The Future of Information Retrieval Evaluation

Douglas W. Oard

Abstract Looking back over the storied history of NTCIR that is recounted in this volume we can see many impactful contributions. As we look the future, we might then ask what points of continuity and change we might reasonably anticipate. Beginning that discussion is the focus of this chapter.

1 Introduction

In his book *The Third Wave*, Alvin Toffler placed what many have called the Information Age alongside the two most consequential transformations in human society, the introduction of agriculture and the industrial revolution (Toffler, 1980). That information retrieval will continue to play a central role in the coming years thus seems undeniable. One point of continuity between the current era and the flowering of science that helped to foster the industrial revolution is Lord Kelvin's admonition that "if you can not measure it, you can not improve it." Hence, the central role of information retrieval evaluation seems assured as well. That is not to say, however, that we will continue to measure our results in the same ways. Indeed, it seems reasonable to expect that information retrieval evaluation will continue to co-evolve along with changes in the information ecosystems that it serves. This chapter reflects on both the emergence of shared task evaluation and on present trends in information retrieval evaluation.

Douglas W. Oard
University of Maryland, College Park, MD USA e-mail: oard@umd.edu

2 First Things First

Shared task evaluation arose in information retrieval from the convergence of two broad lines of work. The first was the test collection tradition in information retrieval that dates back to the early Cranfield collections of the 1960's (Cleverdon, 1991). The central idea in a test collection is to model the behavior of a user by selecting some representative set of documents[1] to be searched, generating representative search topics, generating representative queries for those search topics, and finally generating relevance judgments for some useful set of query-document pairs.

It was the need for relevance judgments that ultimately led to the creation of shared task evaluation for information retrieval. Many early collections were exhaustively judged (i.e., all query-document pairs had a relevance judgment), but as the document collections became larger exhaustive judgments proved to be infeasible. The challenge of larger collections was compounded by the emergence of search topics for which relatively few documents in the collection would be relevant. It was those topics seeking rare documents that made random sampling unsuitable as a means of dealing with increasing collection sizes. The approach that was ultimately adopted, pooling, relied on a form of purposeful sampling in which samples were drawn only from document sets in which existing retrieval systems had difficulty distinguishing between document that were relevant and documents that were not. Ranked retrieval was becoming an increasingly widespread object of study at the time the idea of pooling was first tried in the Text Retrieval Conference, so this approach to sampling was generally operationalized as merging sets of documents that were highly ranked by one or more of several representative ranked retrieval systems (Voorhees and Harman, 2005). It was this need for contributions of results from a number of representative systems that led to the emergence of shared task information retrieval evaluation.

In the movie *The Right Stuff* about the early American space program, one of the characters observes on the importance of financial support with the pithy quote “No bucks, no Buck Rogers.” Shared task evaluation requires resources for planning and coordination, but most essentially for creating the relevance judgments. This side of the equation came from the Defense Advanced Research Projects Agency (DARPA) in the United States, where the voice of Lord Kelvin was strong. The competition for funding within DARPA was adjudicated in part using the “Heilmeier Catechism,” a set of questions to be answered by any new program, one of which is “What are the mid-term and final ‘exams’ to check for success?” DARPA had started a human language technology program, focusing initially on speech recognition, in 1986. Central to that program was a focus on evaluation. By 1990, DARPA was ready to expand its focus to include information retrieval. Hence was born the TIPSTER program, which in turn supported the early years of the Text Retrieval Conference (TREC).

[1] Although it is conventional to refer to documents, the term is often used inclusively to refer to other types of information objects as well.

As is sometimes the case when innovating, shared task evaluation rapidly evolved well beyond its initial focus on measurement. TREC did indeed produce test collections. Importantly those collections were shown to be reusable to a useful degree, thus permitting test collections developed in one year to be used in subsequent years as a basis for testing refinements to the system design. This approach, which came to be called evaluation-guided research, emerged in parallel in several research communities (e.g., information retrieval, speech recognition, and named entity recognition). It would be well recognized by machine learning researchers today as an early instance of supervised learning (albeit one with substantial human intervention in the early days). A second important thing that TREC did was that it produced baseline results to which future results could be compared. This facilitated the entry of new research teams, who could compare their systems against established baselines. A third innovation, was the emergence in 1996 of TREC's more narrowly focused "tracks" to support specific research goals. These three innovations – collections, comparisons, and communities – together serve as a useful frame for examining not just shared task evaluation in TREC, but approaches to information retrieval evaluation more generally.

Much has been written about the benefits of shared task evaluation, but when considering alternatives it is equally important to consider its limitations as well. Perhaps most obviously, shared task evaluation is expensive. For example, the cost of the first 18 years of TREC was calculated to be $29 million USD (Tassey et al, 2010), which is clearly well beyond what many individual researchers could support on their own. Two natural results of this are that some process for making investment decisions is needed, and those decisions must initially be made before seeing what the results will be. Those facts, in turn, tend to result in multi-year commitments to a research program so that insights generated in one year can be capitalized upon in the subsequent years. As a result, shared-task evaluations have a limited capacity to start on new lines of work. Perhaps even more importantly, the need for some decision process, whether centralized or consensus-based, results in there being some gatekeeper role beyond the individual researcher that must judge whether a broad line of research merits the community's attention. Moreover, schedule considerations result in proposals needing to be made early – typically more than a year before the first results will become available. None of these limitations are show stoppers for research problems that that require large-scale "team science" experimentation, but there are many settings (e.g., commercial research on problems with immediate operational implications, or a single student working alone on a novel problem in a 3-year Ph.D. program) for which shared-task evaluation is not sufficiently responsive.

A second critique of shared task evaluation is that it can generate a tendency towards convergence in methods, perhaps thereby delaying the exploration of important alternative approaches. To see an example of this we need look no further than the current widespread interest in neural "deep learning" methods. This sort of bursty convergence in which new techniques are rapidly explored by the community has benefits, but the degree of convergence that in engenders has risks as well. Importantly, this risk is not unique to shared task evaluations—it is simply the flip side of any approach in which researchers come together as a community to compare results in an evaluation-guided research setting.

3 The Shared Task Evaluation Ecosystem

In the two decades that followed TREC's creation, shared-task evaluation expanded at an impressive pace. Some notable examples (with the year in which they started) include:

- TDT (1996): The Topic Detection and Tracking (TDT) evaluation formed as a parallel evaluation venue to TREC to focus on streaming news content in text and speech (Wayne, 2000)
- NTCIR (1999): The focus of this volume, NTCIR formed as a counterpart to TREC with a focus on East Asia.
- CLEF (2000): Initially called the Cross-Language Evaluation Forum, CLEF initially spun out from the TREC CLIR track (Braschler and Peters, 2004).
- INEX (2002): The Initiative for Evaluation of XML Retrieval (INEX) formed independently to focus on retrieval of structured documents, and ultimately became a task in CLEF (Lalmas and Tombros, 2007).
- TRECVID (2003): The TREC Video Retrieval Evaluation (TRECVID) is a separate evaluation venue that initially spun out from the TREC Video Track (Smeaton et al, 2006).
- MIREX (2005): The Music Information Retrieval Evaluation eXchange (MIREX) implemented a large-scale infrastructure for evaluation, using algorithm deposit to accommodate copyright concerns (Downie et al, 2014).
- FIRE (2008): The Forum for Information Retrieval Evaluation (FIRE) has a focus on South Asia (Majumder et al, 2018).
- MediaEval (2010): The MediaEval Benchmarking Initiative for Multimedia Evaluation initially spun out from the CLEF VideoCLEF Task (Larson et al, 2017).

No such list could ever be complete, since shared task evaluation exists any time two or more research groups come together around an evaluation task. For example, several evaluations have been conducted in a national context, including in China, France, Russia and South Korea. Moreover, the boundaries between information retrieval and the cognate disciplines of natural language processing and speech processing are porous, and there have been evaluations in those communities that certainly bear on information retrieval research. For example, there have been evaluations of both event detection and summarization in the Text Analysis Conference (TAC),[2] and there has been evaluation of spoken term detection in the Open Keyword Search evaluation,[3] both of which are, like TREC, organized by the National Institute of Standards and Technology (NIST).

[2] https://tac.nist.gov/

[3] https://www.nist.gov/itl/iad/mig/open-keyword-search-evaluation

All of those are TREC-like, in that they are evaluation venues independent of any larger event in which participants actually come together in a workshop-like setting to discuss their results. There are, however, numerous additional examples in which one or both of those characteristics are not present. Cases in which a shared task evaluation is organized in conjunction with a larger event are sometimes called "data challenges." The granddaddy of these data challenges was perhaps SensEval, named for its focus on Word Sense Disambiguation. SensEval initially formed independently in 1998, but then associated itself with a workshop starting in 2001 (and later changed its name to SemEval in 2007, reflecting its broader interest in semantics).[4] The Conference on Computational Natural Language Learning (CoNLL) started a shared task in 1999,[5] followed in 2001 by the Document Understanding Conference (DUC, which despite its name was actually a workshop series, initially held at SIGIR). SemEval and the CoNLL shared task continue as data workshops to this day, having been joined by many others (e.g., the Big Data Cup[6]); DUC ultimately became a standalone venue (as TAC).

If data challenges are one step away from independent shared-task evaluations such as NTCIR and TREC, prize-based competitions represent an even further departure from the independent conference paradigm. Perhaps the best known members of this genre of shared task evaluation are Kaggle[7] and the Netflix Prize (Bennett et al, 2007). The Netflix Prize started in 2007 with the goal of advancing research on large-scale recommender system. Netflix, a provider of streaming video services, offered participants access to a large collection of anonymized usage data, offering a $1 million USD reward for achieving a 10% improvement over the company's best current algorithm. Kaggle was founded in 2010 to capitalize on similar opportunities for a broad range of problems, acting as a forum within which communities could form around specific challenges. Kaggle has in turn given rise to other similar venues, including Tianchi[8] and Innocentive.[9] Prize competitions often operate as a market in which sponsors define the task and then pay the prize in exchange for a license to commercially use the technique that wins the competition. This stands in sharp contrast to the non-commercial ethos of many of the independent shared-task evaluations listed at the start of this section, which focus principally on pre-competitive basic research. Indeed, some of the independent shared-task evaluation venues actively seek to minimize the competitive aspect of shared task evaluation, in part because of concerns that a "winner-take-all" perspective might depress participation by teams who would otherwise be able to contribute diversity to the document pools that will be judged for relevance.

[4] https://aclweb.org/aclwiki/SemEval_Portal

[5] http://www.conll.org/previous-tasks

[6] http://cci.drexel.edu/bigdata/bigdata2019/BigDataCupChallenges.html

[7] https://www.kaggle.com/

[8] https://tianchi.aliyun.com/

[9] https://www.innocentive.com/

4 A Brave New World

In the movie *The Wizard of Oz*, Dorothy observes at one point that “we’re not in Kansas anymore.” So it is with information retrieval evaluation as well—there are now many more things under the sun than just shared-task evaluation. At least four alternatives can be discerned, each of which has its own strengths and weaknesses.

The first to emerge were project data repositories. Perhaps the best known of these is the Linguistic Data Consortium (LDC) at the University of Pennsylvania,[10] which was founded in 1992 with support from DARPA to serve as a repository for the human language technology community. LDC, and similar organizations around the globe (e.g., the European Language Resources Association, ELRA,[11] or the Linguistic Data Consortium for Indian Languages, LDC-IL[12]) permit researchers to deposit test collections that they have created that may in the future be of use to others. In this way, what were once internal evaluations on data generated within a project can become shared, and over time can emerge as a shared-task reference to which future work can be compared. Perhaps the most successful example of this general approach is the University of California Irvine Machine Learning Repository (Dua and Graff, 2017), which provides test collections that serve as standard references among machine learning researchers (notably including some text classification researchers).

Project data repositories help with community formation and with providing a basis for comparisons, but (at least when serving solely as repositories) they do not create collections. That’s where crowdsourcing comes in. Shared task evaluations in the TREC heritage predate the World Wide Web, but as user-generated content became more pervasive in what came to be called Web 2.0, crowdsourcing emerged as an alternative way of obtaining relevance judgments (Alonso, 2019). Crowdsourcing can be used in many ways in the evaluation of information retrieval systems, but perhaps the most obvious alternative to the approach used in shared-task evaluation is to simply pay crowdworkers to create relevance judgments. Because queries are often treated as independent in information retrieval test collections, the relevance judgment task is easily distributable across multiple crowdworkers. At least two concerns arise when this is done. First, crowdworkers may be less well trained or less attentive to their task than relevance assessors who work at a central facility as their primary job would be. This concern has spawned a line of work on assessing the accuracy of crowdworkers. Second, one common approach to managing those risks, having several crowdworkers vote on the correct relevance label, has the effect of subtly redefining relevance (for purposes of evaluation) away from the opinion of an individual and toward the consensus of a group. Balanced against these concerns, however, are the speed, scalability, and relative affordability of crowdsourcing. Moreover, the diversity of available crowdworkers can provide access to people with needed skills (e.g., language skills or some types of topic expertise) that

[10] https://www.ldc.upenn.edu/

[11] http://www.elra.info/en/

[12] http://www.ldcil.org/

simply might not be available otherwise. For these reasons, crowdsourcing can offer transformational advantages to isolated researchers who, for reasons of location, funding, or problem specificity simply can not plausibly create a shared-task evaluation. Note, however, that crowdsourced test collections need not remain isolated once they have been created, since they can be shared though data repositories.

Creating test collections is, however, just one of at least two ways in which crowdsourcing can be used for information retrieval evaluation. An alternative is to study the actual use of a system using crowdworkers. Test collections have many desirable attributes, but no test collection captures every important aspect of actual information retrieval tasks. Evaluating information retrieval systems in actual use has traditionally been a focus of user studies, and crowdsourcing offers an opportunity to extend the user study beyond the researcher's laboratory across the Internet to meet the users where they are. This opens new opportunities to intermix research using test collections (which are optimized for affordably repeatable evaluation under controlled conditions) and user studies (which offer higher fidelity evaluation, but at incremental cost each time an experiment is run).

There are, of course, limits to the user studies that can be run with crowdworkers. In addition to the obvious limits imposed by affordability considerations, fidelity is always a concern when paying a user to perform a task that you have designed. One way of addressing both of these concerns is to perform what has come to be called online evaluation (Radlinski and Craswell, 2010). The basic approach is simple. First, build a system that becomes so popular that there will be a large number of users whose behavior you can study. Then design experiments in which some aspect of the system (the independent variable) is changed, and the effect is observed by observing some behavioral signal (the dependent variable). Variants on this idea include A-B testing and interleaving. Of course, the first step there—creating systems that have a large user population—can be a tad expensive! But once such a system is available, a very large number of experiments can be run at low cost. Naturally, this approach is popular among commercial services that have a large user base. Batch evaluation measures have also been tuned using query logs, thus more closely linking online and offline (i.e., batch) evaluation (Ferrante et al, 2014).

5 Trendlines

One thing that should be clear from the story to this point is that independent shared-task evaluations such as NTCIR are now just one part of an increasingly diverse and specialized evaluation ecosystem. But that is just one of many trendlines that together will continue to reshape the future of information retrieval evaluation. This section reviews several others.

It is fashionable today in many contexts to remark on convergence. What used to be separate devices (e.g., phones, computers and televisions) now are one. What used to be stored on separate media separate media (video, images, documents, datasets) are now all stored as digital files. What used to be separate functions (computing and communication) are now becoming nearly inseparable. All of these are examples of convergence. We are seeing examples as well of convergence across fields. Information retrieval researchers use speech and language technologies that in an earlier time would have been thought of as separate fields. Database researchers work with semi-structured data that the information retrieval community would recognize as structured documents. Data scientists analyze interaction patterns to help optimize the user experience. Interactive information retrieval research draws in equal measure on insights from information retrieval and human-computer interaction. Work on fairness, accountability and transparency in machine learning finds application in designs of information retrieval systems that are informed as much by social as by technical goals. This convergence of disciplines creates new opportunities, but at the same time it challenges the notions we have developed over time about what is, and what is not, information retrieval.

If convergence disrupts what it is we think we do, the Internet is perhaps even more disruptive because it changes where we can do it. In an earlier era, information retrieval research suffered from what we might call the tyranny of geography. There were a few places in the world where top flight information retrieval research was going on, and it was much easier to get into the field if you could get to one of those places. Today, information retrieval is taught in many places, and indeed well over half the world's population has access to free online courses on the topic. Cloud computing has gone some distance towards democratizing access to high-end computing, and the widely available low-end computing infrastructure has capabilities that were unavailable anywhere on Earth just a few decades ago. We have by no means completely erased the tyranny of geography at this point in history, but it is quite clearly on the wane.

Solving one problem often reveals another, and so it is with the competition for our attention. For essentially all of human history, and with rare exception, information was scarce and human attention was relatively abundant. No one with an Internet connection can fail to notice that the situation today has sharply reversed, and that it is information that is abundant, while it is human attention that is now scarce. If we view our job as helping to separate the wheat from the chaff, it should be clear that this trendline suggests that we'll have no shortage of important problems to work on.

Another trendline worthy of remark is that the nature of gatekeeping is shifting. Long ago we had to choose between a Web track, a filtering track, an interactive track, or whatever other ideas were put forward, because venues like NTCIR simply could not do everything. It's still not possible to do everything, but the emergence of options such as crowdsourcing and online evaluation greatly expand the range of information retrieval evaluations that can be conducted. That's not to say that there will be no gatekeepers. Peer review, for example, will continue to play some role with regard to what gets published. But to the extent that come of gatekeeping can be shifted from before the work is done to after the results become available, that could help to enhance the diversity of the research ecosystem.

One foundational assumption in information retrieval is that information wants to be found, and that our job is to find it. That's actually probably not true for much of the information in the world, however. Examples abound of information that should not be found. In Europe, the right to be forgotten is a right not to have specific information about you found. In many countries with legislation that promoted freedom of access to government information, specific exemptions identify types of information that should not be disclosed. We have debates today about which types of information governments or commercial entities should be allowed to use, and for what purposes. Article 12 of the Universal Declaration of Human Rights declares privacy to be a human right, with all of the complexity that operationalizing the meaning that such a statement entails. In an earlier era, information retrieval research encountered restrictions on access from time to time, and in such cases the response of researchers was generally to focus instead on the many cases in which access control was not a problem.

We are perhaps now nearing the limits of that strategy. Consider the fact that almost all of the words produced on the planet—probably upwards of 99%—are spoken, not written. Couple that with the fact that well over half that speech is produced in the presence of a networked recording device (e.g., a mobile phone). And couple that with the fact that both the speed and accuracy of technology for automatically transcribing that speech has improved by leaps and bounds in recent years. At present, we are largely disregarding all of that content simply because we have no idea how to protect those parts that need to be protected. This has implications for research, of course, but it has implications for evaluation design as well. We have grown up in an era in which we all learned to respect copyright when dealing with test collections. We now need to learn how to deal with sensitive content that will in some cases prevent us from distributing test collections. That does not mean that we won't be able to do shared task evaluations, but it does mean that we'll need to think anew about how best to do them. The Netflix Prize, for example, ended because of a privacy lawsuit.

It has been said that "data is the new oil," a catchy phrase intended to illustrate that there is money to be made. At one time, most information retrieval researchers worked in universities. Today, the the balance has shifted very strongly in favor of industry. That's good news, because that's where the money is, so there is now vastly more research on information retrieval being published than ever before. Its also good news because industry has access to evaluation opportunities that simply can't be replicated elsewhere, most notably with online evaluation. And its also good news because all this commercial activity is helping to bring new problems to the attention of the information retrieval research community.

6 An Inconclusion

It is traditional to end a chapter with a conclusion, but when writing about the future perhaps it would be wise to recognize that the evidence we see today is not sufficiently conclusive to allow us to see that future with clarity. Herewith, therefore, some inconclusive remarks. Josef Schumpeter is best known for his description of creative destruction, a process by which innovations result in the displacement of earlier enterprises that had been built to leverage earlier innovations (Schumpeter, 1942). As the convergence examples above indicate, creative destruction is at least as vibrant today as it was when Schumpeter was writing. Independent shared task evaluations such as NTCIR were created in an earlier era, to fill a role that has since been augmented, and perhaps partially replaced, by other approaches to information retrieval evaluation. It therefore seems timely to consider the question of what role NTCIR, and other independent shared task evaluations, may play in the future. Fortunately, the very name of NTCIR, the NII Testbeds and Community for Information Access Research, can help to guide that discussion.

N is for NII, the National Institute of Informatics. NII, like NACSIS before it, has been a source of leadership, not just in information retrieval evaluation, but in the emergence of a vibrant information retrieval research community in Japan specifically, and in East Asia more generally. Ultimately, NII is made up of people, and it is the choices made by those people that will define the future leadership role of that institution. With wise choices, that N will remain a capital letter.

T is for Testbeds. As explained throughout this chapter, the testbeds of the sort NTCIR has created (principally, test collections) are one part of what is now a rich ecosystem of evaluation methods. There will surely continue to be demand for test collections, but shared task evaluations like NTCIR are no longer the only affordable way in which test collections can be created, and we now live in a world in which a broader range of testbeds can be affordably constructed. We therefore may see the T in NTCIR decline somewhat in its impact, perhaps becoming a lower case t.

C is for Communities. For all the trendlines that portend change, one thing that seems unlikely to change any time soon is human nature. Humans are social animals, and research is a social enterprise. We need ways of bringing people together around new problems, ways of helping new people to join those communities, ways of creating the kinds of shared understanding that are needed to learn from each other how best to solve those problems, and ways of defining what it would mean to succeed at solving those problems. Shared task evaluations like NTCIR serve all of those functions. The C in NTCIR seems destined to remain a capital letter.

I is for Information access. As noted at the start of this chapter, we live in an information age, and it therefore seems unlikely that focus of NTCIR on information would be likely to diminish. The same might not be said for access however, since we are now seeing some convergence of research on (at least) information access, information creation, information understanding, information manipulation, and information policy. So the I in NTCIR seems sure to remain capitalized, but we may see some shifts in what it stands for.

R is for Research. We might think of research in three ways. The most obvious is to think narrowly in terms of some specific type of research, such as evaluation guided research or statistical hypothesis testing. An alternative is to think of research more inclusively, as any systematic way of generating new and generalizable knowledge. And a third alternative would be to think even more broadly about research, as an undergraduate student might, as self-directed learning about new things. Many people who do not see themselves as researchers in the first or second sense need to do research in the third sense. One way or another, the R seems likely to remain since it is central to the self-image of NTCIR, but perhaps the meaning of that R will shift somewhat over time.

Well, there we have it. It seems that we can look forward to a world in which NtCIR remains, and all we will need to do is to figure out what it actually stands for!

Acknowledgements This work has been supported in part by NSF grant 1618695. The author is grateful to Bonnie Dorr, Nicola Ferro, Donna Harman, Gareth Jones, Noriko Kando, Martha Larson, Prasenjit Majumder, Paul McNamee, Jim Mayfield, Mandar Mitra, Carol Peters, Mark Sanderson, Ian Soboroff, Ellen Voorhees, and Charles Wayne for discussions over the years that have helped to shape his thinking on this topic.

References

Alonso O (2019) The practice of crowdsourcing. Synthesis Lectures on Information Concepts, Retrieval, and Services 11(1):1–149

Bennett J, Lanning S, et al (2007) The Netflix prize. In: Proceedings of KDD cup and workshop, New York, NY, USA., vol 2007, p 35

Braschler M, Peters C (2004) Cross-language evaluation forum: Objectives, results, achievements. Information retrieval 7(1-2):7–31

Cleverdon CW (1991) The significance of the cranfield tests on index languages. In: Bookstein A, Chiaramella Y, Salton G, Raghavan VV (eds) Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum)., ACM, pp 3–12, DOI 10.1145/122860.122861, URL https://doi.org/10.1145/122860.122861

Downie JS, Hu X, Lee JH, Choi K, Cunningham SJ, Hao Y (2014) Ten years of MIREX (music information retrieval evaluation exchange): Reflections, challenges and opportunities. In: Wang H, Yang Y, Lee JH (eds) Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014, pp 657–662, URL http://www.terasoft.com.tw/conf/ismir2014/proceedings/T119_342_Paper.pdf

Dua D, Graff C (2017) UCI machine learning repository. URL http://archive.ics.uci.edu/ml

Ferrante M, Ferro N, Maistro M (2014) Injecting user models and time into precision via markov chains. In: Geva S, Trotman A, Bruza P, Clarke CLA, Järvelin K (eds) The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’14, Gold Coast , QLD, Australia - July 06 - 11, 2014, ACM, pp 597–606, DOI 10.1145/2600428.2609637, URL https://doi.org/10.1145/2600428.2609637

Lalmas M, Tombros A (2007) Evaluating XML retrieval effectiveness at INEX. SIGIR Forum 41(1):40–57, DOI 10.1145/1273221.1273225, URL https://doi.org/10.1145/1273221.1273225

Larson M, Soleymani M, Gravier G, Ionescu B, Jones GJ (2017) The benchmarking initiative for multimedia evaluation: Mediaeval 2016. IEEE MultiMedia 24(1):93–96

Majumder P, Mitra M, Sankhavara J, Mehta P (eds) (2018) Proceedings of the 10th annual meeting of the Forum for Information Retrieval Evaluation, FIRE 2018, Gandhinagar, India, December 06-09, 2018, ACM, DOI 10.1145/3293339, URL https://doi.org/10.1145/3293339

Radlinski F, Craswell N (2010) Comparing the sensitivity of information retrieval metrics. In: Crestani F, Marchand-Maillet S, Chen H, Efthimiadis EN, Savoy J (eds) Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010, ACM, pp 667–674, DOI 10.1145/1835449.1835560, URL https://doi.org/10.1145/1835449.1835560

Schumpeter JA (1942) Capitalism, Socialism and Democracy (1942). Routledge

Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval, ACM, pp 321–330

Tassey G, Rowe BR, Wood DW, Link AN, Simoni DA (2010) Economic Impact Assessment of NISTs Text REtrieval Conference (TREC) Program. RTI International

Toffler A (1980) The Third Wave. Morrow

Voorhees EM, Harman DK (2005) TREC: Experiment and evaluation in information retrieval. MIT Press

Wayne CL (2000) Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In: Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece, European Language Resources Association, URL http://www.lrec-conf.org/proceedings/lrec2000/pdf/168.pdf