Relatively noise-free and well-defined: clear definition of "document", few errors, standard language. Documents are independent of each other.
5. Ground truth creation	Users' domain and topic knowledge, language skills. User goals and roles.	Pooling; Manual relevance assessments using (binary liberal) relevance criteria by expert assessors.
6. Result presentation	Result presentation; user-system/ information interaction.	Ranked list of document ID's. Interaction purely based on rank.

Table 1. (*Continued*)

7. Interaction	User actions and system responses; Complexity and predictability of interaction; Users' goal-orientation and motivation, likelihood abandoning system; Restrictions.	Simple interactions of one-shot queries and deep scanning of results. Interaction is minimal and driven by the user. Patient user, no time restrictions. The encountered documents do not affect user behavior.
8. Result use	N/A	Not considered in the experiment.
9. Evaluation criteria & metrics	User goals; success/ failure criteria; motivation. Restrictions.	Ranking and recall in the absence of time or effort related restrictions; Finding as many relevant documents as possible. Operationalized as MAP.

It does not however capture the general success criteria for arbitrary other use cases. For example, a system where users access information objects for entertainment with no clear task-related information need in mind and where the browsing itself is part of the use and enjoyment of the system and where one of the central goals of interaction may be participating in a community of users, and possibly contributing to that community and to the collection needs to be evaluated using entirely different metrics [e.g., 29]. Main success criteria for such system would be e.g. high levels of user engagement manifested as users returning to the site; long sessions with protracted browsing; user adoption of site terminology and categorization schemes; and numerous user actions, such as up-votes, comments, and share actions in response to returned item lists.

To contrast with the stereotypical Cranfield experiment, Table 2 presents a (constructed) example experiment for evaluating the search component of a social video search service in the context of the typical sessions of system use. To some extent, different use case features are considered than in the Cranfield experiment presented in Table 1. The major differences are in how the components of the experiment are instantiated, when the evaluation is based on a different type of a user task or goal: The users' general task is to spend a short period of time on the service, finding something interesting to view, and interacting with their peers. Result use is an internal part of the search session, rather than something which occurs after the session. The search interaction is a success from user perspective if the user experience was pleasant and involved active participation in the social context.

The information access component is then evaluated based on (simulated) sessions [24] of information access and use with a variety of user actions included in the session model; with a test collection of linked data ranked by actionability - the number of views, comments, votes and shares the documents have received; and measured based on a model of social interaction and gains in a time based evaluation.

The topics describe unspecific and through the search session evolving information needs - the search topic per se is not necessarily very important, but serves as an entry point to the service, where social relevance weighs heavily. Requests reflect the users' evolving understanding of the current vocabulary and conceptual model presented by the system, while the interaction patterns in general reflect the actionability of the encountered documents (social potential; peers' preferences and actions).

Table 2. Evaluation task summary for social video search: an experiment focusing on the effect of the search component on the perceived social gain and enjoyability of sessions of system use

Component	Use case features considered	Instantiation of the component
1. Test subjects	N/A	No test subjects. Users modeled through ground truth creation, interaction model and evaluation criteria.
2. Topics	User role(s), goals, clarity of information need.	Topics describe entry points to the service. They might be topically more or less specific: from known item search to very general. Each topic contains a few alternative entry points: query words, concepts or directions to search.
3. Requests	User goals. Users' service proficiency and domain knowledge. Supported search strategies, query modalities and query formulation means.	Keyword queries of varying length and quality, evolving through sessions. Reformulation guided in the interaction model as probability of query reformulation given a result, and the entry points listed in the topics.
4. Data	Repository: media, genre, language, tech. quality, source dynamics. Data volume, complexity.	Linked data with documents and related likes, comments, tags.
5. Ground truth creation	Users goals and success criteria	Extracted from test data based on user engagement: documents ranked based on the number of responses or actions they have triggered.
6. Result presentation	Result presentation; user-system and user-information interaction.	Only vaguely modelled through possible user actions in the interaction model.

Table 2. (*Continued*)

7. Interaction	Possible user actions and system responses; Complexity and predictability of interaction; Users' goal-orientation and motivation; Restrictions (time, cost, effort, social); Probability of changing role.	Modeled as probability of an encountered document triggering user actions (query formulation, browsing, perusing result, viewing video clips, commenting, up-voting, sharing).
8. Result use	Probability of user changing role. System features for enriching, use and sharing of content. User goal.	Result use is an inseparable part of the interaction. Viewing content, up-votes, comments, recommendations.
9. Evaluation criteria & metrics	User goals; success/failure criteria; motivation. Restrictions.	Actionability. Level of user engagement, time spent interacting with the results. Evaluated based on a model of costs and gains (good/bad time; social gain).

Note that not all models needed for conducting this experiment are necessary in place yet: a useful model of unpredictable interaction sequences of many possible user actions might be difficult to define. Isolating or correctly modeling the roles of the different user actions or system components for the flow or success of the interaction might be difficult based on our current knowledge. Modeling the social gain connected to different user actions, or combining the dual success criteria of social gain and having enjoyable time requires understanding of the user population and of social dynamics. These difficulties point to areas where more basic research is needed on how and why users interact with information.

If one were to evaluate the social video service search component using a standard Cranfield experiment as described in Table 1, measuring performance with respect to user goals and success criteria (social gain and having a pleasant time) would not be possible. One could evaluate how well the ranking component ranks topically relevant video clips. Changing the ground truth creation, one could evaluate how well the ranking component ranks socially relevant video clips (given that we could model social relevance satisfactorily). A different metric could be used for operationalizing the evaluation criteria for measuring e.g. the topical diversity of the top results with highest social relevance. These evaluations could be both useful and motivated in many situations, not least for the sake of their viability. They do not however evaluate the same thing as the experiment described in Table 2. Being aware of these differences is important both when designing experiments and when reporting (or reading about) them, and this is where the suggested use case framework can be useful: The goal of the use case framework is to support the analysis of the use case, to suggest possible ways of connecting use cases to experimental designs and to make explicit

how the choices and simplifications made in experimental design affect the applicability and realism of the evaluation results.

5 Towards a Framework

Most experimental designs by necessity compromise between the breadth and the depth of their coverage: an experiment that aims to cover all users and all usages of a system, typically says very little concerning the systems' performance given any specific users or usages. On the other hand, the results from in-depth studies concerning the system usage patterns of specific user groups working on specific tasks are most often difficult to generalize or to transfer to other situations.

The variation in the basic interaction sequences occurring in information access systems is however limited enough to be modeled through a set of predefined interaction sequence templates. Instances of information access usage can thus be described as use cases within a use case framework and related to other instances through their shared interaction sequences. A carefully constructed model of the relationships between the interaction sequences can then notably reduce the complexity of the "evaluation landscape" by bringing together the at first glance different information access use cases that ultimately are characterized by shared interaction patterns and goals and consequently, shared evaluation criteria.

Such a framework facilitates the generalization and re-use of evaluation results of the limited in-depth evaluations in other contexts and thus provides a platform on which evaluation criteria and evaluation results can be described, debated and validated. As more use cases are described, evaluated and validated within the use case framework, the knowledge of characteristics of use cases - with respect to evaluation and success criteria - will be enriched, and the connections between distinctive use case features and patterns of interaction and success criteria become clearer.

6 Conclusions

There are many different approaches to evaluation of information access systems. Selecting the most appropriate approach must be done with attention to the use case, but also on the target (component, complete service), and the perspective of the evaluation (goals of end users, goals of customers, and goals of service providers). Essentially, all types of evaluations benefit from carefully modeling the success criteria and interaction patterns for the evaluated systems. While focusing on improving the performance of isolated system components is motivated in some phases of technology development, such evaluations should not be agnostic about the end user benefits achievable (or not) by further improvements of the components.

We do not claim that all information retrieval evaluations should add a number of variables concerning users with preferences and strategies for interacting with information retrieval systems in their experimental setting: the controlled and manageable experimental settings are one of the main strengths of the laboratory model. Instead, we claim that all information retrieval evaluations should be explicit about what they

evaluate and what they believe is the applicability of the results. If the context and the purpose of the evaluation is not carefully considered, it is difficult to choose the correct evaluation measures to be used. Better description of the context of a specific evaluation also makes it easier to organize and re-use the results and thus supports the growth of knowledge and technology take-up.

Acknowledgements. This work was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 258191 (PROMISE Network of Excellence).

References

1. Ahlgren, P.: The effect of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database. Academic dissertation. Valfrid, Sweden, 166 p. (2004)
2. Azzopardi, L.: The Economics in Interactive Information Retrieval. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 15–24. ACM (2011)
3. Baskaya, F., Keskustalo, H., Järvelin, K.: Time Drives Interaction: Simulating Sessions in Diverse Searching Environments. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 105–114. ACM (2012)
4. Baskaya, F., Keskustalo, H., Järvelin, K.: Modeling Behavioral Factors in Interactive Information Retrieval. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 2297–2302. ACM (2013)
5. Bates, M.J.: Information Search Tactics. Journal of the American Society for Information Science 30(4), 205–214 (1979)
6. Bates, M.J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. Online Review 13 (October 1989)
7. Bennett, J.L.: Interactive bibliographic search as a challenge to interface design. In: Walker, D.E. (ed.) Interactive Bibliographic Search: The User/Computer Interface, pp. 1–16 (1971)
8. Bennett, J.L.: The user interface in interactive systems. ARIST 7, 159–196 (1972)
9. Borlund, P.: The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information Research. An International Electronic Journal 8(3) (2003)
10. Broder, A.: A taxonomy of web search. ACM SIGIR Forum 36(2) (2002)
11. Cleverdon, C.W., Keen, M.: Cranfield CERES: Aslib Cranfield research project - Factors determining the performance of indexing systems. Technical report (1966)
12. Cockburn, A.: Agile software development. Addison-Wesley (2002)
13. Constantine, L., Lockwood, L.: Software for use: A Practical guide to the models and methods of usage-centered design. Addison-Wesley (2006)
14. Fuhr, N., Belkin, N., Jose, J., van Rijsbergen, K.: Interactive Information Retrieval. Dagstuhl Seminar Proceedings: number 09101. ISSN 1862-4405. Schloss Dagstuhl -Leibniz-Zentrum fuer Informatik, Germany (2009)
15. Hansen, P., Järvelin, K.: Collaborative information retrieval in an information-intensive domain. Information Processing and Management 41(5), 1101–1119 (2005)

16. Hansen, P.: Work task-oriented studies on IS&R processes. Developing theoretical and conceptual frameworks to be applied for evaluation and design of tools and systems. In: Fisher, K., Erdelez, S., McKechnie, L. (eds.) Theories of Information Behaviour. ASIST Monograph series, pp. 392–396. ASIST, Medford (2005)
17. Hearst, M.: "Natural" Search User Interfaces. Communications of the ACM 54(11), 60–67 (2011)
18. Ingwersen, P., Järvelin, K.: The turn: Integration of Information Seeking and Retrieval in Context. Springer, Dortrecht (2005)
19. Jacobson, I.: Object-oriented development in an industrial environment. In: Procceedings of OOPSLA 1987: Sigplan Notices, 22(12) (1987)
20. Jacobson, I., Christerson, M., Jonsson, P., Overgaard, G.: Object-Oriented Software Engineering: A Use Case Driven Approach. Addison-Wesley (1992)
21. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)
22. Kanoulas, E., Carterette, B., Clough, P., Sanderson, M.: Evaluating Multi-Query Sessions. In: Proceedings of 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1053–1062 (2011)
23. Keskustalo, H.: Towards Simulating and Evaluating User Interaction in Information Retrieval using Test Collections. Ph D Dissertation. Univeristy of Tampere: Acta Universitatis Tamperensis 1563 (2010)
24. Keskustalo, H., Järvelin, K., Pirkola, A., Sharma, T., Lykke, M.: Test Collection-Based IR Evaluation Needs Extension Toward Sessions - A Case of Extremely Short Queries. In: Lee, G.G., Song, D., Lin, C.-Y., Aizawa, A., Kuriyama, K., Yoshioka, M., Sakai, T. (eds.) AIRS 2009. LNCS, vol. 5839, pp. 63–74. Springer, Heidelberg (2009)
25. Kuhlthau, C.: Inside the Search Process: Information Seeking from the User's Perspective. Journal of the American Society for Information Science 42(5), 361–371 (1991)
26. Kumpulainen, S., Järvelin, K.: Barriers to Task-Based information access in molecular medicine. Journal of the American Society for Information Science and Technology 63(1), 89–97 (2012)
27. Liu, J., Belkin, N.: Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In: Proceedings of the 33th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 26–33. ACM (2010)
28. Moffat, A., Zobel, J.: Rank-Biased Precision for Measurement of Retrieval Effectiveness. ACM Transactions on Information Systems 27(1) (2008)
29. Murdock, V., Clarke, C., Kamps, J., Karlgren, J.: Proceedings of SEXI 2013 - Workshop on Search and Exploration of X-Rated Information at WSDM 2013 (2013)
30. Rose, D., Levinson, D.: Understanding user goals in Web search. In: Proceedings of the 13th International ACM Conference on World Wide Web, pp. 13–19. ACM (2004)
31. Sanderson, M.: Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval 4(4), 247–375 (2010)
32. Smucker, M., Clarke, C.: Time-based Calibration of Effectiveness Measures. In: Proceedings of 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 95–104. ACM
33. Spink, A., Saracevic, T.: Interaction in information retrieval: Selection and effectiveness of search terms. Journal of the American Society for Information Science 48(8), 741–761 (1997)
34. Su, L.T.: Evaluation Measures for Interactive Information Retrieval. Information Processing and Management 28(4), 503–516 (1992)

35. Tague-Sutcliffe, J.: The pragmatics of information retrieval experimentation, revisited. Information Processing and Management 28(4), 467–490 (1992)
36. Vakkari, P., Hakala, N.: Changes in Relevance Criteria and Problem Stages in Task Performance. Journal of Documentation 56(5), 540–562 (2000)
37. Voorhees, E.M.: The philosophy of information retrieval evaluation. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 355–370. Springer, Heidelberg (2002)
38. White, R., Dumais, S., Teevan, J.: Characterizing the influence of domain expertise on web search behavior. In: Proceedings of the Second International Conference on Web Search and Data Mining, pp. 132–141. ACM (2009)
39. White, R., Jose, J., van Rijsbergen, K., Ruthven, I.: Evaluating Implicit Feedback Models Using Searcher Simulations. ACM Transactions on Information Systems 23(3), 325–361 (2005)
40. Wirfs-Brock, R.: Designing Scenarios: Making the Case for a Use Case Framework. Smalltalk Report (November-December, 1993)