Adopting Systematic Evaluation Benchmarks in Operational Settings

Jussi Karlgren

Abstract Evaluation of information systems in commercial and industrial settings differs from academic evaluation of methodology in important ways. Those differences have to do with differing organisational priorities between practice and research. Some of those priorities can be adjusted, others must be taken into account, to be able to include evaluation into an operational development pipeline.

1 Evaluation in an Operational Setting Differs from an Academic Setting

Some of the differences between operational and academic settings are obvious, some less so. (”Industrial” or ”operational” will here be understood to include all kinds of applied uses of information systems, including non-commercial and public contexts of use).

Firstly, an information access service is seldom the primary objective of an industrial project. The industrial project is built to be used for some concrete purpose and information access is a component, frequently an important one, in some process to contribute to that purpose. The ultimate objective of the information access system is to be a sustainable component in that process, for the length of time that process contributes interestingly to the overall goals of that project, be it to generate revenue or goodwill or general happiness.

Secondly, the objective for an industrial project is to perform some task adequately. There is rarely need for optimising performance beyond what is necessary to satisfy the requirements posed on a system. This is in contrast with academic projects, where the goal is to improve and optimise some method, some algorithm, or some performance for some fixed and well specified task. Such improvement and optimisation may not be in the interest of an operational service, in face of limited

Jussi Karlgren
Gavagai and KTH, Royal Institute of Technology, Stockholm

resources: funds, competent personnel, or attention, all of which are scarce in most industrial contexts.

These two differences have an impact on evaluation methods.[1] What stands in the way of systematic and continuous formal evaluation of information system quality in industry is that evaluation in academic projects focusses on less complex, idealised tasks than what industrial applications or technology can accommodate and evaluation metrics and methods from academic research projects typically reduce an information need challenge into something very clear-cut and clean. Thus, the evaluation schemes proposed in laboratories frequently appear to be irrelevant to understanding the quality of the operational service being offered to customers or end users.

Simplicity and crispness do not reflect the reality of deployed systems in practical use: systems may have many instances, sometimes non-identical; usage may be distributed across numerous nodes; the data under consideration may vary; and the users may have very different objectives than is assumed in an evaluation scheme. Operational data can be messy, incomplete, and distributed over numerous systems, where academic test collections have been cleaned, simplified, and organised to the point that they no longer adequately represent the complexity and variability of the operational realities (Imhof and Braschler, 2015). One key factor in making evaluation schemes relevant is to acknowledge the simplification from industrially relevant task to testable output from a system. How then are operational tasks different from those used as models for benchmarking evaluation?

1. The information need may be complex and involve combinations of information items, which makes search technology but one component in a larger whole: “Is this political question worth taking a stand on?” “What factors appear to worry potential customers for our product at what stage in their purchase path?” “What factors in the pension system cause most confusion for our senior citizens?” “Does this group of people pose a risk for public safety?” “Will it be easy or difficult to recruit college graduates to this business area next Fall?”
2. Establishing whether a need is fulfilled or not may be more challenging than in a topical retrieval experiment. The analysis may involve several steps beyond the retrieval or identification of candidate items, and the relevance of such items may be impossible to assess at search time. The determination of what is important, relevant, valuable, or not may be made by someone other than the person who formulates the information need. Sometimes no result is the most positive result, but a *no items found* result page is unsatisfying and not what most analysts hope for. “What published work might be relevant to assessing the novelty of this potential patent application?” “Did our customers notice that we mis-labeled the content of our product and corrected it and if they do, do they care?”

[1] This point has been made in several recent projects such as CHORUS, TrebleCLEF, or PROMISE, where industrial and research interests have met in think tanks and workshops to share experiences (Braschler, 2009; Braschler et al, 2012, e.g.) This insight is also integral to several CLEF evaluation workshops.

3. The real world data and process may be complex and dynamic compared to the analysis of documents from a relatively static benchmarking database. A test set built on a static model may not generalise well. “Do items posted on that video streaming site infringe on our copyright?” “Is the pricing of this tradeable asset moving in some direction?” “How should we set the initial odds for this bet in our book?” “Will the data from our newly acquired division merge well with what we have been working on before?”
4. The presentation, packaging, and delivery may be complex; the fulfilled information need may not be operational or actionable: even a well executed retrieval or filtering task may not actually deliver what is useful for the organisation. In most organisations, providing more information for decision making means more work, not less, and this may cause some consternation for decision makers at the receiving end. In general, queries such as “What are some of the more interesting trends in our market area that are likely to influence our sales five years down the road?” will provide more useful data than “How many mentions did our brand get in social media and in what sentiment were they expressed?” and in a streaming Big Data access scenario, the individual data points are less interesting than patterns in their flow and changes in those patterns.
5. In real life tasks, human system users are adaptable and have great readiness to accommodate even to clumsy systems in order to accomplish or further their goals. Applications built from overly simplistic assumptions about user needs may still be functional as tools, and they influence the usage and inform the expectations of users. The cost of introducing new tools, retraining personnel and readjusting processing pipelines may be considerably more complex than coping with noisy or otherwise substandard output from an information system.

2 Openness and Accessibility

While academically accepted testing may be attractive for marketing reasons to achieve authority or status, organisations may be skittish to make test results public or use public test sets for reasons related to contractual obligations, commercial risks (real or perceived), or user privacy. If tests are performed in-house, the interpretation of test results may be difficult: if management poses unrealistic goals, which is not unheard of, those in the organisation who are responsible for engineering efforts may be unwilling to provide quantitative data to avoid argumentation and thus unwilling to openly evaluate systems for which they may have less responsibility. And crucially, many academic test sets, if relevant and interesting, are only available to non-profit or research organisations. *A challenge for those who define evaluation schemes and procedures is to make them available for all, and to allow for testing without publication of results.*

3 Reliability vs Validity

A method—an algorithm, a computational approach, a memory model etc—may be interesting for research purposes: it may provide insights into human information processing, it may demonstrate interesting characteristics of a collection or the items in it, or it may at some time in the future be the basis for other methods of interest. That method may even score well on various quantitative tests, improving results given by previous approaches. That same method may still be completely uninteresting for practical purposes. A test, however formalised and solid, however robust in its ranking of various experimental conditions, does not guarantee usefulness.

This distinction between *reliability* and *validity* has a long history in the behavioural sciences. Evaluation of information access has for many years been systematic and quantitative, using well-established and commonly accepted benchmarks to compare approaches and methods. These benchmarks, however well normalised and graded, do not guarantee validity of the test. The validity on the test hinges crucially on the task it is patterned to emulate. If the evaluation concerns some behaviour of some component which at the end of the system pipeline makes little or no difference for satisfying the requirements of users, it will have little validity. By contrast, if we want evaluation efforts to predict subsequent take-up of some solution in practice, the evaluation scheme and the metrics it offers need to have high *validity*.

The link between benchmarking a component and assessing its eventual effect on user satisfaction and thus potential for industrial take-up is confounded by a large number of variables, some of which are very challenging to model with any level of confidence in evaluation efforts. If no such linkage can be demonstrated, it is unlikely that the results of an evaluation scheme will convince an industrial system designer to pay attention to that specific evaluation result.

This is where some representation which demonstrates the connection between a system component and its performance on the one hand and user satisfaction on the other will come in handy. In discussions at Conference and Labs of the Evaluation Forum (CLEF), and other related conferences and workshops *use cases* have been proposed as one such potential representation. A use case is a relatively informal or semi-formal description of a system's behaviour and usage intended to capture its functional requirements by describing the interactions between outside agents and the system. Everything should be described in terms with which primary users reach their goals and the description should be useful for system development and evaluation purposes. The objective of using use cases is to make such descriptions simple, lightweight, and incrementally amendable[2].

Use cases for information access evaluation can be written to make hypotheses about user preferences, goals, expectations, and satisfaction explicit. Use cases may

[2] A use case is *not* a set of scenarios, nor need it be a formal UML schema. Currently, the term use case is often used to mean a vaguely stated area of potential application or a usage scenario for a technology. A use case should be more specific to be useful for system development, and in this case, evaluation. (Jacobson, 1993)

be put together with various levels of ambition, competence, and insight. There is no need to aim for perfection, but once formulated, they will enable practitioners and system architects to examine those hypotheses and to assess if an evaluation scheme is relevant to what they are putting effort into and whether it conforms to the behaviour they can observe in their customers and clients. *Use cases (or some similar semi-formal approach) can be used to bridge the gap between benchmarking and validation.*

4 The Implicit Use Case of Benchmarking

It is worth noting that the lack of an explicit use case does not mean that there is no use case in the background. The Cranfield paradigm (Cleverdon et al, 1966) compares the capability of information retrieval algorithms to identify and rank topically relevant documents given a well-defined information need under controlled test settings. This, together with appropriate gold standards and scoring practices, has given the information retrieval development efforts a level playing field of immense usefulness. The entire point of that test framework is to 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. These interaction-related factors will oftentimes be the most important determinants for the user experience of a system, especially if the information retrieval system is but a component in a larger service. *To catch the attention of industrial parties and to ensure validity of their metrics, academic experiments must formulate use cases which capture aspects of interest in deployed tasks.*

5 Organisational Thresholds for Introducing Systematic Evaluation in Industrial Projects

The above factors—use case discrepancy, complexity vs measurability, satisficing vs optimisation, lack of resources—all contribute to lack of interest for systematic and routine evaluation of information systems in practical settings, even where it would be motivated. They all contribute to organisational thresholds, which have repeatedly been brought to the fore at discussions in workshops and panels on evaluation in industry (Forner et al, 2013; Kazai et al, 2016; Kanoulas and Karlgren, 2017).

Enterprises often lack the resources, above all in terms of engineering personnel, to develop evaluation practice and to keep track of best practice in evaluation research. New graduates who may have performed rigorous evaluation in educational and graduation projects have small possibilities to change existing routines and prac-

tices in the organisation they are recruited to work in. Retaining and encouraging the experimental and daring technology culture from the educational background of new entrants is a challenge for any development-oriented organisation, but can if well formulated, have the beneficial side effect to be a persuasive recruitment strategy.

Commercial or related practical realities do not prioritise quality metrics of the type discussed in this volume. Enterprise needs are different from the most generalised needs of the implicit benchmarking use case (Kruschwitz et al, 2017, e.g.). Customers or other end users make multi-factor decisions based on technical and administrative fit to other existing systems and on a multitude of technical factors such as platform independence, scalability, consistency, coverage, and reliability of service, where content quality of output is only one of several features of interest. At the time when a major introduction decision is made, it is likely to be of high priority, but monitoring it continuously fades to the background as the system is installed and deployed. Feedback from end users is handled by customer service and sales staff who have a different focus than engineering staff would. Concrete bug reports will be sent from support staff or sales staff to engineering staff, but more general views of quality of service are routinely covered through workarounds, customer training, or new product releases, the effect of which are more notable for the customer than search component quality. Organisational gaps between customer opinion and engineering staff makes quality monitoring less organisationally useful: using the customer feedback pipeline to motivate continuous quality improvement, not only assurance, will add urgency to quality testing and evaluation. This means turning observations from evaluation metrics into development tickets with concrete goals for improvement of output. *A challenge for industrial and other applied organisations is to encourage a culture of continuous improvement in their technology departments and to provide an information pipeline to support it.*

The focus of a system in production is on its entire output. This is in the end evaluated through sales and customer satisfaction, metrics which have the attention of executive management of an organisation. Component-wise evaluation is done by engineering departments, through systematic testing, most notably through unit testing. Unit testing, the systematic and routine quality testing of components which are subject to development and change, is most often binary in nature: a module passes or fails a test. Quality testing of information retrieval components, by contrast, will yield a score ranging somewhere in the middle between complete failure and perfect ideal performance. The output of such tests is less obviously actionable: an evaluation score from a retrieval test typically does not generate a bug report but may instead invite tuning or improvement efforts. How much effect such an effort has on the bottom line of the organisation can be difficult to assess, and there are no obvious cut-off thresholds that can be set at the outset to categorise scores into failure vs success. *Industrial sites will need help from academic practitioners to interpret evaluation scores, related to best practice, rather than optimisation.*

In many operational contexts the number of testable components can be prohibitively large. If the engineering effort of a corporation or public office ranges over dozens of different systems many of which have proprietary information ac-

cess components, some of which are internal to the system, some outward facing, their testing cannot easily be coerced into the same framework. Engineering in a large organisation can be driven by innovation and development efforts as well as maintenance and upkeep. The former efforts involve feasibility decisions and extensive testing; the latter, frequently, assume technology to be stable. This may not always be true, especially in face of changing influx of data and scalability concerns: if the original metrics to motivate a decision have been lost or discarded along the line, reintroducing them will be a challenge and involve a serious amount of work and effort. Evaluation needs to be viewed as part of system monitoring, not solely as a decision making criterion. *Preserving evaluation metrics from development processes and keeping them in place during the operation life cycle phase of a system saves effort.*

6 How to Make Evaluation Practice Relevant for Industry

The main lessons to be learnt from examining the gap between academic and operational evaluation are that to make the former more relevant and the latter more systematic and actionable, the operational priorities of a system development process need to be taken into account and adjusted where necessary.

- Evaluation schemes and procedures must be conveniently available and integrable to allow for testing without publication of results.
- Evaluation target notions, methods, procedures, and metrics must have validity with respect to tasks.
- Validity can be achieved through e.g. formulation of use cases which capture aspects of interest in deployed tasks.
- Evaluation schemes must be sensitive to the distinction between optimisation and best practice.
- Many evaluation schemes, while useful benchmarks for academic research, will not be useful for industrial sites.
- Industrial sites will need help from academic practitioners to interpret evaluation scores.
- Industrial organisations must recognise that development and deployment decisions feed into the entire life cycle of a system.
- Industrial organisations must encourage a culture of continuous improvement.
- Industrial organisations must provide an information pipeline and procedures to support such a culture.

References

Braschler M (2009) Best practices in system-oriented aspects for multilingual information access applications. In: Proceedings of the eChallenges 2009 Conference

Braschler M, Rietberger S, Imhof M, Järvelin A, Hansen P, Lupu M, Gäde M, Berendsen R, de Herrera AGS (2012) Best Practices Report, Deliverable 2.3. PROMISE project

Cleverdon CW, Mills J, Keen M (1966) Aslib Cranfield research project—Factors determining the performance of indexing systems. Tech. rep.

Forner P, Bentivogli L, Braschler M, Choukri K, Ferro N, Hanbury A, Karlgren J, Müller H (2013) PROMISE technology transfer day: spreading the word on information access evaluation at an industrial event. SIGIR Forum (1):53–58

Imhof M, Braschler M (2015) Are Test Collections "Real"? Mirroring Real-World Complexity in IR Test Collections. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixth International Conference of the CLEF Association (CLEF 2015), Lecture Notes in Computer Science (LNCS) 9283, Springer, Heidelberg, Germany, pp 241–247

Jacobson I (1993) Object-oriented software engineering: a use case driven approach. Pearson Education India

Kanoulas E, Karlgren J (2017) Practical issues in information access system evaluation. SIGIR Forum (1):67–72

Kazai G, Ingersoll G, Lin J (2016) "Evaluation is for conference papers. I need to build a real life product!". SIGIR 2016 Industry Track Panel, Pisa, Italy

Kruschwitz U, Hull C, et al (2017) Searching the enterprise. Foundations and Trends in Information Retrieval 11(1):1–142