Use Cases as a Component of Information Access Evaluation

Jussi Karlgren
jussi@sics.se

Anni Järvelin
anni@sics.se

Gunnar Eriksson
guer@sics.se

Preben Hansen
preben@sics.se

Swedish Institute of Computer Science
Box 1263, S-164 29 Kista, Sweden

ABSTRACT

Information access research and development, and information retrieval especially, is based on quantitative and systematic benchmarking. Benchmarking of a computational mechanism is always based on some set of assumptions on how a system with the mechanism under consideration will provide value for its users in concrete situations – and those assumptions need to be validated somehow. The valuable effort put into those validation studies is seldom useful for other research or system development projects. This paper argues that use cases for information access can be written to give explicit pointers towards benchmarking mechanisms and that if use cases and hypotheses about user preferences, goals, expectation and satisfaction are made explicit in the design of research systems, they will can more conveniently be validated or disproven — which in turn makes the results emanating from research efforts more relevant for industrial partners, more sustainable for future research and more portable across projects and studies.

Categories and Subject Descriptors

H.5 [**INFORMATION INTERFACES AND PRESENTATION**]: User Interfaces—*benchmarking, evaluation*; H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

Keywords

Use cases, evaluation, validation, benchmarking, information access

General Terms

Theory

1. BENCHMARKING AND VALIDATION

Information access research and development, and information retrieval especially, whatever the media type under consideration, is based on quantitative and systematic *evaluation* as the main vehicle of research. Most typically, the evaluation follows the Cranfield model.[7] This is a *benchmarking* practice. A test set of pre-assessed target documents is used as a *benchmark* or *gold standard* for some collection, under the assumptions that an *information need* can be formulated satisfactorily and appropriately; that documents can be assessed as being *relevant or not* (or more or less relevant) for some given information need; that the relevance of a document with respect to that information need is *independent* of other documents in the collection, based solely on the qualities of that document; and that search engine users want to find as many relevant documents on the topic as possible.. A system can then be evaluated after how well it delivers results in conformance with the benchmark.

In other words, the benchmarks following the Cranfield paradigm focus on testing and comparing the information retrieval algorithms' capability of identifying and ranking topically relevant documents given a well- defined information need. The strength of this evaluation methodology is that it creates a controlled test setting and stable evaluation measures for meaningful comparisons of the information retrieval engines' performance on these tasks. It is good practice and has served the field well over a period of time within which information retrieval has positioned itself as one of the most important application areas of information technology and computer science. The benchmarks nevertheless abstract evaluation away from variation of factors such as the goal of the user, situation, context, user preferences or characteristics, interaction design, network latency and other such system-external qualities, systematically and intentionally ignoring factors relating to human behaviour and human interaction with information systems.

Benchmarking is thus only one part of evaluation. Validating the starting points is as important: the users are using information retrieval systems to support their actions in daily life, for entertainment, education and in professional tasks. Investigating if the developed tools and technologies (and the design principles behind them) actually work for the tasks they are envisioned to address — if the machinery delivers performance when it is moved from the workshop into the production environment it is designed for — is also needed.

In our field, this means testing a system in the field or in field- like conditions. User studies often implement an end-to-end system and have a number of test subjects use the system for a brief while in a laboratory environment with more or less realistic tasks assigned to them. This sort

DESIRE'11, October 28, 2011, Glasgow, Scotland, UK.

of study may be useful to evaluate the ergonomics of some specific interface widget, but they certainly are very unlikely to provide purchase to establish the validity of the starting points of a design for a task: the confounding factors in a subsequent production environment majorise the variables studied in a typical test setup.

Performing valid user studies well is a craft in itself. The human- computer interaction field has a lively discussion on topics related to validation of system-building hypotheses. There are several examples of excellently designed and executed studies in our field, but most of them are closely bound to one specific system or design and very seldom provide sustainable results which could be reused for other system design processes, and general case user studies comparatively seldom address information access specifically. The gap between engineering and application-oriented research in information retrieval and the craft of designing and building appealing and habitable interaction models based on a general understanding of people going about their business in their everyday activities needs to be closed somehow.

This paper argues for the practice of formulating *use cases* for this purpose. Use cases may be put together with various levels of ambition, competence, and insight, but once formulated, interaction specialists can debate and test the validity of the use case; information system specialists can set parameters for system benchmarking, based on crucial characteristics of the use case; and industrial and commercial stakeholders such as broadcasters, media editors and archivists, consumer electronic device manufacturers and personal information management researchers, product designers, or even net activists can build and design their systems according to results given by the use case, if they find it conforms to the behaviour they can observe in their customers and clients.

2. USE CASES EXPLAINED, BRIEFLY

A *use case* is a relatively informal or semi-formal description of a system's behaviour and usage which is intended to capture all the functional requirements of a system by describing the interactions between the outside actors and the system to reach the goal of the *primary actor* [11, 12, 8, 17]. In other words, in a use case the system, the actors, the goal of the primary actor, and the sequence of actions between the system and the actors are defined to capture and organize the functional requirements of the system. Use cases are tools for developing systems, and user actions as formalised in the use case — most often using UML, the Unified Modeling Language — are mapped onto system components and system development objects for the purposes of system development and evaluation.

Scenarios are not use cases but instances of them: often several scenarios are necessary to track the various paths through a given use case for a system. A use case is typically organized around a *main success scenario* which records the simplest path through the use case, the one in which everything goes right and the goal is reached without difficulty. [8]

3. THE IMPLICT USE CASE OF INFORMATION RETRIEVAL

Cranfield-style studies performed so far have not been agnostic with respect to use cases. While the notion of a use case has not been explored to any great extent in information access research[1], there are implicit assumptions made concerning the users and their goals in the notion of retrieval being a topical and task- based activity for focused, active, and well-spoken users. This implicit use case has informed both the evaluation and design of information retrieval systems. In a typical information retrieval benchmark, the main success scenario would consist of the primary actor submitting a query to the system and the system returning a ranked list of documents to the user. No alternative scenarios are usually considered and the rest of the interaction between the system and the user is left outside of the scope of the benchmark.

Essentially, the Cranfield studies can work well to establish the usefulness of systems with respect to some human activities if the activities in question fit this implicit use case. If the activities do not fit, evaluations will fail to establish success criteria. When information access technology moves from its current prototypical domain of topical text retrieval, the implicit information retrieval use case becomes less useful as a backbone for evaluation. The advent of multimedia as a large information carrier may be the most obvious example, as multimedia is different, used differently, by different users, and for different reasons than text. Therefore, benchmarking must change to capture the most important criteria for success for a variety of multimedia information access systems, using e.g. appeal, confidence, and satisfaction to complement completeness and precision as target notions.

Aside from the intrinsically problematic nature of the implicit assumptions, a yet bigger issue will be the practical ramifications for systems engineering from leaving the use case implicit. If the purpose, the usage and the users of the systems that are being evaluated are not explicitly and clearly defined it makes the validation of the starting points of such evaluations difficult and hinders the meaningful comparison of the evaluation results. This is where use cases show promise of being a useful tool for evaluation of future generations of information access systems. They can be a practical tool to bridge the divide between benchmarking and validation and they can guide the design of benchmarking efforts by requiring the evaluation design to make explicit the intended usage of the evaluated system, and how it provides value for its users.

4. VARIATION ACROSS USE CASES IN INFORMATION ACCESS

Defining the central features of the primary actors, their goals and the surroundings where the system under discussion is going to be used will help understanding – from the user point of view – what the important success criteria for information access system are. It will also make it easier to compare the different evaluation tasks and test collections to each other, provided that the research community will be able to agree on a set of features that should be discussed. During the course of the European CHORUS coordination action a number of salient features of use case were collected and presented in the CHORUS project deliverables and the

[1]The term "use case" is frequently used in papers on information access technology, but usually it is used to refer to informal descriptions of how useful a certain system component might be.

FACTOR	TYPICAL VALUES	RELATION TO EVALUATION
User		
user group	single user; collaborative (synchronous/asynchronous; established/adhoc group)	
expertise in domain or system usage	novice vs occasional vs expert	result presentation and learning curve
System		
network and platform	home / office / mobile/ ...	network latency; size of result
Source		
media	text, audio, video, images, graphs, 3-D objects, maps, diagrams, data collections	gold standard set-up
business model	subscription, pay-per-view, no cost	cost calculation
repository	size, ownership, quality, provenance	browseability; quality and trust
permanence	collection vs stream	reproducibility
Session		
query	specification, example, set, implicit	formulation effort
initiative	push vs pull; lean-forward vs lean-backward	optimisation vs satisficing
context	none, implicit, user-specified, individual user model, stereotypical user model	fit over time to user model
goal	known-item search, overview, question answering, entertainment, socialisation, information refinement, monitoring	target notion: relevance / satisfaction / confidence
timeliness	real-time vs offline process	response time
persistence	single-shot, durational, repetitive	learning curve
result	single item, list (exhaustive or selection; ranked, ordered, organised), summary (report, overview, visualisation), answer (extraction, db fill), notification, browsing interface	recall-precision trade-off

Table 1: Some Use Case Dimensions of Variation

CHORUS survey [6, 1, 15]. Some of the features can be found in the Table 1.

Recent strands in the study of interactive retrieval have begun to move beyond the modelling of sessions as simple retrieval of items from a collection, emphasizing the importance of modelling context beyond the query itself in understanding the goals of the user. The dimensions of variation are familiar to the field — there is a large body of literature on the character and variation of interacting with information retrieval systems, both theoretical and based on lab and in-situ studies, starting from the 1970's at least (e.g. [3, 4, 16, 2, 5]) and continuously addressed today in the digital library field: this paper will not attempt a survey beyond those recently published (e.g. [14]). Table 1 is to indicate how some of those variational dimensions can be used to guide the description of a use case enhanced with explicit hypotheses about user goals and behaviour to inform validation activities in system evaluation.

4.1 User factors

Factors directly related to the user or users have obvious implications for evaluation. Two examples here will suffice: firstly, recent studies in collaborative information retrieval [9] show how collectives of collaborating users break some of the patterns of single-user interaction with an information system. Evaluation of results cannot necessarily be done using metrics for invididual retrieval. Secondly, the expertise of a user in domain or in the search system has immediate effect on evaluation: if a system is intended for professional users, a lab study with one session will not evaluate the long-term suitability of the solution in a professional setting and a probe study may be more appropriate and the system behaviour must be measured over a longer time depth or over a session rather than over a single search request. [10]

4.2 System factors

Factors related to the technology used for interaction with the system, both as regards interaction device as well as the infrastructure for information transport will influence the presentation, the flow and the optimal configuration of information delivery. For instance, the size of screen and the convenience of input from the user – e.g. keyboard or voice input – will influence what result sets are likely to be most acceptable to the users.

4.3 Source factors

In interaction with information sources different from the prototypical text document collection a number of central factors of user satisfaction and thus evaluation change. If the interaction is with inherently streamed data, a database of retrospective material will become unrealistic and the current requirement of benchmarking to be reproducible on the same data set (such as it is formulated e.g. in the call for papers to this very conference) counterproductive. A more suitable requirement could conceivably be to require the benchmark results to be stable and predictable given some sampling procedure on the data stream. Additionally, if the source repositories are commercial and require users to pay for access to each item, the evaluation must incorporate a cost factor.

User characteristics	Session characteristics	Hypothetical goal
Adhoc web search		
occasional user, individual, active, well-spoken	explicit brief query specification, pull, rapid turnaround, no feedback	satisfy topical fact-based information need
Personal TV		
occasional user, passive social context?	implicit query, recommendation push, lean-back	while away time, minimise intellectual effort gain social relevance, entertainment
Anomaly monitoring – coast guard		
professional, cooperative?	explicit query (selected from a set of predefined queries) lean-back	full overview of navigational patterns, monitoring

Table 2: Some Example Use Cases with Scenarios

5. EXAMPLE USE CASES IN INFORMATION ACCESS

In Table 2 some very simple formulations of use cases are given with annotations as to salient characteristics and questions for evaluation. The user actions are annotated with tentative user goals. This formulation of use cases is intended to show that if explicit hypotheses about user preferences are given, these hypotheses can guide evaluation both in choice of benchmarking metric and in making validation goal-directed. The choice of which hypotheses to work on is naturally a question which this model leaves up to the system engineering team – if the team is happy with the hypotheses they can be left as stated for others to address at some later point.

For each use case each gray-box action of the system needs to be evaluated with respect to user goals. Those goals need to be validated; the components benchmarked.

5.1 Topical web search

- User formulates and enters a number of search terms *"This represents my information need appropriately"*
- System searches index for matching documents and presents ranked list
- User peruses list and selects relevant documents for reading
  "These documents will fulfil my need with an approriate reading effort / enjoyment ratio".

The ranked list can be benchmarked with respect to match to topical search terms. The query entry interface can be evaluated using human factors methodology. The hypotheses that users are able to formulate appropriate queries needs to be validated. The hypothesis that users are able to find the most appropriate documents in a ranked list also needs validation. The second validation will impact the benchmarking metrics for the ranked list. (cf e.g. [13]). The use case is illustrated in Table 3 giving the sequence of actions for the main success scenario annotated by the hypotheses related to them.

5.2 Lean-back entertainment — watching TV from a couch

- User relaxes *"I do not wish to expend any effort."*
- User activates system *"What is on?" "What is my peer group viewing?"*
- System presents programming after inspecting stored user profile and current viewing pattern of peers.
- User, by minimal short coding of preference, accepts or discards offering. If user discards offering another program is presented. If user accepts user profile is updated and the information is transmitted throughout the peer group.

The presentation of programs with respect to user profile can be benchmarked in comparison with other systems. The hypothesis that users prefer not to expend any or minimal effort in choosing programming needs to be validated since it will impact the design of the initial interaction with the system and the interaction at the confirmation or rejection point; the hypothesis that they care about the viewing patterns of their peer group needs to be validated since it will impact the benchmarking of the information access component; the hypothesis that they do not mind the television entertainment system storing and sharing their viewing patterns needs to be validated since it will impact the quality of information the system has at its disposal for retrieving programming; and finally the hypothesis that dwell time is a reasonable measure for user satisfaction[2] needs to be validated for the iterative search and the user profile update functionality to kick in appropriately.

5.3 Anomaly monitoring

Coast guard

- User activates monitor and specifies geographical area and shipping lanes to monitor. *"I need to watch the lanes for accidents."*
- System polls sensors in area.
- User waits. *"I do not want to engage with system."*
- System alerts if movement of observed vessel is anomalous and predictive of further anomaly.

The hypothesis that a user does not want to engage with the system needs to be validated, as does the hypothesis that a user is able to select among the offerings of the system. Other factor such as pricing may affect the coverage of the system described. It is easy to think of other anomaly monitoring use cases, such as a consumer market analyst monitoring user reactions to some trademarks or a sports team

[2]For instance, to investigate if the user might be doing *"zzz"* rather than *"mmm"*.

fan monitoring gossip and events concerning her favourite team. Whether these scenarios and the user goals are similar enough to warrant a shared evaluative framework can only be established if the use cases are specified further and the hypotheses and underlying assumptions are validated. If this is done, the results from one study (since it is unlikely one single research or development effort would result in a system to serve all scenarios above) can be used for other efforts as well, in toto or piecewise.

6. INDUSTRIAL RELEVANCE

One of the success criteria for an evaluation of an information access solution is the ability to predict subsequent take-up of the solution in practice. If the evaluation centers on benchmarking of a component which does not affect end user satisfaction, the evaluation may be scientifically interesting but have little practical value. The connection between benchmarking and take- up may be confounded by a large number of variables which may be difficult to model and the final quality of the complete system may hinge crucially on something completely different than the variables measured by benchmarking of its components. There is no reason to settle for anything but the best components, but if their effect cannot be measured in practice, it will be difficult to convince a commercial system designer to invest any effort in the improvements. Here, a validated use case with clear and explicit hypotheses of usage goals and linked to evaluation benchmarks will be a much more convincing argument than a benchmark alone

7. CONCLUSIONS AND QUESTIONS TO DISCUSS AT THE PRESENT WORKSHOP

This paper has posed a number of arguments.

Firstly, it would be desirable to see a more informed usage of the use case-related methodology in the field of information access. Currently, often when "use case" is mentioned, it is used to describe a vaguely stated area of potential application or a usage scenario for a technology rather than a use case in the technical sense.

Secondly, this paper argues that the large body of work on interactive information access could be harnessed to more productive and sustainable use if its hypotheses were brought to explicitly bear upon the design of information systems. This will be necessary when system use transcends the implicit use case of topical retrieval which has been predominant in research so far. This is a question for further discussion at the workshop.

Thirdly, this paper argues that use cases for information access can be written to give explicit pointers towards benchmarking mechanisms. How this can be done with a minimum of unnecessary effort is a question for discussion at the workshop.

Fourthly, this paper argues that use cases for information access can be written to make hypotheses about user preferences, goals, expectations, and satisfaction explicit. This will enable other researchers from other fields, notably user studies experts, to study those hypotheses, validating or disproving them, without incurring the expense and effort of studying information access usage from first principles. This will provide our field a framework for bridging validation and benchmarking. This will enable the information access field to recruit interested parties from neighbouring areas where user studies are the prime focus and interest. It will make our research results sustainable and portable from one research effort to another, and it will make our field more relevant to practicioners who will be able to take our results as given for their system design purposes or as indicators of which system solutions they will be able to deploy in their application area for their customers. That increased level of relevance for practical system design is crucial if the field of information access research is to remain relevant to information access system purveyours, as it has for the past fifty years.

Questions left unanswered is how to publish and prioritize between the necessarily more numerous evaluation tasks — both benchmarking and validation tasks — which a use case formulation will give rise to, how to make the linkage between a user action as given in a sequence model and an evaluation excercise appropriately deterministic, and how to generalise from a number of use case formulations to find common evaluation frameworks in the most profitable manner.

	User		System	
	Action	Hypothesis	Action	Hypothesis
Main success scenario: iterate until user is satisfied or gives up	Issues a query.	User is able to verbalise need		
			Searches index, presents a ranked list of items	
	Peruses ranked list, selects relevant items	User is able to identify relevant docs		
	Reformulates query.	User is able to identify direction of improvement		

Table 3: Sequence model example

Acknowledgments

This work was supported by the European Commission under the Promise network of excellence and draws results from the 2006-2009 Chorus coordination action.

8. REFERENCES

[1] R. Bardeli, N. Boujemaa, R. Compañó, C. Dosch, J. Geurts, H. Gouraud, A. Joly, J. Karlgren, P. King, J. Köhler, Y. Kompatsiaris, J.-Y. LeMoine, R. Ortgies, J.-C. Point, B. Rotenberg, Å. Rudström, O. Schreer, N. Sebe, and C. Snoek. CHORUS deliverable 2.2: Second report - identification of multi-disciplinary key issues for gap analysis toward EU multimedia search engines roadmap. November 2008.

[2] N. J. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications*, 9(3), 1995.

[3] J. L. Bennett. Interactive bibliographic search as a challenge to interface design. In D. E. Walker, editor, *Interactive bibliographic search: The User/Computer Interface*, pages 1–16. AFIPS, Montvale, New Jersey, 1971.

[4] J. L. Bennett. The user interface in interactive systems. *ARIST*, 7:159–196, 1972.

[5] P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3), 1997.

[6] N. Boujemaa, R. Compañó, C. Dosch, J. Geurst, Y. Kompatsiaris, J. Karlgren, P. King, J. Köhler, J.-Y. LeMoine, R. Ortgies, J.-C. Point, B. Rotenberg, Å. Rudström, and N. Sebe. CHORUS deliverable 2.1: State of the art on multimedia search engines. November 2007.

[7] C. W. Cleverdon and M. Keen. Cranfield CERES: Aslib Cranfield research project - Factors determining the performance of indexing systems. Technical report, 1966.

[8] A. Cockburn. *Agile software development*. Addison-Wesley, 2002.

[9] P. Hansen and K. Järvelin. Collaborative information retrieval in an information-intensive domain. *Information Processing and Management*, 41(5):1101–1119, October 2005.

[10] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, first edition, 2005.

[11] I. Jacobson. Object-oriented development in an industrial environment. *Proceedings of OOPSLA '87: Sigplan Notices*, 22(12):183–191, 1987.

[12] I. Jacobson, M. Christerson, P. Jonsson, and G. Övergaard. *Object-Oriented Software Engineering: A Use Case Driven Approach*. Addison-Wesley, 1992.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October 2002.

[14] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2), 2009.

[15] P. King and Y. Kompatsiaris. Towards a use case ontology for multimedia information retrieval. 2008.

[16] T. Saracevic. Evaluation of evaluation in information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 138–146, New York, NY, USA, 1995. ACM.

[17] G. Övergaard and K. Palmqvist. *Use cases - Patterns and blueprints*. Addison-Wesley, 2004.