Against Recall:
Is it Persistence, Cardinality, Density, Coverage, or Totality?

Justin Zobel Alistair Moffat Laurence A. F. Park

Department of Computer Science and Software Engineering,
The University of Melbourne,
Victoria 3010, Australia
jz,alistair,lapark@csse.unimelb.edu.au

Abstract

The concept of recall has been one of the key elements of system measurement throughout the history of information retrieval, despite the fact that there are many unanswered questions as to its value. In this essay, we review those questions and explore several further issues that affect the usefulness of recall. In particular, we ask whether it is reasonable to expect to be able to measure recall; whether some researchers are conflating the concepts of recall and answer set cardinality; and whether it is plausible that a user would rely on a belief that a system is "high recall" to deeply explore an answer list. Combined with earlier observations about the unknowability of recall, and the lack of a plausible user model in which recall is a measure of satisfaction, we conclude that use of recall as a measure of the effectiveness of ranked querying is indefensible.

Introduction

Precision and recall are widely used as the basic elements from which many measures of information retrieval (IR) system effectiveness are compounded. In typical current retrieval scenarios, systems return rankings, that is, document lists ordered by some scoring criteria. *Precision* measures the density of relevant documents in the head of this list, that is, amongst the first k documents, while *recall* measures the proportion of the total set of relevant documents that appear in the k-element prefix. Varying k gives a relationship between recall and precision.

Explicit use of precision is common in papers on search effectiveness. Explicit quantitative use of recall, in its formal definition, is now rare; in the 2008 SIGIR proceedings, there are only a few mentions in 50 or so search-oriented papers. However, implicit use of recall – or quantities that are factors in it or derived from it – is widespread, in particular because it is a component of mean average precision.

In qualitative settings recall is still a concept that authors often appeal to. In reality we believe that such references are to one of several related, but more useful concepts: persistence, cardinality, coverage, density, and totality. We first define what we mean by these concepts.

Persistence

One metaphor for search behavior is to consider the user's *persistence* – the likely number of documents in the ranking that they will examine. A high-persistence user examines more documents than a low persistence user; it follows that they will expect their searches to result in higher recall, even though they are unlikely to know what that real level is.

A simple model of persistence is given by Moffat and Zobel [2008], in which users are hypothesized to always inspect the first document in the ranking, and thereafter to proceed to the next document with probability p; thus the probability of inspecting exactly k documents is $(1-p)p^{k-1}$. There are obvious variations on this formulation, such as to allow the probability p to vary with k or with discovery of relevant material; however, the formulation as given has the attractive properties of loosely capturing a typical behavior and of being tractable to mathematical analysis, such as precise determination of upper and lower bounds in the presence of uncertainty.

Whether any particular formulation is an accurate model of user behavior is open to debate. However, at a more intuitive level, the notion of persistence captures a broad property of search: persistence is the extent to which users continue to inspect answers to a search, having started at the top of the ranking. In particular, while a user may occasionally skip a few documents or a few pages of results – for example, to get a sense of whether more relevant material is likely to be found – they do not skip to the end or middle of the ranking. Indeed, in many typical search contexts, from the user's perspective most rankings have a beginning but no end. Hence, in a high persistence search, the user may inspect many pages of results, but starting from the front of the ranking; while for a low persistence search the user may not even look at every snippet in the first page of results, and will quickly stop and issue a fresh query.

Note that as we have defined it, persistence is independent of relevance, and it relates to the number of documents accessed, not the number of them that are relevant.

Cardinality

Another useful concept is the *cardinality* of a search: the count of relevant documents seen so far in processing of an answer list. Cardinality gives a loose measure of the amount of information the user has seen. In the context of a query with a known number of relevant documents, cardinality is proportional to recall, but it is a fallacy to apply this observation when the number of relevant documents is unknown and claim that high cardinality implies high recall; we explore this issue below in the context of density.

Cardinality is widely used as an surrogate for recall. Both are precisely defined in terms of numbers of relevant documents observed, but recall is normalized, and what a researcher thinks of as being a "high recall search" may well in fact be one in which many relevant documents are found, in other words, be a high cardinality search. In the context of a user scanning a ranked list of documents from the start, cardinality is also directly proportional to precision.

Coverage

An orthogonal property is that of *coverage*, that is, the fraction of the query that is answered from the observed search results. Based on this measure, a user may be happy with a single document, if it, for example, provides a fact or an answer to a specific question. Alternatively, the user may seek topic coverage, such as a set of related facts, or information on several different aspects of a topic. They may seek a key fact, but only be able to discover which fact is key through other information returned in the search. They may continue to inspect documents until they are satisfied that the coverage

of information on the topic is likely to be complete. Or they may simply continue until they feel they have learnt enough to be informed on a topic, without concern as to whether their knowledge is comprehensive. Coverage is thus related to the classification of search types into information seeking, exploratory, and so on.

It is attractive to use coverage as a metric in the context of web search, as it concerns how much users have learnt, rather than how many documents they have seen. However, coverage is not always easily measured. For a query, relevance assessors need to identify and categorize each fact or aspect of relevance present in the document corpus, and then annotate each relevant document to list the facts that it addresses. We can, then, use coverage as a motivation – as a thing that we in principle wish to assess, and also, with some certain difficulty, as a measure. However, it should be obvious that recall and coverage are at best loosely related; complete recall implies complete coverage (at least, within the limits of the collection being used). In all other situations, knowledge of recall gives no information about coverage, and vice versa.

Also worth noting is that maximizing recall is very different to maximizing coverage. For example, in the context of search on what might be described as *open* collections, such as a vast number of web pages from wildly diverse sources, the same information is likely to be present in many documents, and the documents later in a ranking are often either duplicates of earlier documents, or contain no additional information. Thus, in processing a ranking, we may continue to discover relevant documents, but without adding to our knowledge, and hence without adding to the coverage.

Density

We define *density* as the localized precision in the ranking, that is, the fraction of relevant documents within some defined nearby section. Given that search systems have as their goal to bring relevant documents to the top of the results ranking, we expect density to decrease the further down the ranking we proceed.

In the absence of any other information, when the density is high at some point in the ranking, it is an indication that recall is likely to still be low, since relevant documents can be expected to continue regularly occurring as the ranking is inspected. This inference is correct regardless of the cardinality that has been reached through to that point.

Conversely, low density near the top of a ranking may or may not be evidence of low (eventual) cardinality, since there is no way of knowing whether relevant documents will continue to be found infrequently, or whether their rate of occurrence will drop away completely. Hence, for some search tasks, low density is a discouragement, and likely to precipitate a change in tactics, perhaps cessation of the search process, or the issuing of a followup query using different terms. Or, for search tasks in which the objective is high cardinality, low density may simply mean that the inspection of documents must continue – low density might indicate that an increased level of persistence is required.

Finally, a low density segment of the ranking that follows a high density segment is usually a signal that cardinality (and thus recall) is approaching its maximum for this query, and may be interpreted by the user as suggesting that there are relatively few relevant documents that have not yet been discovered – a notion that is explored in the next section.

Totality

It is usual for certain "high recall applications" to be cited to rebut suggestions that recall is of little importance. Examples that are routinely given include searching for precedents in legal cases; searching for medical research papers with results that relate to a particular question arising in clinical

practice; and searching to recover a set of previously observed documents. While we agree that these are plausible search tasks, we dispute that they are ones in which recall provides an appropriate measurement scale. We argue that what distinguishes these scenarios is that the retrieval requirement is *binary*: the user seeks *total recall*, and will be (quite probably) equally dissatisfied by any approach that leaves any documents unfound. In such a situation, obtaining (say) 90% recall rather than a mere 80% recall is of no comfort to the searcher, since it is the unretrieved documents that are of greatest concern to them, rather than the retrieved ones.

If recall is to be of any interest as a measure, the user needs to have a calibration, or estimation mechanism, available. For example, if a user examines 20 (or maybe 200) documents in a ranking beyond the last that is identified as being relevant, the low density that they observe may be used as evidence in favor of a hypothesis that there are no more relevant documents to be found. But the user will then test that hypothesis before accepting it as true, by looking at further documents in the ranking; by running other queries using term synonyms suggested by the documents already seen; by running the same queries using other search services; and by exploring links or citations to and from the documents already encountered. They will stop their search only when they have confidence in their hypothesis of totality. That is, in practical situations, total recall is not something that is established by a single query; it is something that is hypothesized (but never fully proven) as a result of a range of techniques all failing to identify further relevant documents. Indeed, it is difficult to see how a user could have confidence that a search had achieved high recall without using further queries and evidence to establish that the pool of relevant documents has been exhausted. The idea that a single search will be used to find all relevant documents is simplistic.

User experience

Having argued that recall should be differentiated from all of persistence, cardinality, coverage, density, and totality, what then is left? Is there still a role for recall in information retrieval applications, measured precisely as the fraction of the relevant documents that have been retrieved?

Cooper [1968, 1973], Saracevic [1995], and Moffat and Zobel [2008] suggest that the answer is "no". They argue that the user experience is the ultimate measure of search utility, and point out that in the great majority of search situations – the exception being Boolean querying – the user cannot know the number of relevant documents available. A key issue then becomes that recall cannot possibly correspond to user satisfaction, since the user cannot know (even within orders of magnitude) how many relevant documents are available. As Cooper [1973, page 95] notes, "surely a document which the system user has not been shown in any form ... does that user neither harm nor good."

Given the differences between persistence, cardinality, coverage, density, and totality that we explored above, all of which have some direct effect on the user's searching experience, we again suggest that recall is an unnecessary – and perhaps even distracting – measure of search effectiveness.

Using a high recall system?

Suppose that a user is is confident that the engine they are using is designed for high recall (whatever that actually means), and also that they understand the concept of recall: in particular, that recall is distinct from cardinality. They appreciate that the number of relevant documents per query is a quantity that wildly varies, and that some queries will have only a few answers, while others will have hundreds. (This range of numbers of relevant documents is more or less that observed in the TREC Ad-Hoc experiments, in which queries were chosen to have reasonable numbers of relevant

documents to allow the assessment process to produce useful results. Whether randomly sampled queries would have similar bounds is unknown, but seems unlikely.) Suppose further that the user is determined to issue a single query only, and will inspect the resultant ranking until satisfied that no more relevant documents will be found in this ranking, or in the ranking for any other query that they have might performed instead.

For the moment, put aside that such behavior seems innately contradictory, and that it requires that the user be supremely confident that the query is so all-encompassing and well-formulated that it is capable of matching all relevant documents; and forget that this belief means that the user must be willing to inspect large numbers of irrelevant documents rather than re-probe the collection with a fresh query, no matter what is revealed by the documents they inspect. Independent of these concerns, the hypothetical high-recall user must still have a stopping or satisfaction criterion for ending their inspection of the answer list. It must be some property of the list of answers; presumably, that the tail of the inspected part of the list is so sparsely populated with relevant documents that the user can assume that no more answers are forthcoming. As full inspection of the list is not feasible, it is not obvious that any other criteria could be used.

That is, at some stage the user must stop processing the list of answers, and must do so based only on that which has been observed, namely, cardinality (and hence precision), and local density. Suppose the answer list is broken into digestible pages of say 10 answers each. Does a high-recall user stop when the most recent page is found to be full of relevant documents, or empty? If it is the former, they can have no belief of totality, and thus no confidence that recall can be high. If it is the latter, they base their decision solely on the observation that density has become low, which is not actually related to recall.

Yet contrasts between "high recall" and "high precision" systems seem to imply that low density (and thus low precision) is evidence that the system is high recall. From the perspective of a user, this seems highly implausible.

Pragmatic concerns

Whether recall is knowable in any large but bounded collection is unclear. The TREC assessment process was designed to identify a complete or near-complete set of relevant documents for every query, through pooling the results from the participating systems; these are assumed to be numerous and diverse, although in practice systems represented a rather smaller number of distinct techniques. Queries in many of the tracks, in particular the first years of the Ad-Hoc track, were hand-chosen in an attempt to ensure that numbers of answers were within a preferred range. Even so, for some queries not only were the sets of relevant documents clearly incomplete (see, for example, Zobel [1998]), but the range of possible sizes of these sets had a high level of uncertainty. That is, despite a substantial effort to gather complete sets of documents – for reasons such as allowing reliable reuse of the collections – and despite compromises to the integrity of the experiments such as non-random query selection, for these queries recall remains essentially unknown.

If the set of documents is effectively unbounded, as is the case on the web, then, after a finite judging process, some remaining unjudged documents must be relevant. On even a moderately sized static web crawl, the possibility of finding all relevant documents for a typical query seems remote.

Note that many composite measures of retrieval effectiveness, such as average precision (AP, or MAP) and normalized discounted cumulative gain require knowledge of the number of relevant documents as part of their formulation, and that, in the absence of comprehensive relevance judgments, any system comparison using these metrics must therefore include an unknown amount of imprecision. We should not be using measures on which we cannot place confidence bounds.

Conclusions

Measurement of recall may be of interest for those legal and medical applications where all relevant documents are required. However, the relationship to ranked search is at best tenuous, and arguments for recall in these expert cases has little relevance to the kinds of search that are investigated in the vast bulk of the IR literature. In summary,

- Recall is unrelated to persistence.
- Recall is unrelated to cardinality.
- Recall is unrelated to coverage.
- Recall is unrelated to density.
- Totality is a recall-related concept, but it is not plausible that a single search will be used to find all relevant documents, and total recall is established only after a range of techniques have failed to find further relevant documents.
- Recall is unrelated to the quality of the user experience.
- Measurement of recall is neither feasible nor meaningful in web search.

That is, we are aware of no justification for implicit or explicit use of recall as a measure of search satisfaction. Claims made about recall in any particular searching context should, we believe, be accompanied by a clear statement addressing the concerns we have raised. Recall should not be used without explicit justification.

References

W. S. Cooper. Expected search length: A single measure of retrieval based on the weak ordering action of retrieval systems. *American Documentation*, 19(1):30–41, January 1968.

W. S. Cooper. On selecting a measure of retrieval effectiveness: Part I, the 'subjective' philosophy of evaluation. *Jour. of the American Society for Information Science*, 24:87–100, March 1973.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):(2)1–27, December 2008.

T. Saracevic. Evaluation of evaluation in information retrieval. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proc. Eighteenth Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 138–146, Seattle, Washington, July 1995. ACM Press, New York.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. Twenty-First Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.