Principles for Robust Evaluation Infrastructure

Justin Zobel[*], William Webber[*], Mark Sanderson[**], and Alistair Moffat[*]

[*]Department of Computer Science & Software Engineering, The University of Melbourne, Australia,
jzobel,wwebber,ammoffat@unimelb.edu.au

[**]School of Computer Science & Information Technology, RMIT University, Australia,
mark.sanderson@rmit.edu.au

ABSTRACT

The standard "Cranfield" approach to the evaluation of information retrieval systems has been used and refined for nearly fifty years, and has been a key element in the development of large-scale retrieval systems. The resources created by such systematic evaluations have enabled thorough retrospective investigation of the strengths and limitations of particular variants of this evaluation approach; over the last few years, such investigation has for example led to identification of serious flaws in some experiments. Knowledge of these flaws can prevent their perpetuation into future work and informs the design of new experiments and infrastructures. In this position statement we briefly review some aspects of evaluation and, based on our research and observations over the last decade, outline some principles on which we believe new infrastructure should rest.

Overview

Test collections have been a driver of information retrieval (IR) research for half a century. Since the effort of creating a collection often greatly exceeds that of running an experiment, the availability of test collections has allowed researchers to contribute ideas and measure their effectiveness, even if they lack the resources to construct such collections themselves. The collections have also allowed the thorough comparison and verification of IR experimental results, since the use of common metrics and data have meant that researchers can readily compare their work, and undertake retrospective studies.

Prior to the first TREC event in 1992, existing test collections had common "Cranfield" characteristics: they consisted of data, queries (or information needs), and relevance judgements. By current standards those early collections were small, but they were thoroughly curated; generally had exhaustive relevance judgments; and, in size (often less than a megabyte), were at the limits of the distribution and storage mechanisms of their era.

With TREC, a new element was introduced, that of scale. In one leap, the number of documents involved in test collections, and the volume of data that needed to be manipulated, grew by a factor of a hundred or more. Exhaustive judgements were impractical, and pooling became the mechanism used to identify candidate relevant documents. The use of blind evaluation was another new element, as *runs* of search results were created prior to relevance judgements being undertaken.

Archives of these runs were constructed as a side effect of these TREC-era mechanisms, and are of significant ongoing value. The archived runs provide an invaluable resource on which to investigate questions of "how best to evaluate systems", including ones that involve "what if" scenarios. The TREC (and similar) run archives have underpinned much of our own research on system measurement and allow, for example, systems to be re-evaluated in the light of new measurement techniques.

The effort invested in TREC was also a spur to other developments. One was the appearance of public domain systems capable of handling large volumes of data (including MG, Lemur, Indri, Zettair, and Terrier); another was the consolidation of a large number of incomparable measurement techniques into a small number of measures, including average precision (AP) and then normalized discounted cumulative gain (nDCG).

It is clear that shared testing environments (such as TREC's test collections), as well as resources such as shared, public-domain IR systems, are critical to research in this field. It is our view that other elements are also critical; in particular, we need:

- Environments for publishing new data, runs, and systems;
- Shared, statistically based tools for measuring and recording experimental outcomes
- Social frameworks that make openness the norm; and
- Provision of mechanisms by which restricted or private data can be evaluated, accessed, or inspected.

On those rare occasions where experiments must be conducted without any form of sharing, guidelines on how such experiments are

best reported should be developed.

Accurate and robust measurement is essential to the progress of science. We need good instruments in order to be able to determine whether our systems are improving and whether and which small innovations are of benefit. And we need mechanisms that provide some level of reproducibility of research results; potentially, we need to encourage publication of attempts to reproduce results. Investment in appropriate infrastructure, and in the ongoing refinement and critique of evaluation methodologies, is of benefit to the whole community.

We now outline some of the principles on which we believe infrastructure should be based.

Experimental design should be based on statistical principles

Such an observation may seem obvious, but many of the experiments reported in IR papers are devoid of any critical statistical insight. Moreover, we have found that the appropriate use of statistical methods not previously explored in the context of IR can strengthen experimental results.

For example, we can with hindsight conclude that the query sets of the TREC corpora are too small. Statistical power analysis shows that only relatively large effects can be reliably observed over fifty queries. While there are good economic reasons for query sets to be so small (for example, because a small set allows the depth of assessment needed for reusable sets of relevance judgements), their size may have hindered research progress, because researchers may have been unable to demonstrate significance for small, but nevertheless consistent, improvements. Similarly, search methodologies that improve only subsets of queries cannot be effectively tested with such sets.

This statement may appear to be in contradiction to the results in well-known papers that have determined the sufficient number of topics to have in a test collection: Buckley and Voorhees suggested a minimum of 25,[1] while Carterette et al. in the context of using shallow pooling concluded that around 200 were enough.[2] However, the major search engine companies have test collections with several thousand queries and we assume that they do so for sound statistical reasons. Using power analysis, a standard statistical method, we have found that a hundred queries is almost certainly insufficient to detect a small but reliable improvement.[3]

There has been insufficient recognition of sources of experimental error and how they can be managed or quantified. These include, for example, unreliable and incomplete judgements, biases in query selection, inconsistency due to system-to-system variation, and inappropriate aggregation of results across queries. Some of these factors remain neglected, and yet are clearly factors that affect the veracity of any final conclusions to be drawn in a comparative evaluation. More broadly, the sources of kinds of error vary between collections, leading to unknown levels of uncertainty.

It is critical that standard measures be used. The plethora of metrics available for reporting has led to researchers presenting multiple tables of systems scores in published work, as if these different measures reflect orthogonal concepts. In fact, most effectiveness measures correlate strongly with each other, and what is a good system by one metric is usually also a good system when measured by another. So, for example, reporting precision at depths ten and twenty, and average precision at depth one hundred, does not contribute three times as much evidence to a claim of system superiority as does reporting average precision alone. In fact, it probably contributes little additional information at all.

Indeed, one could argue that system designers only have one fundamental goal – namely, to populate the initial part of the ranking with as great a population of proposed-to-be-relevant answers as possible, spread across the spectrum of possible query interpretations – and hence that only one fundamental measure should be required. What that measure should be is, of course, then open to debate; but this debate is distinct from the requirement on a system designer to defend their claims of "improved performance", and can be carried out independently of particular systems and particular test collections.

Effective reproducibility requires public data and open systems

A key tenet of the scientific method is that of disclosure – for centuries researchers have been expected to explain *how* they did an experiment, as well as what the measured *result* of it was, so that claimed outcomes could be verified by others. One only has to consider the cold fusion debacle in 1989 to see why independent verification of claimed research outcomes is important. The IR research community, however, does not have the same enthusiasm for publishing repeat experimentation as the field of research that Ponds and Fleishman worked in.

In the computing disciplines, claims about result quality or *effectiveness* depend on software and data alone (whereas claims about computational effort or *efficiency* also depend on hardware). If both software and data are made available to others as an integral part of publication, or least if there is the expectation of their availability post-publication, then it should be straightforward to independently verify claims about effectiveness.

Provision of software as part of the review process is required for some journals in other fields. The journal *Bioinformatics* is one of several in computational biomedicine where code must be provided as part of the paper submission process; the ACM *Journal of Experimental Algorithmics* has a similar requirement. These rules enforce a culture of open code, and to a lesser extent data, allowing the whole community to benefit from one team's effort.

A complicating issue is that software is often developed in-house by research teams, and represents a significant component of the innovative cost of undertaking research. Once developed, software is often considered to be proprietary – certainly so when developed

[1] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability", *Proc. SIGIR*, pages 33–40, Athens, Greece, 2000 (dx.doi.org/10.1145/345508.345543).

[2] B. Carterette, V. Pavlu, E. Kanoulas, J. Aslam, and J. Allan, "Evaluation over thousands of queries", *Proc. SIGIR*, pages 651–658, Singapore, Singapore, 2008 (dx.doi.org/10.1145/1390334.1390445).

[3] W. Webber, A. Moffat, and J. Zobel, "Statistical power in retrieval experimentation", *Proc. CIKM*, pages 571–580, Napa, USA, 2008 (dx.doi.org/10.1145/1458082.1458158).

in an industry lab, and often so even when developed by academic research initiatives. That is, innovative software is both costly to create, and also embodies a competitive advantage that its authors may be reluctant to surrender too quickly. On the other hand, for publicly funded research, there is an argument that the resource was publicly funded for the public good, and does not belong to the individual researcher or team. This is the argument that has led to mandating of publication of some data arising from publicly funded research in, for example, genomics.

Another complicating issue is the quest for realistic, and ideally real, data. In IR, many types of data are desirable: there is the underlying text itself, with public web pages the most obvious example, but corporate intranet data also being of considerable interest; there are the query sessions that users submit against that text; and there are the interaction records that describe how users reacted to certain combinations of text and query, for example, via click-through logs and other such data. Effort might also be put into the task of generating relevance judgements, an evaluation of some or all of the documents with respect to some or all of the queries, to determine the extent to which they represent answers to one or more possible interpretations of that particular query.

Of particular issue in this regard is that the community or public data typically used by academic research groups tends to be of a smaller scale than is available to commercial entities, and often is less up-to-date. Customers of commercial systems must be assured of their privacy (even when that service is provided free of charge), which means that any data released in connection with user behaviour – such as logs of query sessions, or interaction logs – must either be heavily anonymized or be subject to rigorous legal control. A recent attempt to bypass those problems and collect query and interaction data directly from volunteers via a research system failed to reach a critical mass of users, and the initiative has now been discontinued.

The net position is that, on the one hand, researchers in commercial laboratories are likely to be able to carry out more precise experimentation than their university-based colleagues, but are also less likely to have their work validated independently.

We must accept that some experiments will involve private systems, private data, or both. However, if there is no map from such data and systems to material that other researchers might access, the experiment has in effect been conducted in an isolated universe, and the lessons for the shared universe are likely to be limited. Work in which all materials are kept private and description of which is minimal requires unusually strong arguments to justify publication.

The corollary, then, is that those who for whatever reason work within a private framework are under an onus to also experiment on some public resource, to allow meaningful comparison. Perhaps, for example, external data can be used; or a subset of the private data published; or comparable experiments undertaken using a public system. If the raw data needed for replication of the experiments cannot be released, then perhaps the aggregated data required for verifying the analysis can. At the very least, researchers working in such a private environment must describe their system and data in sufficient detail to allow other researchers to conduct repeat experiments on similar data sets and systems.

Going beyond these ad-hoc solutions, we need to create mechanisms under which commercial organizations are comfortable with outside use of their data. This may include trusted independent sites at which the data is maintained, and can be accessed under license; agreed ethics frameworks on use and reporting to which researchers can subscribe; agreements under which researchers can use data at the organizations free from restriction on reporting of findings; and so on.

We further note that hypotheses can only be robustly confirmed from real data; simulated data, derived from a model, only allows learning about the model. While simulated data can allow exploration of parameters and so on, ultimately some confirmation on real data is essential.

Data can be derived in a quasi-artificial way from real collections; for example, partitioning of collections was used as a basis of a long sequence of experiments in federated retrieval. However, different partitioning methods led to results of varying value: in numerous instances, early results were not substantiated on more realistic partitionings, and the later work exposed strong limitations in the artificial constructs used earlier on.

Requiring the release of data or code would be contentious. But we should at least be able to require authors to report the availability of experimental materials at submission time. Materials could be available publicly; for research use only, under whatever conditions; or not at all. Such a requirement forces researchers, public and private, to think about and plan for data release, and organizations to decide upon and commit to a materials-release policy at submission time. It would also alert researchers to their ethical responsibility to extract and maintain experimental data.

No fixed requirements for availability need be set. But a reporting system would at least make the availability, and potential for verifiability and reproducibility, explicit to reviewers and readers. Over time, it should add persistent pressure for greater transparency and availability of research materials.

Progress needs to be measurable

TREC and its sibling organizations, such as CLEF, NTCIR, and FIRE, have a dual purpose: they use the collective outputs of the international IR community to build test collections, and at the same time compare the effectiveness of the systems built by that community. Because of its mandate to keep building test collections, TREC every year produces new versions of the collections, complicating the task of cross-year comparisons. With the benefit of hindsight, it can be seen that these evaluation organizations should have included the charting of progress as part of their role for the community.

In the field of speech recognition, five years ago Deng and Huang produced an exemplary summary of progress.[4] A graph in this paper shows how the research community applied itself to a series of data sets almost every year, reducing the error rates made by speech

[4] L. Deng and X. Huang, "Challenges in adopting speech recognition", *Communications of the ACM*, 47(1), 2004, pp. 69–75 (dx.doi.org/10.1145/962081.962108).

recognition systems. After a few years' progress on a particular data set, it was discarded by the community and replaced with new data representing a more challenging problem: in the early 1990s, the focus was on carefully gathered recordings of reading; when error rates fell to a few percent, the emphasis switched to broadcast and conversational speech. The longitudinal analysis of Deng and Huang shows steady reduction in errors on different tasks. There is no equivalent summary of IR research.

Nor, indeed, does IR have an equivalent record of progress. Some data sets have remained in use for decades, with little measurable gain. We noted at the start of this statement the importance of the TREC run archives in our own work. An example of this is our experiments with standardization, which were used to estimate how much true performance gain there had been in systems over the last two decades. Without standardization, inter-year comparison is impossible, but by introducing standardization and reference systems we found (unhappily) that there was no evidence of improvement from 1994 to 2008. This post-hoc analysis would have been impossible without the archive. Also unhappily, we were able to confirm it by inspection of published results, finding that claimed 'gains' appeared to be largely due to poor choice of baselines.[5]

In addition, evaluation metrics develop over time. For instance, precision dominated the earliest TRECs, followed by AP, while nDCG is widely used today. Metric values for earlier runs cannot be calculated without a run archive, preventing direct comparison of effectiveness between new and historical runs.

It was for these reasons that we created `EvaluatIR`.[6] This site provides an archive of runs against a large range of text collections; includes tools for measuring effectiveness of a new, uploaded set of runs; compares sets of runs using a range of statistical methods; and provides an ongoing record of effectiveness of new and historical methods. However, uptake of this site has been slow.

The national and international guidelines for researchers, such as the *Australian Code for the Responsible Conduct of Research*,[7] emphasize the need for experimental data to be retained for a period of years, and that at a minimum it be available to other researchers for the purposes of verification. While 'data' in this context means the results of an experiment, rather than its subject, output such as digested tables are inadequate. What is required is the detail that indisputably confirms that an experiment took place. In practice, in IR, this will usually mean the data and code.

Summary

Information retrieval research has a laudable history of production of publicly available experimental materials, and of the diffusion of standard experimental techniques and measures. There have also, though, been significant missed opportunities: to place experiments on a sound statistical basis; to establish standards for the release of data and code; and to record and measure progress over time.

To return to the question of why we need such an infrastructure: we desire reproducibility, that is, the ability to run comparable (or, in the limit, identical) experiments and achieve results consonant with those of the original research. Reporting on the availability of research data and code allows the lesser goal of verifiability, that is, the ability to interrogate and re-analyze results to check their plausibility. A reader or reviewer who finds a result surprising, extreme, or implausible needs to be able to go to the experimental data, at whatever level it is available, and interrogate it.

The problems tackled by retrieval researchers continue to evolve. We are seeing in particular a shift of emphasis away from the holistic retrieval problem to work on new domains and tasks. This change and diversification in direction poses several challenges to the community, if it is to maintain its tradition of public data and standard methodologies. How do we resource, for example, recommender systems? How can we create a common, shareable dataset for investigating implicit user feedback? At the same time, the emergence of these new fields offers an opportunity not merely to extend the discipline's existing achievements, but fix its past omissions. For instance, the difficulties of shareable datasets may be addressed by establishing requirements for the release of experimental data; new sub-fields may be given greater direction by couching their tasks as problems to be measurably solved, rather than collections to be endlessly iterated over; and proper power analysis at the outset of an common experimental program can establish the scale and configuration of resources needed to properly support it.

We close with an observation on recent work published by Baggerly and Coombes,[8] who reverse-engineer microarray studies on responsiveness to cancer treatment, studies which omit the data and details needed for direct reproduction. Baggerly and Coombes find a catalogue of errors: sensitive and resistant labels for subjects switched; data columns offset by one; faulty duplication of test data; incorrect and inconsistent formulae for basic probability calculations; and so forth. They comment that 'most common errors are simple ... [and] most simple errors are common'. And these are not in obscure papers, but in large-team studies, which have lead to patent grants and clinical trials – trials in which, for example, errors in the original papers meant that patients were being given contra-indicated treatments. While the consequences of poor experiments in IR are not necessarily as grave, we intend that the work of our community be useful and substantial, and public infrastructure is required to ensure that similar problems are not perpetuated. To take another perspective, work is only of value if the gains it describes can be verified and incorporated by others, and we need public infrastructure and shared standards to achieve this goal.

[5] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel, "Improvements that don't add up: Ad-hoc retrieval results since 1998", *Proc. CIKM*, pages 601–610, Hong Kong, China, 2009.

[6] `evaluatir.org`

[7] `www.nhmrc.gov.au/_files_nhmrc/file/publications/synopses/r39.pdf`

[8] K.A. Baggerly and K.R. Coombes, "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology", *Annals of Applied Statistics*, 3(4), 2009 (`arxiv.org/pdf/1010.1092`).