The Relationship between Recall and Precision

Michael Buckland* and Fredric Gey
School of Library and Information Studies, University of California, Berkeley, Berkeley, CA 94720

Empirical studies of retrieval performance have shown a tendency for Precision to decline as Recall increases. This article examines the nature of the relationship between Precision and Recall. The relationships between Recall and the number of documents retrieved, between Precision and the number of documents retrieved, and between Precision and Recall are described in the context of different assumptions about retrieval performance. It is demonstrated that a tradeoff between Recall and Precision is unavoidable whenever retrieval performance is consistently better than retrieval at random. More generally, for the Precision–Recall trade-off to be avoided as the total number of documents retrieved increases, retrieval performance must be equal to or better than overall retrieval performance up to that point. Examination of the mathematical relationship between Precision and Recall shows that a quadratic Recall curve can resemble empirical Recall–Precision behavior if transformed into a tangent parabola. With very large databases and/or systems with limited retrieval capabilities there can be advantages to retrieval in two stages: Initial retrieval emphasizing high Recall, followed by more detailed searching of the initially retrieved set, can be used to improve both Recall and Precision simultaneously. Even so, a tradeoff between Precision and Recall remains.

Introduction

When tests of the performance of retrieval systems were first developed, an empirical tendency was noticed for Recall (completeness of retrieval) and Precision (purity of retrieval) to be inversely related: One might have high Recall or high Precision, but not, it seemed, both at the same time. There appeared to be a trade-off, even though having high values of both at the same time would be preferable. This article examines the nature of the relationship between Precision and Recall and explains why a trade-off between Precision and Recall is unavoidable under certain conditions. The strategy of performing retrieval in two stages is also considered because it offers a possibility of improving Recall and Precision simultaneously. Examination of the relationship between Recall and Precision also leads to the identification of a particular class of curves (tangent parabolae) that resemble typical empirically found retrieval results.

The relationship between Recall and Precision was described by Cleverdon (1972), and later studied by Heine (1973), Bookstein (1974), Robertson (1975), and more recently by Gordon and Kochen (1989). In modeling Recall and Precision, Heine, Robertson, and Gordon and Kochen have assumed continuous functions for Precision and Recall, while Bookstein described Recall and Precision in terms of a two-Poisson discrete model. Robertson (1975) discussed the implications of thinking about Recall and Precision as functions of other independent variables.

Definitions

The customary definitions of Precision and Recall are based on the following traditional (albeit questionable) assumptions:

(a) Binary relevance judgments, namely that every retrievable item is recognizably "Relevant" or recognizably "Not relevant."

Hence, for every search result all retrievable items fall into one and only one of four cells in a matrix defined by the two distinctions: (i) Retrieved or Not Retrieved; and (ii) Relevant or Not Relevant (See Table 1). *Generality* is the proportion of documents in the entire document collection that are judged to be relevant.

For any given retrieved set, *Recall* is the number of retrieved Relevant items as a proportion of all Relevant items, i.e., $N_{ret \cap rel}/N_{rel}$. Recall is, therefore, a measure of effectiveness in retrieving (or selecting) performance and can be viewed as a measure of effectiveness in including relevant items in the retrieved set. One-hundred percent Recall can always be achieved by retrieving (examining) the entire database, but this defeats the purpose of a retrieval system. High Recall is not always needed, since people commonly do not want all relevant items, often preferring only one or a few relevant items. However, the ability to achieve high Recall (100% or close to it) efficiently is clearly desirable.

For any given retrieved set, *Precision* is the number of retrieved Relevant items as a proportion of the number of

*To whom all correspondence should be addressed.

Received November 1, 1991; revised June 1, 1992 and December 11, 1992; accepted June 9, 1993.

TABLE 1. Retrieval matrix.

	Relevant	Not Relevant	TOTAL
Retrieved	$N_{ret \cap rel}$	$N_{ret \cap \overline{rel}}$	N_{ret}
Not Retrieved	$N_{\overline{ret} \cap rel}$	$N_{\overline{ret} \cap \overline{rel}}$	$N_{\overline{ret}}$
TOTAL	N_{rel}	$N_{\overline{rel}}$	N_{tot}

retrieved items, i.e., $N_{ret \cap rel}/N_{ret}$. Precision is, therefore, a measure of purity in retrieval performance, a measure of effectiveness in excluding nonrelevant items from the retrieved set. High Precision—like high Recall—is desirable. The ideal would be to achieve 100% on both at the same time.

(b) Retrieval is seen as an expansive process: Searches are (or can be) expanded to retrieve more and more items, thereby increasing Recall.

Alternative assumptions are possible: One might use degrees of relevance, and one might prefer to iterate searches, using relevance feedback. However, this study examines Recall and Precision in traditional terms. For an introduction to this area see Lancaster (1979). For a recent overview of retrieval evaluation see Harman (1992).

Recall

In order to understand the relationships between Recall and Precision, it is important and useful to understand the *theoretical* behavior of these measures under various assumptions. If we know how well (or how poorly) Recall and Precision can behave in theory, we can evaluate actual retrieval systems in practice in the light of these limits. In Figure 1, we display four theoretical cases of Recall performance. For ease of illustration, we assume that 100 items are relevant in a retrievable set of 1000, a relatively rich case of one-tenth of the retrievable items being relevant. The fraction of the collection which is relevant, $G = N_{rel}/N_{tot}$, is the parameter, *Generality,* the probability that a document in the collection is relevant. In our illustration, we have used the arbitrary case $G = 0.1$.

For the ideal case of *perfect retrieval* we have the characteristic that all relevant documents are retrieved *before* the first nonrelevant document. In the figure, the first 100 documents retrieved would all be relevant, leading to a steeply rising straight line hitting a maximum (of 100 relevant documents, 1.0 Recall) when the first 10% of documents have been retrieved. Since, at that point, all relevant documents have already been retrieved, Recall must remain 1.0 while any or all remaining documents are retrieved. In all cases, the slope of perfect retrieval (with respect to proportion of documents retrieved t) is $1/G$, the inverse of the Generality, until all relevant documents have been obtained.

Choosing documents (or document surrogates) randomly from a database would mean that the next item retrieved is as likely as any other document to be relevant. In other words, the next 10 retrieved items would, in our example, on average, contain one relevant document (10%). Thus, for *random retrieval,* plotting successive values of Recall yields a straight line from the origin to the upper right corner of the figure.

On a graph of the Recall curve using cardinal numbers of Recall (e.g., 0–100) and for Number of documents retrieved (e.g., 0–1000) the slope of the Recall curve for random retrieval is determined by the Generality (if 100 out of 1000 documents are relevant, the generality is $100/1000 = 0.1$, and the slope of the Recall curve for random retrieval is 0.1). If the axis and abscissa have both been normalized to the interval $[0, 1]$, then the angle of the random retrieval line is invariably 45°.

One can imagine another ideal limit, that of *perverse retrieval,* in which all of the nonrelevant items are retrieved before the first relevant one. In this example, until 90% of the documents have been retrieved, the next document will always be nonrelevant, and the numerator of the Recall equation is zero, and hence, Recall remains zero until no more irrelevant documents remain, at which time the system has no choice but to retrieve relevant items. At this point, system behavior becomes like that of perfect retrieval, Recall rises straight and rapidly to the upper right position where all documents have been retrieved, and Recall reaches 1.0. Again, the slope of this last part of the curve is the inverse of the Generality.

In all cases, therefore, Recall performance is bounded by the parallelogram defined by perfect and perverse retrieval. Note also that since Recall is a cumulative process—retrieved documents are never unretrieved—the Recall curve must start at the origin, must end in the top right corner, and can only move to the right or diagonally upward toward the right. Further, to be worthwhile,

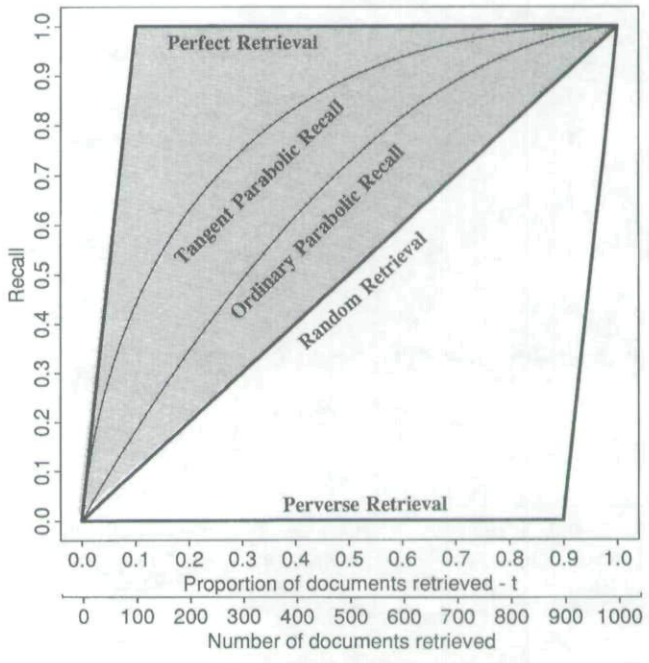

FIG. 1. Recall under various retrieval assumptions.

a retrieval system should perform better than random retrieval, but, realistically, is unlikely to perform perfectly. Therefore, we should expect the Recall curve of any real information system to lie in the shaded triangle that denotes better-than-random retrieval but less-than-perfect Recall performance. Figure 1 contains two examples of curves reflecting what we call "realistic retrieval."

Mathematics of the Recall Curve

So far, we have concentrated on verbal and graphical descriptions of Recall. We can also consider what shape we should expect realistic Recall curves to be. The formal constraints are that such curves must go from the origin to the top right corner, must remain within the region defined as feasible (the shaded region in Fig. 1), and we should expect retrieval performance to be best at the beginning of the search and to deteriorate as the search is expanded.

Parabolic Recall

Gordon and Kochen (1989) discussed various forms of realistic Recall curves, notably parabolic and logarithmic curves. The mathematics is simplified if we set (as Gordon and Kochen did) $t = x/N_{tot}$ to transform the horizontal axis to the range 0 and 1. The proportion of total documents retrieved then becomes t.

In particular, *parabolic Recall* would postulate that the Recall curve would be a quadratic equation[1]

$$R^*(x) = R^*(N_{tot}t) = R(t) = a_0 + a_1t + a_2t^2$$

constrained to lie within the triangle bounded by perfect retrieval and random retrieval. Since Recall at the origin is zero (no documents have been retrieved),

$$R(t = 0) = a_0 = 0$$

Since we always have Recall ≤ 1, and since Recall must be 1 when all documents have been retrieved,

$$R(t = 1) = 1 = a_1 + a_2$$

or

$$a_2 = 1 - a_1$$

Finally, because of the constraint that the slope is between 1 and $N_{tot}/N_{rel} = 1/G$, we find:

$$1 \leq R'(t = 0) = 2a_2t + a_1 = a_1 \leq \frac{1}{G}$$

This yields a family of parabolae for all values of a_1 between 1 and $1/G$. An "ordinary parabolic Recall," plotted in Figure 1, which satisfies these conditions, as well as the constraint that $R(t) \leq 1$, is

$$R(t) = 1.9t - 0.9t^2$$

Tangent Parabolic Recall

Not all parabolae satisfying the criteria $1 \leq a_1 \leq 1/G$ and $a_2 = 1 - a_1$ also satisfy the constraint that $R(t) \leq 1$. For example, one might wish to find a parabola which is tangent to perfect Recall at the origin. This criterion, considered by Gordon and Kochen, means that "the first document retrieved is relevant." The unique parabola which also satisfies this criteria is

$$R(t) = 10t - 9t^2$$

However, if we plot (Fig. 2) this "simple tangent parabola" and examine its values, we find it violates the constraint that Recall must (by definition) remain less than or equal to 1.0. The parabola rises rapidly and attains the value 1.0 after less than 13% of the documents have been retrieved.[2] The parabola then rises to a maximum value of 2.71 before descending to 1.0 when all documents have been retrieved. Thus, it *seems* that if one accept the assumption that the first retrieved document must be relevant (i.e., that the parabola is tangent to perfect retrieval at the origin), then one must reject quadratic Recall behavior as unrealistic.

It there a way out of this dilemma? To explore this we can turn to analytic geometry and try to see if a nonlinear transformation can supply us with a parabolic curve which would have $0 \leq R(t) \leq 1.0$, its first derivative $R'(t) > 0$ (i.e., Recall is monotonically increasing), and its first derivative at the origin equal to the slope of perfect retrieval, i.e., $R'(0) = 1/G$. One such transformation is

$$t^{\#} = (1 - 2G)\, t^2 + 2G\, t$$
$$R^{\#}(t) = -t^2 + 2\, t$$

which maps pairs (t, R) into pairs $(t^{\#}, R^{\#})$. Since the transformation is quadratic, the resulting transformed equation is also quadratic. We call this case *tangent parabolic Recall,* and it is plotted in Figures 1 and 2.

This parabola is but one of a family of tangent quadratic curves which can be drawn to pass through the origin and peak incident to a point along the perfect retrieval curve between $t = G$ and $t = 1.0$, yielding a piecewise Recall curve. This satisfies the condition that 100% Recall might be attained before all documents have been retrieved.

[1] In order to minimize confusion about distinctions of scale, from this point forward, scales will be displayed and slope computations made in terms of t—proportion of total documents retrieved. If slopes were computed with respect to x—absolute number of documents retrieved, they would be different. Finally, all reference to the absolute scales will be dropped from the equations. However, the equations used for Recall and Precision are correct, even though expressed in the proportional scale.

[2] It was similar behavior (although of a different family of parabolae) which led Gordon and Kochen to reject a parabolic model of recall. One can, of course, utilize a piecewise definition (as Gordon and Kochen have), modeling with the parabola where $R(t) \leq 1$ and $R(t) = 1$ for the remainder of the recall curve.

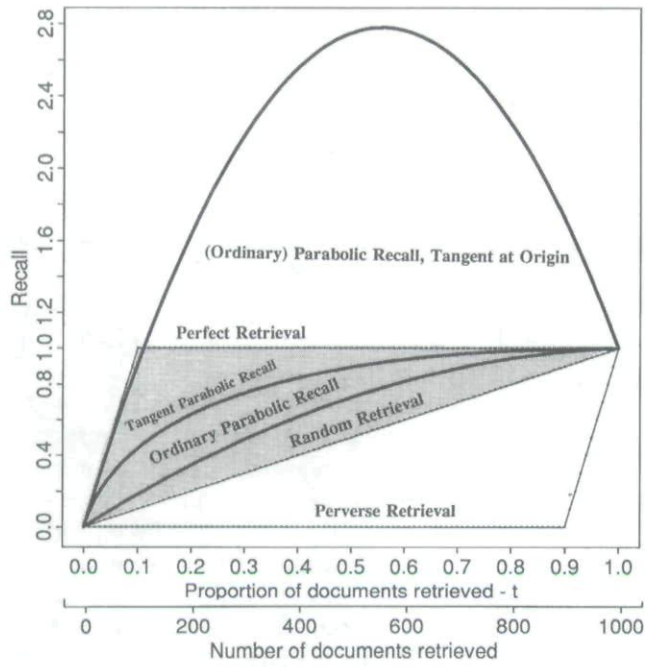

FIG. 2. Parabolic Recall under various assumptions.

Precision

Corresponding values for Precision in the cases of Random, Perfect, and Perverse retrieval are plotted in Figure 3.

With *random retrieval,* the probability of the next retrieved item being relevant will reflect the overall proportion of relevant items in the retrievable set, the Generality. Hence, in this hypothetical case where 100 out of 1000 items are relevant, Precision will be tend to be a flat horizontal line at 10%, however, many items are retrieved.

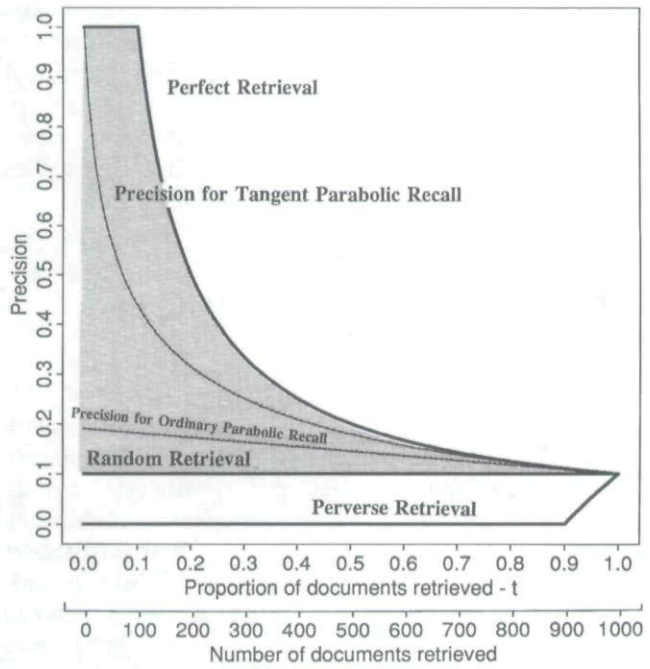

FIG. 3. Precision under various retrieval assumptions.

With *perfect retrieval,* all retrieved items within the first 100 items retrieved, will, by definition, be relevant. Hence, within that range, Precision will be always be 1.0, a horizontal straight line. If retrieval were to be continued after these 100 items, than all additional retrieved items would necessarily be nonrelevant. After the first 100 items, while Recall remains at 1.0, Precision declines hyperbolically to the limit of $G = 0.1$ when the entire retrievable set has been retrieved.

All curves reflect the fundamental relationship

$$P^*(x) = \frac{R^*(x)N_{rel}}{x}$$

or

$$P(t) = \frac{R(t)G}{t}$$

which is how the hyperbolic descent occurs in perfect Precision.

Correspondingly, with *perverse retrieval,* in the range 1–900 retrieved, all will be nonrelevant and Precision will remain at zero, a horizontal straight line from the origin. After 900 items, the remainder are all relevant and Precision climbs monotonically to the limit of 10% when every item has been retrieved.

For our example of (ordinary) parabolic retrieval, the relationship between Precision and Recall is

$$P(t) = \frac{(1.9t - 0.9t^2)G}{t} = G(1.9 - 0.9t)$$

i.e., Precision for parabolic Recall must be a straight line. More generally, for any ordinary polynomial Recall function of t, Precision will be a lower order polynomial function of t. Since, for tangent parabolic Recall, both the horizontal and vertical axes have undergone a nonlinear transformation, Precision will remain a nonlinear function of number of documents retrieved. Our impression is that empirical data tend to be nonlinear and, if so, tangent parabolic curves offer a better working model than ordinary parabolic curves.

Again, for all possible retrieval results, Precision must necessarily be within the region bounded by perfect Precision and perverse Precision. Further, for all retrieval systems whose performance is better than Random retrieval, Precision must be within the shaded region bounded by perfect Precision and random Precision. Since all performance lines for better-than-random performance must start from the vertical axis above 10%, and must converge on the limit of 10% when all 1000 items have been retrieved, Precision, for all "realistic" retrieval systems, must tend to be a downward sloping curve.

Precision versus Recall

We noted the empirical finding that Precision and Recall appear, in practice, to be inversely related: improvement in either tends to be associated with poorer performance of

the other. This is unsatisfying and we need to ask whether it is possible to evade this inconvenient pattern and how.

Since, in our hypothetical examples, Precision and Recall have both been plotted against a common scale, the number (or proportion) of documents retrieved (Figs. 1 and 3), it is possible to plot Precision and Recall against each other. Figure 4 is a graph of Precision versus Recall showing a replotting of the examples given in Figures 1 and 3: Perfect retrieval, Random retrieval, Perverse retrieval, and the two examples of Realistic retrieval. Such graphs can be drawn for any example of retrieval performance we care to imagine.

A Practical Example: CACM Query 25

One question which might be raised is whether the models introduced mirror actual retrieval. Figure 5 contains three plots of actual retrieval behavior versus perfect retrieval for query number 25 from the well-known CACM test document collection, described by Fox (1983) and used in Buckley and Salton (1988) in term-weighting experiments. The retrieval method utilizes the well-known *Cosine* similarity measure from the vector space model of information retrieval. As can be seen, both Recall and Precision behavior for the query follow the general pattern described by our "tangent parabolic recall" model; Fig. 5c, which plots Precision versus Recall, shows jagged changes of direction (slope) because of the discrete nature of actual retrieval. In Figure 5d, the recall performances for all 52 CACM queries have been plotted and superimposed on the same graph. This shows that they all fall within boundaries we have defined as "realistic," i.e., less-than-perfect retrieval and better-than-random retrieval. Since all recall points fall within the boundaries of perfect and random retrieval, we can, for this collection, conclude that the vector space retrieval model yields "realistic" retrieval results as defined above.

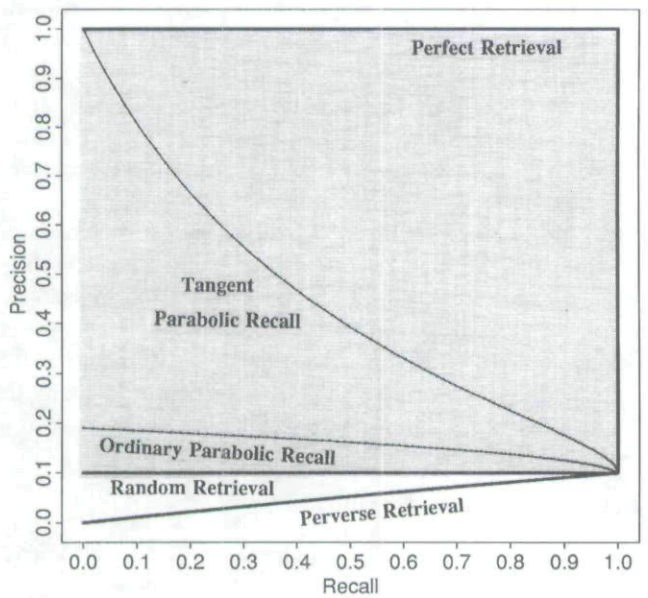

FIG. 4. Recall–Precision under various retrieval assumptions.

Consistency in Realistic Retrieval

Consistently realistic retrieval performance, defined as retrieval performance consistently better at each point of the recall curve than recourse to random selection, forms a convex Recall curve always above the Random retrieval line. In all such cases, when Precision versus Recall is plotted, it forms a downward sloping curve and a trade-off between Precision and Recall is entailed, as Gordon and Kochen noted in 1989. This is the trade-off between Precision and Recall that has been found empirically.

More generally we can relax the assumption that retrieval performance is consistently realistic. What if, for example, retrieval performance started well, deteriorated to worse than random, then improved? During the latter improvement, might Precision and Recall improve together? Graphs can be plotted of Precision versus Recall for any imaginable retrieval performance. Examination of such graphs for various hypothetical cases shows that the expected trade-off does not occur under some circumstances. The condition to be met, for the Precision–Recall trade-off to be avoided, is that, *as the total number of documents retrieved increases, retrieval performance must be equal to or better than overall retrieval performance up to that point.* This condition is always met with perverse retrieval and with any retrieval performance that had been consistently worse than random retrieval. That this should be the condition for avoiding a Precision–Recall trade-off is also explicable in terms of the basic mathematics involved: Recall is, by definition, a cumulative measure of the search performance up to any given point. Precision, likewise, is ordinarily defined cumulatively, as the proportion of documents that are relevant among the total number of documents retrieved thus far. This proportion will not decline (i.e., constitute a trade-off) if, as the number of retrieved documents increases, stable (or improving) retrieval performance were to maintain (or increase) the ratio of relevant to nonrelevant documents. (An actual example of temporary remission of the trade-off is indicated by arrows in Fig. 5a and c). Neither better-than-random nor an improvement in retrieval performance constitutes a sufficient condition for the trade-off to be avoided.

Two-Stage Retrieval—High Recall Followed by High Precision

The flexibility permitted by increasingly affordable information technology suggests an approach to mitigating or avoiding the trade-off between Precision and Recall. We now examine a two-stage retrieval strategy whereby two searches are performed: An initial search emphasizing high Recall, then a second search of the retrieved subset seeking to improve Precision within that subset, as proposed, for example, by Porter (1983), and being explored by Buckland et al. (1992).

To repeat the same search on the initially retrieved subset would be pointless, since the same results would be

expected. A second search on the subset would only make sense if it were different. A different search is likely to be both feasible and desirable for at least two reasons:

(a) *Technical considerations.* Experimental retrieval techniques, such as vector space matching or relevance feedback are simply not yet available on large existing bibliographic services, which are not easily modified. The newer retrieval techniques can be provided on workstations, at least experimentally or for occasional use on an exception basis. Next generation techniques have generally been tested on databases of modest size. They may not scale up well to larger files, but their use on downloaded subsets should not be difficult.

(b) *Predictive power.* We agree with the view of Belkin and Croft (1987) that retrieval is essentially a matter of selection by matching, whether by full or partial matching, of a search statement with representations of items. This is consistent with Wilson's view of bibliographic searching, whether manual or online, as being concerned with "fitting the description" (Wilson 1968). The searcher has some more or less well-formulated notion in mind of what is wanted and seeks to find one, a few, or all available items that match ("fit") that notion more or less well. There is a description implicit in retrieval, which can be made explicit only imperfectly.

Existing online retrieval systems vary in their ability to express and handle complex descriptions. Even where considerable expressive powers are provided, studies of the use of online retrieval systems consistently reveal that little of the system's capability for handling complex descriptions is used. Even simple Boolean statements get surprisingly little use. A searcher actually interested in English-language descriptions of industrial activity in Dresden, Germany, in the 1930s is quite likely to start with a drastically simplified search statement such as FIND SUBJECT DRESDEN.

In the case of online library catalogs, even systems with relatively expressive search capabilities, not all of the features of the retrievable items can be searched. Sometimes one can search (or limit a search) by language, for example, but rarely, if ever, by country of publication or a variety of other fields in standard bibliographic records.

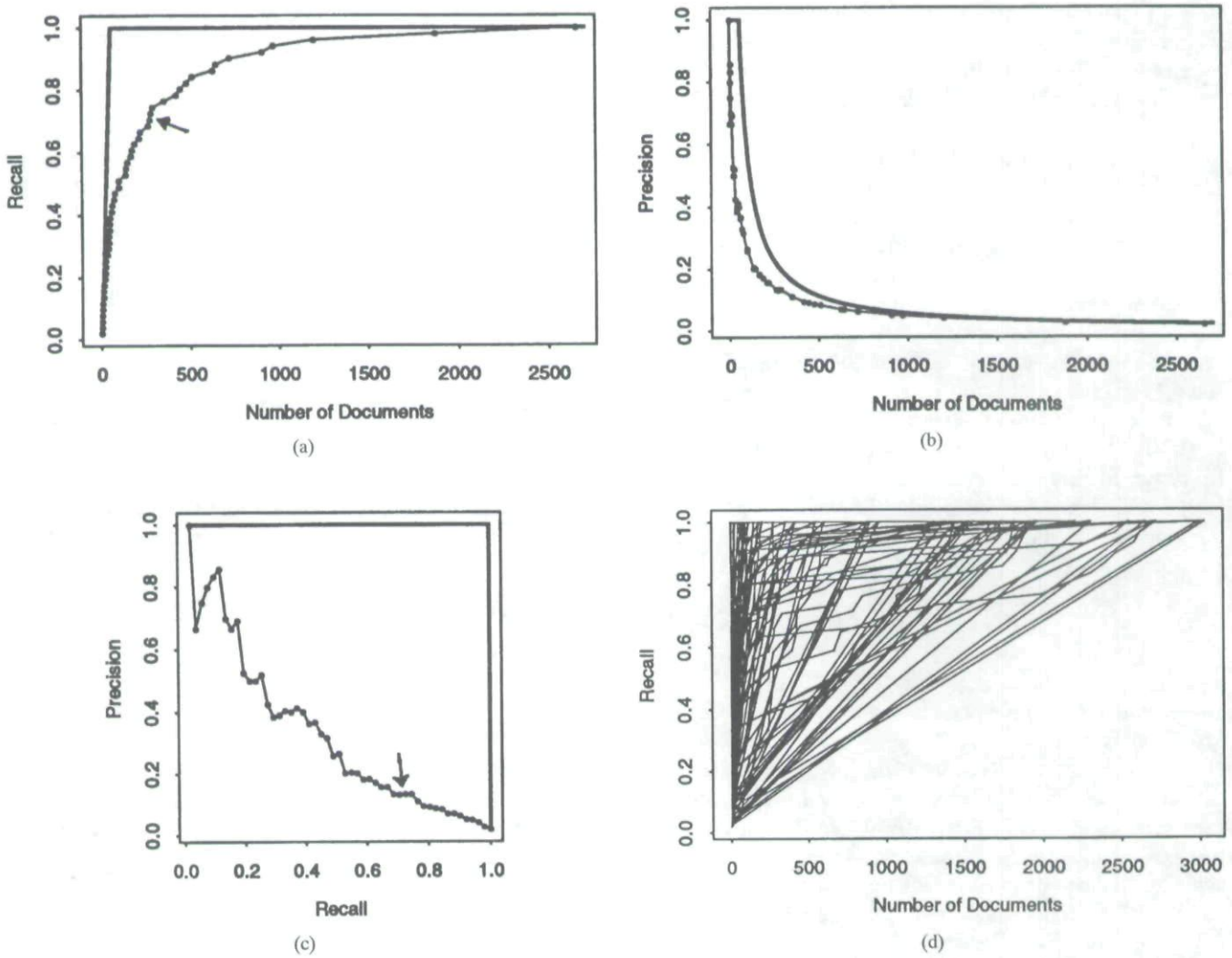

FIG. 5. (a) CACM query 25: Actual Recall versus Perfect Recall. (b) CACM query 25: Actual Precision versus Perfect Precision. (c) CACM query 25: Recall–Precision versus Perfect Recall–Precision. (d) CACM collection, all queries: Actual Recall.

Since searching is a predictive activity, we can note that there are two approaches to achieving better predictions. One is to use *stronger predictors,* i.e., search statements that are more reliable at retrieving what is wanted. Another is to use *additional clues* that cumulatively improve search success. Typically, there is considerable scope for the use of additional clues since in existing bibliographic systems: (i) not all the descriptive data can be searched; (ii) complex search capability (when provided) is little used; and (iii) actual search statements tend to simplify drastically the description that is implicit or explicit in the users mind—even though the fitting of that description is the purpose of retrieval. The downloading, indexing, and searching of retrieved subsets provides, at least in principle, the ability to use a more complete match to what the user wants: preferably in English, published around 1930, currently conveniently available, and so on in accordance with the invoked or assumed preferences of the user.

What difference might two-stage retrieval make? In Figure 6, we examine an arbitrary, high-Recall retrieval result, S, in a hypothetical realistic (i.e., better-than-random but less-than-perfect) retrieval system, shown as Recall curve D. Within this subset, the line for Perfect retrieval (B) would remain the same, but the line for Random retrieval within the selected subset would necessarily be steeper (A′) because it is limited to this selected subset.

If, as suggested above, additional search techniques and/or use of a broader range of clues were to result in more effective retrieval (within the subset) than had been the case in the original search, then the result would be better discrimination between relevant and nonrelevant items (or a better ranking by degree of relevance) as shown by line D′, running above and/or to the left of D. Within the subset, since retrieval performance remains equal to or better than random (now line A′), the trade-off between Precision and Recall will necessarily remain. However, since Precision is the ratio of retrieved relevant items to all retrieved items, to the extent to which line D′ is to the left of line D, Precision is improved without loss of Recall. Likewise, to the extent to which line D′ is above line D, Recall has been improved without loss of Precision. Using the fundamental relationship between Precision and Recall introduced above.

$$P(x) = \frac{R(x)N_{rel}}{x}$$

we find, for example, at points T and T′ that $x = n_T < x = n_{T'}$ and $R = r_T = r_{T'}$, and hence, $P(n_T) > P(n_{T'})$. Moreover, since for points T and T″, $x = n_T = n_{T''}$ and $R = r_T > r_{T''}$, then $P(n_T) > P(n_{T''})$. More generally, we can conclude that if secondary retrieval from a subset can result in a shift in the retrieval curve from line D to line D′—in the direction of the arrow—then the effect is to achieve the desired simultaneous improvement in Precision and Recall. Nevertheless, a trade-off between Precision and Recall remains.

An additional comment can be made on the relationship between the secondary search (line D′ in Fig. 6) and the original search (line D). Because the secondary search is, we have assumed, a better ordering of the items within the subset S, the *marginal* retrieval performance—the probability that the next item retrieved will be relevant—of the secondary search (line D′) is superior to the marginal retrieval performance of the original search (line D) near the origin, but inferior as the subset becomes fully retrieved (as shown by the flattening of line D′ relative to D as they both approach point S). It does not follow that the secondary search ceases to be preferable to the original search at this stage, because the *cumulative* retrieval performance of the secondary search dominates: it is always superior to the original search. Instead, the decreasing marginal performance of the secondary search indicates an increasing likelihood that, if additional relevant items are wanted, the optimal point for conducting a fresh search on a larger or supplementary subset (or on the entire database) has been reached.

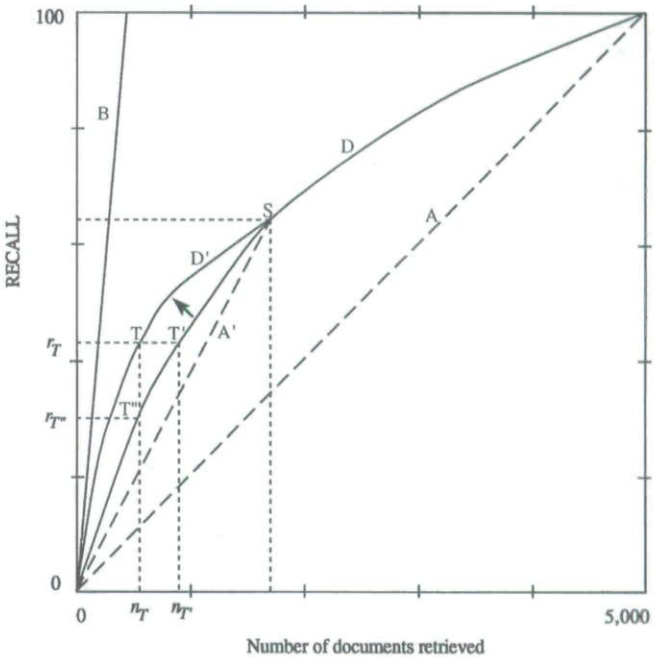

FIG. 6. Recall for two-stage retrieval.

Summary

Recall and Precision and, in particular, Recall–Precision plots, have been used for many years to characterize document retrieval performance. In this article, the relationship between Recall and Precision has been approached conceptually to delineate the theoretical limits to retrieval performance in terms of "benchmark" retrieval behaviors. It has been shown that there is a definable region for all feasible retrieval results. For all cases of consistently better-than-random retrieval, Recall curves tend to follow an increasing curve rising from the origin, and a trade-off between Precision and Recall is inherent, not just an inconvenient empirical finding. More generally, a trade-off between Precision and Recall is entailed unless, as the

total number of documents retrieved increases, retrieval performance is equal to or better than overall retrieval performance thus far.

There is a fundamental relationship between Precision and Recall which, for a given model of Recall, constrains the behavior of Precision. In particular, if Recall is modeled by a polynomial function of proportion of documents retrieved, then Precision is modeled by a lower order polynomial function of the same variable.

The quadratic model of Recall has been examined and refined. We have demonstrated a simple geometric transformation which can produce quadratic Recall and satisfies tangency to perfect retrieval at the origin and yields reasonable looking Recall–Precision tradeoffs.

Two-stage, or, more generally, multistage retrieval procedures, whereby a retrieved set is used for a subsequent, more detailed search, is likely to achieve the goal of improving *both* Precision *and* Recall simultaneously even though the trade-off between them cannot be avoided.

Acknowledgments

We would like to thank Terry Ligocki of the University of California, Berkeley, Mathematics Ph.D. Program for explaining and illustrating the geometry of the rotated parabola. Fred Cisin, a master's student at the School of Library and Information Studies, suggested the case of perverse retrieval. The reviewers were very helpful in pointing out where the mathematics could be improved and the ideas clarified. This work was supported in part by U.S. Department of Education HEA IID Grant R197D00017 for "Prototype for an Adaptive Library Catalog" and by the School of Library and Information Studies, University of California, Berkeley.

References

Belkin, N. J., & Croft, W. B. (1987). Retrieval techniques. *Annual Review of Information Science and Technology, 23,* 108–145.

Bookstein, A. (1974). The anomalous behavior of precision in the Swets model, and its resolution. *Journal of Documentation, 21,* 374–380.

Buckland, M. K., Butler, M. H., Norgard, B. A., & Plaunt, C. (1992). OASIS: A front-end for prototyping catalog enhancements. *Library Hi Tech, 10,* 7–22.

Cleverdon, C. W. (1972). On the inverse relationship of Recall and Precision. *Journal of Documentation, 28,* 195–201.

Fox, E. (1983). Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts. *Computer Science Technical Report 83-561.* Ithaca, NY: Cornell University.

Gordon, M., & Kochen, M. (1989). Recall–Precision trade-off: A derivation. *Journal of the American Society for Information Science, 40,* 145–151.

Harman, D. (Ed.) (1992). Special issue: Evaluation issues in information retrieval. *Information Processing and Management, 28,* 439–528.

Heine, M. H. (1973). The inverse relationship of Precision and Recall in terms of the Swets model. *Journal of Documentation, 20,* 81–84.

Lancaster, F. W. (1979). *Information retrieval systems: Characteristics, testing and evaluation* (2nd ed.). New York: Wiley.

Lancaster, F. W. (Ed.) (1978). Precision and Recall. In *Encyclopedia of Library and Information Science* (vol. 23, pp. 170–180). New York: Marcel Dekker.

Porter, M. F. (1983). Information Retrieval at the Sedgwick Museum, *Information Technology: Research and Development, 2,* 169–186.

Robertson, S. E. (1975). Explicit and implicit variables in information retrieval (IR) systems. *Journal of the American Society for Information Science, 26,* 214–222.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24,* 513–523.

Wilson, P. G. (1968). *Two kinds of power: An essay on bibliographic control.* Berkeley: University of California Press.