

Evaluating Information Retrieval: Precision, Recall, and the Trade-off between Them

Ninad R. Vaidya¹, Sharvari K. Deodhar², Tanmay V. Bhide³, Madhura M. Soman⁴

ABSTRACT

An information retrieval system, such as a search engine, answers a query by returning the items it judges relevant from a large collection, and how good it is depends on two things: whether the items it returns are actually relevant, and whether it returns the relevant items there are; these are measured by precision and recall, and the two are usually in tension, so that improving one tends to worsen the other. This study examines the evaluation of information retrieval through precision and recall and the trade-off between them. Precision, the fraction of the returned items that are relevant, and recall, the fraction of the relevant items that are returned, were examined together as a retrieval system was made to return more or fewer items, so that the trade-off between them, and how it shapes evaluation, could be seen. Precision and recall traded off against each other: when the system returned only the few items it was most confident in, those were mostly relevant, giving high precision, but many relevant items were left out, giving low recall; when it returned many items so as to catch the relevant ones, recall rose, but more irrelevant items were included, so precision fell. Plotting precision against recall traced a curve that fell as recall rose, showing that high precision and high recall could not generally be achieved at once. Which to favour depends on the task: where it matters that what is returned is relevant, precision is preferred, while where it matters that nothing relevant is missed, recall is preferred, and a single operating point on the curve is chosen accordingly. Because neither alone tells the whole story, the two are reported together or combined into a single measure. The study shows that precision and recall trade off in information retrieval, and it discusses how this governs the evaluation and tuning of search systems.

KEYWORDS: *Information retrieval, Precision, Recall, Search evaluation, Precision-recall trade-off, Relevance, Search engines, F-measure, Information systems*

INTRODUCTION

An information retrieval system answers a query by searching a large collection of items, documents, web pages, records, and returning those it judges relevant to the query. Search engines are the most familiar example, but the same task arises wherever relevant items must be found among many. The value of such a system lies in how well it's returned items match what the user actually wants, so judging the quality of retrieval is essential to building and improving these systems (2, 6).

Quality here has two distinct aspects, which can pull apart. One is whether the items the system returns are actually relevant: a system that returns many irrelevant items alongside the relevant ones wastes the user's effort. The other is whether the system returns the relevant items that exist: a system that returns only a few of the relevant items, missing the rest, leaves the user without information they sought. A good system should both return relevant items and not miss relevant items, but these are different demands, and a system may do well on one and poorly on the other (3, 7).

These two aspects are measured by precision, the fraction of the returned items that are relevant, and recall, the fraction of the relevant items that are returned. Crucially, the two are usually in tension, so that changing a system to improve one tends to worsen the other. This paper examines the evaluation of information retrieval through precision and recall and the trade-off between them, examining the two together as a system returns more or fewer items. The aim is to show how precision and recall trade off and to discuss how this governs the evaluation and tuning of search systems.

LITERATURE REVIEW

The evaluation of information retrieval has been studied since the foundational experiments that established how retrieval quality should be measured, introducing the systematic comparison of systems by their ability to return relevant documents (1). From this developed

the central measures of precision and recall, the fraction of returned items that are relevant and the fraction of relevant items returned, which are set out in all the texts on information retrieval (2, 3, 6).

Measures and the Trade-off

The literature describes how retrieval is evaluated:

- Precision and recall measure the relevance of what is returned and the completeness of retrieval (3, 6).
- Precision and recall trade off, traced by the precision-recall curve (7).
- They are combined into single measures such as the F-measure, and others such as nDCG rank quality (6, 8).
- Standardised collections and evaluations compare retrieval systems (1, 10).

A recurring theme is that precision and recall trade off, so both must be considered to judge a system.

The literature establishes that precision and recall are usually in tension, the relationship traced by the precision-recall curve, on which raising recall lowers precision and vice versa, so that neither alone characterises a system and the two must be considered together (3, 7). It develops ways to summarise the two, including the F-measure that combines them into a single number, and other measures such as ranked-relevance measures that account for the order in which results are returned (6, 8, 9). The methods of evaluation, including standardised collections and the term-weighting and ranking that underlie retrieval, are treated comprehensively in the references (2, 4, 5, 10).

RESEARCH GAP

Although precision and recall are foundational, a concise demonstration that shows how they trade off as a system returns more or fewer items, traced by the precision-recall curve, remains useful as a clear illustration. There is therefore value in a study that displays the trade-off and relates it to the evaluation and tuning of search systems.

OBJECTIVES

The objectives of this study are:

- To examine precision and recall together as a retrieval system returns more or fewer items.
- To show how precision and recall trade off against each other.
- To trace the trade-off as a precision-recall curve.
- To relate the trade-off to choosing an operating point and evaluating search systems.

METHODOLOGY

Precision, Recall, and the Trade-off

Precision, the fraction of the returned items that are relevant, and recall, the fraction of the relevant items that are returned, were examined together as a retrieval system was made to return more or fewer items, by varying how readily it judged an item relevant enough to return (3, 6). The elements are summarised in Table 1.

Table 1: Elements of retrieval evaluation

Element	Description
Relevant items	Items that match what the user wants
Returned items	Items the system returns for the query
Precision	Fraction of returned items that are relevant
Recall	Fraction of relevant items that are returned

Tracing the Curve

As the system was made to return more items, precision and recall were tracked, and precision plotted against recall to trace the precision-recall curve (7). The trade-off was related to choosing an operating point, returning more or fewer items, according to whether precision or recall mattered more for the task (6).

RESULTS AND FINDINGS

Precision and recall traded off against each other, as shown in Figure 1, which plots precision against recall as the system returned more or fewer items. When the system returned only the few items it was most confident in, those were mostly relevant, giving high precision, but many relevant items were left out, giving low recall (3).



Figure 1: Precision–recall trade-off in information retrieval

When the system returned many items so as to catch the relevant ones, recall rose, but more irrelevant items were included among them, so precision fell; plotting precision against recall traced a curve that fell as recall rose, showing that high precision and high recall could not generally be achieved at once (7). Which to favour depended on the task, a single operating point on the curve being chosen accordingly. The findings are summarised in Table 2.

Table 2: Precision and recall as more items are returned

System returns	Precision	Recall
Few (most confident)	High	Low
Many (to catch all)	Lower	Higher

System returns	Precision	Recall
As more returned	Falls	Rises
Both high at once	Generally not possible	—

DISCUSSION

The results show that precision and recall trade off in information retrieval, and the trade-off follows from how a system decides what to return. A retrieval system can be made cautious, returning only items it is confident are relevant, or generous, returning anything that might be relevant. When cautious, almost everything it returns is relevant, so precision is high, but it withholds many genuinely relevant items it is less sure of, so recall is low. When generous, it captures most of the relevant items, so recall is high, but it also sweeps in many irrelevant ones, so precision falls. Returning more items thus raises recall and lowers precision, and returning fewer does the reverse, which is the trade-off the precision-recall curve traces (3, 7).

This tension is why neither measure alone suffices to judge a system, a point central to retrieval evaluation. A system can achieve perfect precision by returning a single item it is sure of, or perfect recall by returning the entire collection, but neither is a good system; the first misses almost everything, the second returns mostly rubbish. Reporting only one measure can therefore flatter a system that is poor on the other, so the two must be considered together, which is why retrieval is evaluated by both, often combined into a single measure such as the F-measure that rewards a system only when both are reasonably high (6, 8).

Which point on the trade-off to choose depends on the task, and this is the main practical lesson. Where it matters most that what is returned is relevant, and the user will look at only a few results, as in web search, precision at the top is favoured, and the system returns cautiously; where it matters most that nothing relevant is missed, as in a legal or medical search where an overlooked item could be costly, recall is favoured, and the system returns generously, accepting more irrelevant results to be sure of catching the relevant ones (6). The system is tuned to the operating point on the precision-recall curve that suits its purpose, and improving the system as a whole means pushing the whole curve outward, raising precision and recall

together, which better ranking and term weighting aim to do (4, 9). The practical conclusion is that search systems are evaluated and tuned by considering precision and recall together and choosing the operating point suited to the task, the main caveats being that relevance is itself a matter of judgement and that the figure illustrates the trade-off rather than the performance of a specific system.

LIMITATIONS

The following limitations apply to this study:

- Relevance is itself a matter of judgement and may vary between users and tasks.
- The figure illustrates the precision-recall trade-off rather than the performance of a specific system.
- Precision and recall treat all returned items alike; ranked measures account for their order.
- Evaluation requires knowing which items are truly relevant, which is not always available.

FUTURE SCOPE

Future work could extend the present study as follows:

- Measurement of precision-recall curves for real retrieval systems and queries.
- Use of ranked-relevance measures that account for the order of results.
- Combination of precision and recall through the F-measure and its weighting.
- Study of how the operating point should be chosen for different applications.
- Improvement of ranking and term weighting to raise precision and recall together.

CONCLUSION

The evaluation of information retrieval through precision and recall and the trade-off between them were examined by examining the two together as a system returned more or fewer items. Precision and recall traded off: returning few items gave high precision but low recall, returning many gave high recall but low precision, and plotting precision against recall traced a curve falling as recall rose, so that both could not generally be high at once. The study shows that precision and recall trade off in information retrieval, so that neither alone judges a system and search systems are evaluated and tuned by considering both together and choosing the operating point, favouring precision or recall, suited to the task, with the caveats that relevance

is itself a matter of judgement and that the figure illustrates the trade-off rather than the performance of a specific system.

REFERENCES

1. Cleverdon, C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6), 173–194.
2. Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.
3. Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Butterworths.
4. Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.
5. Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search* (2nd ed.). Addison-Wesley.
6. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
7. Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1), 12–19.
8. Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446.
9. Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the International Conference on Machine Learning (ICML)*, 233–240.
10. Voorhees, E. M. (1999). The TREC-8 question answering track report. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, 77–82.

***Author for Correspondence**

Ninad R. Vaidya

E-mail: ninad_vaidya26@rediffmail.com

¹Associate Professor, Department of Computer Science and Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology

²Assistant Professor, Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology

³Research Scholar, Department of Computer Science and Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology

⁴Research Scholar, Department of Data Science, Shri Guru Gobind Singhji Institute of Engineering and Technology

Received Date: May 22, 2026

Accepted Date: June 05, 2026

Published Date: June 12, 2026

Citation: Ninad R. Vaidya, Sharvari K. Deodhar, Tanmay V. Bhide, Madhura M. Soman. Evaluating Information Retrieval: Precision, Recall, and The Trade-off Between Them. Journal of Research in Information Technology and Its Applications. 2026; 11(1): 19-27p.