Information Retrieval Performance Measurement Using Extrapolated Precision

William C. Dimm
Hot Neuron LLC
bdimm@hotneuron.com

May 13, 2015

Abstract

Performance measures like the F_1-score make strong assumptions about the trade-off between recall and precision that are not a good fit for some contexts like e-discovery. This paper advocates comparing performance at different recall levels, when necessary, by using a novel method for extrapolating a single precision-recall point to a different level of recall. With this approach, the constant-performance contours are a parameterized family of reference precision-recall curves.

1 Problems with F-Scores and Similar Measures

When evaluating information retrieval (IR) systems where both high recall, R, and high precision, P, are desirable, comparing precision values at the same level of recall avoids the problem of determining how much degradation of precision is acceptable to attain higher recall. Sometimes, however, it is necessary to compare results when the level of recall is different. For example, systems may generate binary relevance predictions rather than relevance scores, or systems may aim for a particular level of recall with the exact level reached only being determined during analysis that occurs after the retrieval. The goal is to make a useful quantitative measure of how one point in the precision-recall plane compares to another.

Before proposing a new approach, it is worthwhile to examine some of the shortcomings of other performance metrics to identify areas where improvement is possible. One popular metric is the F-measure or F-score [vR79]:

$$F_b(R,P) = \frac{1}{\frac{1}{b^2+1}\left(\frac{b^2}{R}+\frac{1}{P}\right)} = \frac{(b^2+1)RP}{b^2P+R} \tag{1}$$

The value for F_b will fall between R and P, with large values of the weighting parameter, b, pushing F_b closer to R. A common choice for the weighting

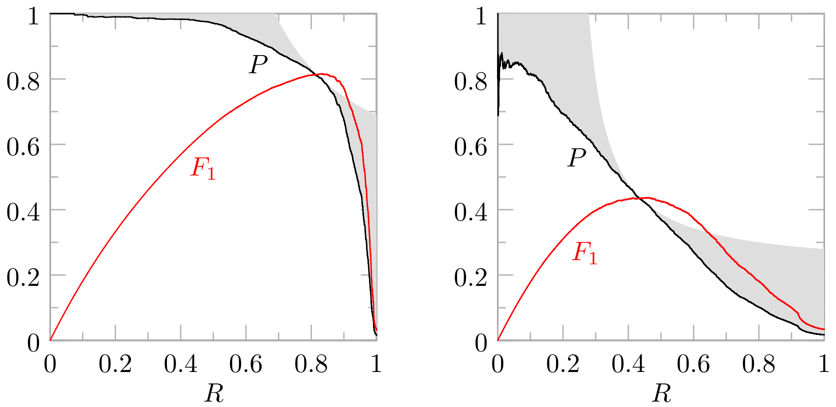

Figure 1: Precision and F_1 as a function of recall for two tasks. The shaded regions indicate points where a system with better precision at that recall would get a lower F_1 than the best F_1 that the graphed system could achieve (at a different recall).

parameter is $b = 1$. F_1 tends to be closer to the smaller of R and P, so one cannot achieve a high value for F_1 while having a very low value for R or P.

Recall and precision can have very different significance to the user of an IR system in some contexts like e-discovery. In e-discovery, the documents that are retrieved by the IR system are often reviewed by humans to make final relevance determinations and identify any privileged information that should be redacted or withheld. The number of documents retrieved, n, is related to the size of the document population, N, the prevalence (or richness) of relevant documents, ρ, and recall and precision, by:

$$n = \frac{\rho N R}{P} \tag{2}$$

where the numerator is the number of relevant documents found, and dividing by P converts it to the number of documents retrieved. High recall is required to satisfy the court, whereas Equation (2) shows that high precision is desirable to minimize the time and expense of human document review corresponding to a particular level of recall.

Although the goal is to compare systems at different levels of recall, it is an informative first step to examine the performance measure at equal recall because Equation (2) provides a way to quantify the value of the results. Figure 1 shows P and F_1 as a function of R for a system performing two IR tasks, but one could imagine the two graphs as comparing two different IR systems performing the same task. At $R = 0.75$, the system on the right would require 6.5 times as much document review (excluding training and testing) as the system on the left, based on the precision values and Equation (2). The ratio of the F_1 scores at $R = 0.75$ is only 3.6. This is a general property of F-scores—they mix P and R together in a way that dilutes the impact of differing precision. The effect is even more significant if more weight is given to the recall. For example, the ratio of F_2 values at $R = 0.75$ in this case is only 2.0. F-scores fail to accurately measure performance differences in the single situation (equal recall) where the right result is known without ambiguity.

Figure 1 shows that F_1 is maximal at $R = 0.83$ on the left and at $R = 0.46$ on the right. If F_1 is truly the right performance measure when comparing different IR systems, it should also be the right measure when comparing different points on the same precision-recall curve. In other words, it would be irrational to stop the retrieval at any R value other than the one that maximizes F_1. In the context of e-discovery, stopping retrieval at the recall level that maximizes F_1 is simply not consistent with common practice or the law. Currently, it is common to aim for recall of at least 0.75 when using supervised machine learning. Producing only 46% of the relevant documents, as F_1 would suggest for the situation on the right side of Figure 1, is likely to lead to objections from the requesting party unless the cost of that 46% is already substantial compared to the value of the case, which highlights another critical problem with F-scores. Rule 26(b)(2)(C) of the Federal Rules of Civil Procedure requires discovery to be limited if "the burden or expense of the proposed discovery outweighs its likely benefit," a principle known as proportionality. The recall level that maximizes F_1 is determined by the document population, the retrieval task, and the performance of the IR system. It is completely blind to the value of the case being litigated or any of the other factors that are considered when determining the limits imposed by proportionality. If the recall levels that people aim for in practice aren't consistent with maximizing F_1, that says quite clearly that F_1's approach to measuring the relative worth of recall and precision is not in line with the actual needs of the people using IR systems. Shifting to a different F-score that weights recall more heavily, like F_2, doesn't fix the problem—it would dictate aiming for $R = 0.59$ to maximize F_2 in the case on the right side of Figure 1, which seems more reasonable, but it would also push up the optimal recall for the left side of the figure to 0.89, and still wouldn't account for the case-specific limitations imposed by proportionality.

F-scores are problematic because they make strong assumptions about permissible trade-offs between precision and recall, and those assumptions cause trouble when they don't align with the actual information need. F-scores are intentionally small when R is very small or very large (since P will inevitably be small at very large recall), leading to a hump-shaped curve that penalizes hitting a recall level that is too far from the recall that the F-score considers to be optimal. This means a superior system may appear, based on an F-score, to be inferior if evaluated at a recall level that is not favorable as far as the F-score is concerned. Regions of the precision-recall plane where that can occur are shaded in Figure 1. The shaded regions are large in Figure 1 because the contours of constant F_1 are not shaped like precision-recall curves—the upper bound of each shaded region is a constant F_1 contour. Although we've focused on F-scores, this analysis applies to other hump-shaped (as a function of R) performance measures like the Matthews correlation coefficient [Mat75].

Another approach to comparing performance at different recall levels is to extrapolate the precision values to the same target recall. As an example of a simple extrapolation scheme, an IR system giving binary relevance predictions could be extrapolated to a lower recall level by taking a random subset of the documents predicted to be relevant, effectively treating the precision as being the same for all recall levels below the recall that was actually achieved. This is similar to the interpolation that is sometimes done with precision-recall curves [OW13]. Although this paper advocates using extrapolated precision as a performance measure, the simple extrapolation scheme is flawed. The performance measure is really just P itself, without any regard for the recall where it was achieved (see Figure 9 for an example of the method's bias). As a thought experiment, consider creating several binary IR systems by taking a single IR system that generates a relevance score and applying different relevance score cutoffs to generate binary results. The system with the highest relevance score cutoff would typically achieve the highest precision (hence, appearing to be superior) by retrieving the smallest number of relevant documents, which isn't a good fit for the goals of e-discovery.

2 Extrapolated Precision

Since proportionality, and therefore factors that are external to the document population and classification task, should dictate recall goals in e-discovery, we seek a performance measure that reflects how well the IR system is working without trying to dictate an optimal recall level for evaluation. In other words, the contours of constant performance should be shaped like typical precision-recall curves, so varying the recall by moving along a precision-recall curve having typical shape will have minimal impact on the value of the performance measure and thus will minimize the risk of drawing wrong conclusions about which system is performing best due to differences in the recall level where the measurement is made. If an IR system has a precision-recall curve with an unusual shape, it will cut across constant performance contours and should achieve higher values for the performance measure in recall regions where the precision achieved is considered anomalously good compared to the rest of the precision-recall curve. Since the performance measure aims to be fairly recall-agnostic, any recall requirements should be imposed externally, like requiring the recall to be above some level or applying a penalty factor to the performance measure if the recall doesn't lie within a specified range.

To achieve the goals outlined in the previous paragraph, we will create a parameterized set of non-overlapping (except at $R = 1$ or $P = 1$) reference precision-recall curves based on a reasonable model of how a precision-recall curve is typically shaped. A single reference curve will pass through each meaningful (i.e., $P > \rho$) point in the precision-recall plane. The parameter value, β, that indicates which curve passes through the IR system's precision-recall point would be sufficient to rank the IR systems that are being compared, but we want a measure that says something meaningful about how much better one system is compared to another, so we'll translate β into an extrapolated

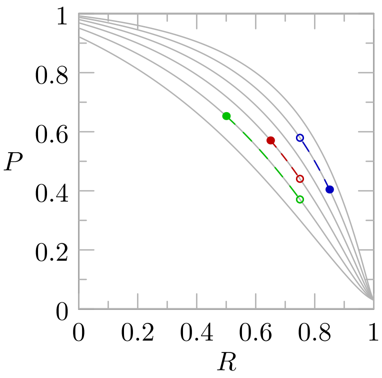

Figure 2: The three solid dots represent precision-recall measurements for three different systems. We find the approximate precision at target recall $R_T = 0.75$ for each point by moving along the reference curve that passes through it.

estimate of the precision at some target recall, R_T, by simply moving along the reference curve as shown in Figure 2. Furthermore, the extrapolated precision has a straightforward and meaningful interpretation—it is an estimate of the quantity we would have measured in the first place, $P(R_T)$, if we could demand that all IR systems reach exactly the same recall for evaluation. If the system is evaluated at a recall that is reasonably close to R_T, the extrapolated precision should be a good approximation to the actual precision at R_T, so it can be used in Equation (2) as a reasonable estimate of review cost. If the system is evaluated at a recall that is far from R_T, the extrapolated precision will still be a reasonable performance measure, it just won't necessarily be a good estimate for $P(R_T)$.

To be clear, the extrapolation will involve estimating $P(R_T)$ from a single point, $P(R)$. This approach does *not* involve knowing the full precision-recall curve, so it can be used on systems that provide binary relevance predictions.

It would be unwise to try to write down a set of reference precision-recall curves directly, because the slope of the precision-recall curve is related to the probability of a document at that recall level being relevant, $p(R)$, and it would be easy to accidentally write down a precision-recall curve that implied a probability function with strange lumps, or that failed to stay between 0 and 1. Instead, we'll write down a reasonable model for $p(R)$ and generate $P(R)$ from it.

The differential equation relating the precision and recall to the probability of a document being relevant is (in the continuum limit):

$$\frac{dP(R)}{dR} = \frac{P(R)}{R}\left[1 - \frac{P(R)}{p(R)}\right] \tag{3}$$

It can be confirmed by direct substitution that Equation (3) is solved by:

$$P(R) = \frac{R}{\int_0^R \frac{dr}{p(r)}} \tag{4}$$

for any well-behaved $p(R)$ that satisfies the normalization condition:

$$\int_0^1 \frac{dr}{p(r)} = \frac{1}{P(1)} \tag{5}$$

The intuition is that the integral in the denominator of Equation (4) is the number of documents retrieved to reach recall R divided by the number of relevant documents in the whole document population.

The next step is to choose a reasonable model for $p(R)$. To clearly differentiate between actual probability and precision curves and our model curves, we'll use x for the model probability and X for the model precision (corresponding to p and P respectively). Also, we'll take the model precision at $R = 1$ to be equal to the prevalence, i.e., $X(1) = \rho$. We aim to find a simple model that can produce a full range of realistic precision-recall curves. The model should also be one where the integral in Equation (4) can be performed analytically.

A ratio of two second-degree polynomials can easily be parameterized to provide monotonic probability curves satisfying $0 \leq p(R) \leq 1$, while accommodating widely varying curvature. The numerator will primarily control the behavior when R is close to 1. Examination of probability curves for several classification tasks (see bottom of Figure 3) suggests that the dominant behavior near $R = 1$ is proportional to $(1 - R)^2$. There must be an additional term in the numerator that approaches zero more slowly as $R \to 1$ or the integral in Equation (4) will diverge. Adding a constant term to represent a small number of documents that are found at random because the IR system has not detected the features that make them relevant is reasonable, and works well for $\rho \approx 0.01$. Some classification tasks with high prevalence show a hint of a $(1 - R)$ term, but it is small and it is probably not worthwhile to try to model it.

We take the denominator to be equal to the numerator plus an additive piece that is normally relatively small when $R \to 0$ since the probability of a document being relevant at $R = 0$ is typically high. Thus far, the form of the model is:

$$x(R) = \frac{1 + \beta^2(1 - R)^2}{1 + \beta^2(1 - R)^2 + c_0 + c_1 R + c_2 R^2} \tag{6}$$

where the $c_0 + c_1 R + c_2 R^2$ part must be non-negative for all $R \in [0, 1]$ to ensure that the probability is never greater than one.

The β parameter will ultimately be used to select the curve from the family of model curves based on the observed value of precision and recall; all other parameters must be eliminated. We take c_2 to be zero for simplicity, and because it doesn't seem to be necessary. The normalization condition in Equation (5) forces a relationship between c_0, c_1, and the prevalence:

$$c_0 = \frac{1 - \rho}{\rho} \frac{\beta}{\tan^{-1}\beta} - c_1 \left[1 - \frac{1}{2} \frac{\ln(1 + \beta^2)}{\beta \tan^{-1}\beta}\right] \tag{7}$$

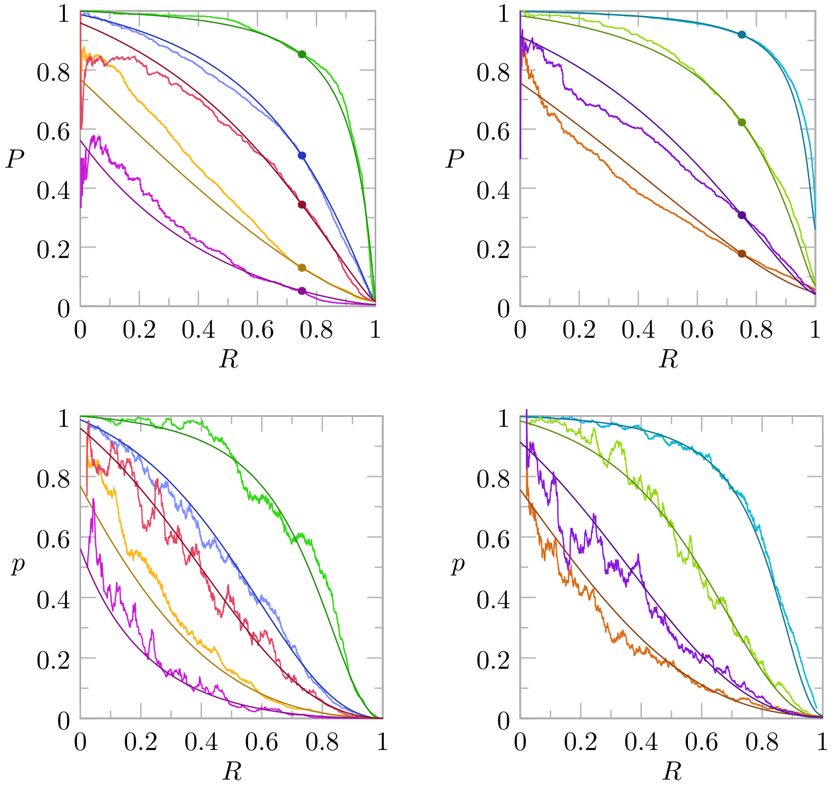

Figure 3: Classification tasks on left have ρ ranging from 0.005 to 0.014. Tasks on right have ρ from 0.040 to 0.255. Graphs along the top show precision-recall curves and model curves, X, where the β parameter for each model curve comes from the measured precision at $R = 0.75$, which is indicated with a dot. Bottom shows the probability curves and corresponding model curves, x. β values range between 16 and 526 on the left, and between 8 and 50 on the right.

To ensure that $x(R) \leq 1$ at $R = 0$ we must have $c_0 \geq 0$, which imposes a restriction on c_1:

$$c_1 \leq \frac{1-\rho}{\rho} \frac{\beta}{\tan^{-1}\beta} \left[1 - \frac{1}{2} \frac{\ln(1+\beta^2)}{\beta \tan^{-1}\beta} \right]^{-1} \tag{8}$$

It is useful to understand the impact of the final unwanted parameter, c_1, before choosing a value for it. The larger we make c_1, within the limit imposed by Equation (8), the smaller c_0 will be, due to Equation (7). The smaller c_0 is, the closer $x(0)$ will be to 1 for a given value of β, according to Equation (6). Probability and precision are equal at $R = 0$, so $X(0) = x(0)$. Larger c_1 pushes x, and therefore precision, values closer to 1 at $R = 0$.

If we take the largest allowed value for c_1, c_0 will be zero and all of the model's precision-recall curves (all values of β) will have precision equal to 1 at $R = 0$. That would be too extreme—we know from experience that precision is sometimes less than perfect at low recall. A model that forces perfect precision at zero recall will behave badly if forced to fit a point with modest precision at very low recall.

On the other hand, if our choice for c_1 makes c_0 too large, the model precision-recall curves will fall far short of perfect precision at $R = 0$, even when precision is observed to be relatively high at high recall. In other words, the precision-recall curves will flatten out instead of climbing toward 1 as $R \to 0$. Normally, the β^2 term in the numerator and denominator of Equation (6) would tend to drive the precision to 1 at low recall as β is increased, but Equation (7) shows that c_0 is superficially proportional to β for large β, due to the normalization requirement, so it fights against achieving perfect precision (though it will lose out to β^2 if β is large enough). Examination of Equation (7) reveals that the strongest part of c_0's dependence on β can be cancelled out with a judicious choice for c_1:

$$c_1 = \frac{1-\rho}{\rho} \frac{\beta}{\tan^{-1}\beta} \tag{9}$$

There are other possibilities for c_1 that vary in how much they reduce c_0, but Equation (9) is simple and produces reasonable results, as shown in Figure 3, which shows how the model compares to some actual precision-recall curves for tasks of varying difficulty and prevalence. For comparison, Figure 4 shows the poor model curves that would result from taking c_1 to be half of the value suggested in Equation (9).

The final model for the probability curves is:

$$x(R;\rho,\beta) = \frac{1+\beta^2(1-R)^2}{1+\beta^2(1-R)^2 + \frac{1-\rho}{\rho}\frac{\beta}{\tan^{-1}\beta}\left[R + \frac{1}{2\beta\tan^{-1}\beta}\ln(1+\beta^2)\right]} \tag{10}$$

and the model for the precision curves is:

$$X(R;\rho,\beta) = \frac{R}{R + \frac{1-\rho}{\rho}\left\{1 - \frac{\tan^{-1}[\beta(1-R)]}{\tan^{-1}\beta}\left[1 + \frac{1}{2\beta\tan^{-1}\beta}\ln(1+\beta^2)\right] + \frac{1}{2\beta\tan^{-1}\beta}\ln[1+\beta^2(1-R)^2]\right\}} \tag{11}$$

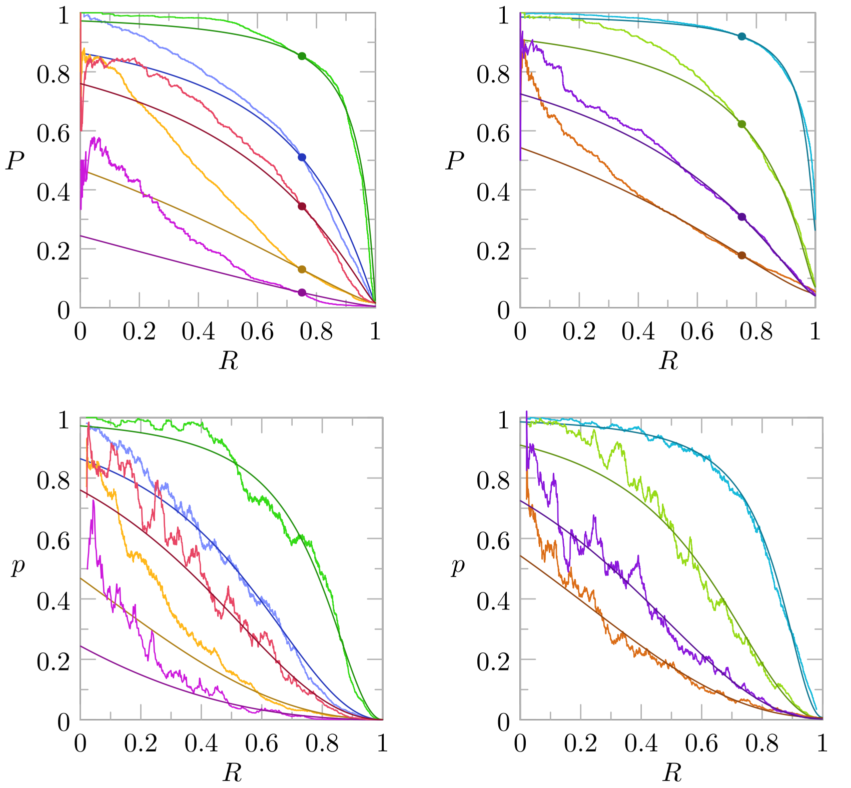

Figure 4: Example of a bad model. Similar to Figure 3, but c_1 is taken to be half of the value recommended in Equation (9), resulting in model curves that are too flat at low recall.

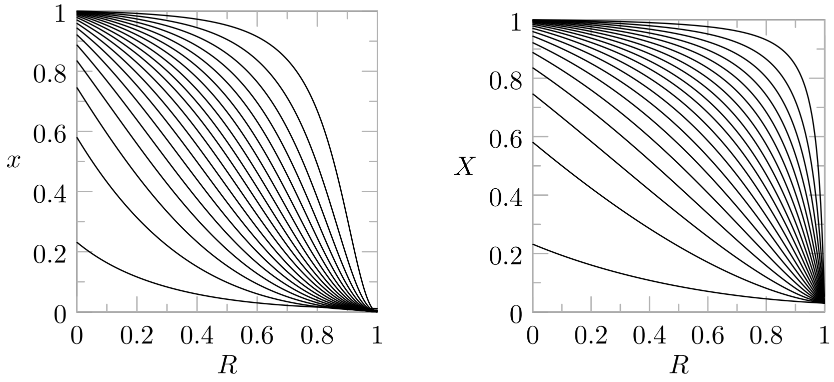

Figure 5: Graph of model probability curves, $x(R; \rho, \beta)$, on the left, and corresponding precision curves, $X(R; \rho, \beta)$, on the right, for $\rho = 0.03$ and various values of β ranging from 2.44 to 833, with β values chosen to give evenly spaced X values at $R = 0.75$.

Figure 5 shows a graph of the model probability function, x, and the corresponding reference precision-recall curves, X, for various values of β.

Finally, we determine which reference curve passes through the single precision-recall point that was measured, $P(R)$, by finding the value of β that satisfies:

$$P(R) = X(R; \rho, \beta) \tag{12}$$

which must be done numerically. Extrapolation should not be attempted if the measured recall or precision are extremely close to 1.0 because those regions contain very little information about how the system will perform at other recall levels, which is reflected in Figure 5 by many different reference curves being very close together in those regions.

To summarize the process, start with a single R and P value for the IR system being analyzed, similar to computing F_1. The prevalence for the document population, ρ, is also required. Solve Equation (12) numerically to find the β that corresponds to the known values of ρ, R, and P. Once β is known, extrapolate to any target recall, R_T, desired by computing $X(R_T; \rho, \beta)$ using Equation (11). Extrapolate results for several different IR systems (potentially measured at different recall levels, but ideally those levels should be close to R_T for accurate extrapolation) to the same R_T so they can be compared in a meaningful manner. The end result is a performance measure that has real meaning—it is an estimate of the precision at R_T, not some strange mixture of precision and recall that is hard to interpret, and it can be used in Equation (2) to estimate the number of documents that would need to be retrieved to reach recall R_T.

$X(R_T)$ can be computed from a single precision-recall point. If an entire precision-recall curve is available, $X(R_T)$ can be computed for each point on the curve and plotted as a function of R (the recall of the precision-recall point

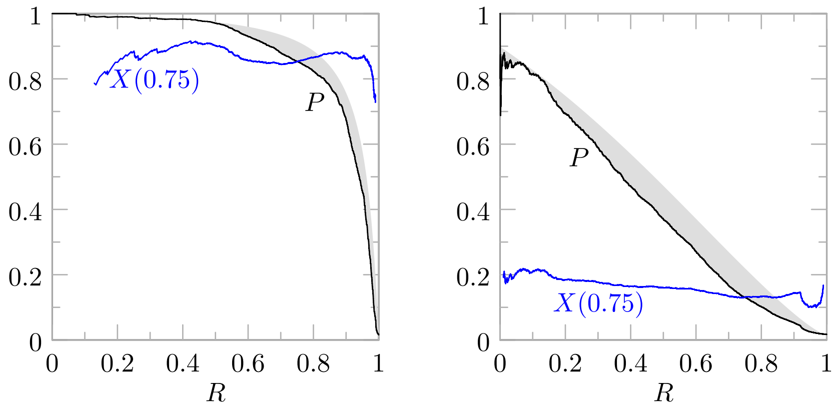

Figure 6: Precision and $X(0.75)$ as a function of recall for two tasks. The shaded regions indicate points where a system with better precision at that recall would get a lower $X(0.75)$ than the best $X(0.75)$ that the graphed system could achieve for any recall (though recall would normally be restricted).

used in the computation), similar to the F_1 plots in Figure 1. Figure 6 shows the plot for $R_T = 0.75$ with the data from Figure 1. The $X(0.75)$ curve must cross the P curve at $R = 0.75$ because $X(0.75)$ is intended to be an estimate of what P would be at $R_T = 0.75$ based on a measurement at some arbitrary R, so any sensible extrapolation scheme should give the exact right answer when the arbitrary R happens to equal R_T. To the extent that the extrapolation scheme is working well, $X(0.75)$ should be perfectly flat, which means that conclusions about IR system quality won't change if systems are evaluated at differing levels of recall. If the actual precision-recall curve deviates from the shape of the reference curves generated by the model, that will be reflected in $X(0.75)$ having bumps or a non-zero slope. For example, the precision-recall curve on the right side of Figure 6 shows a little dip when recall is greater than 0.9, and that dip also appears and is amplified in $X(0.75)$.

Figure 7 shows a comparison of six different supervised machine learning algorithms applied to the same easy IR problem. Dots highlight performance values at $R = 0.75$, which we'll take to be a reasonable recall level for the IR task. If all of the systems could be evaluated at exactly $R = 0.75$, the ranking of the systems would be unambiguous. If the full precision-recall curves were not known for the six systems and we had just six precision-recall points to compare, measured at slightly different recall levels, comparing precision values without any adjustment for the differences in recall might result in the wrong system winning because P varies strongly with R. The center graph in Figure 7 shows that F_1 gives a distorted view of the relative performance of the systems even if they are all measured at exactly $R = 0.75$. The F_1 values at $R = 0.75$ are much closer to each other than the corresponding P values are, giving the misleading impression that the performance differences are smaller than they are. The

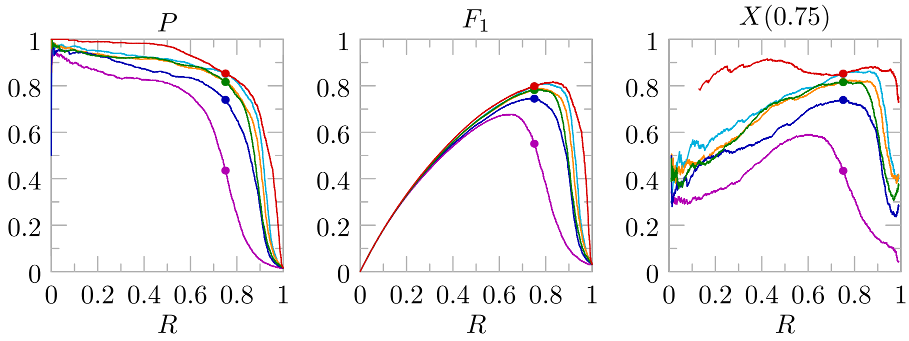

Figure 7: Comparison of P, F_1, and $X(0.75)$ as a function of recall for six different supervised machine learning algorithms trained and tested on the same data for the same relatively easy IR task. Dots highlight values at $R = 0.75$ for comparison.

rightmost graph in Figure 7 shows that the $X(0.75)$ values are exactly equal to the $P(0.75)$ values at $R = 0.75$, and for recall values slightly away from 0.75 the $X(0.75)$ values change very little (with the exception of the worst-performing system, which has a very oddly-shaped precision-recall curve). Evaluating the systems at recall levels slightly away from 0.75 is less likely to result in drawing wrong conclusions about relative performance with $X(0.75)$ than with P.

Figure 8 shows the same systems applied to a more difficult IR problem. The system that would come in dead last based on precision for all values of recall greater than 0.57 would take fourth place in an evaluation based on maximizing F_1, illustrating the fact that maximum F_1 is not an appropriate measure when high recall is required. The F_1 values at $R = 0.75$ are again distorted compared to P at $R = 0.75$. In this case, they are more spread out, but not by a uniform amount. The F_1 values change significantly as you move away from $R = 0.75$, so an evaluation involving recall values that vary slightly could result in incorrect performance rankings based on F_1. Again, $X(0.75)$ is seen to be flatter near $R = 0.75$, so there is less risk of erroneous performance rankings due to varying recall.

Finally, Figure 9 examines extrapolation accuracy over small differences in recall. The left side shows the extrapolation error, $X(R) - P(R)$, where X extrapolates from $P(R+0.05)$. The right side analyzes the simpler extrapolation method of assuming that precision is constant, so it graphs $P(R+0.05) - P(R)$. X is seen to generally have smaller errors than the simple method. As one might expect, the simple method produces predictions that are too low when the precision-recall curve isn't flat.

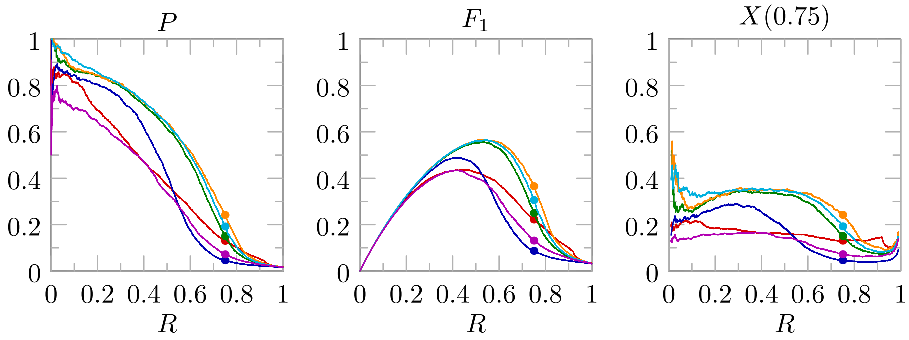

Figure 8: Comparison of P, F_1, and $X(0.75)$ as a function of recall for six different supervised machine learning algorithms trained and tested on the same data for the same relatively difficult IR task. Dots highlight values at $R = 0.75$ for comparison.

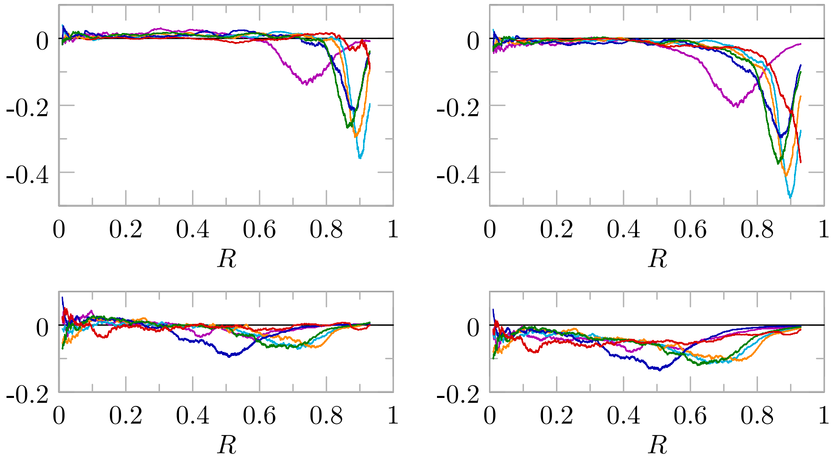

Figure 9: Extrapolated precision minus actual precision. Values are extrapolated from recall $R + 0.05$ down to recall R. Extrapolation on left side uses X. Right side uses the simple method of treating the precision as a constant. Top corresponds to the same data and classification algorithms as Figure 7. Bottom corresponds to data and algorithms from Figure 8.

3 Conclusions

The proposed performance measure based on extrapolating precision values to a target recall does not make strong assumptions about the appropriate trade-off between recall and precision, as F_1 does, so it is sensitive to the IR system's performance while being less sensitive to the recall level where the measurement is made. This reduces the risk of drawing incorrect conclusions about which system is performing best when the measurements are made at different levels of recall. The extrapolation method involves a family of reference precision-recall curves that are seen to have reasonable shapes compared to a modest set of test curves. A few tests of extrapolations over small differences in recall are found to be more accurate than the simple method of extrapolating precision to lower recall by treating precision as a constant. Work remains to be done on computing confidence intervals on the extrapolated precision when it is estimated via sampling.

References

[Mat75] B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975.

[OW13] Douglas W. Oard and William Webber. *Information Retrieval for E-Discovery.* now Publishers, 2013.

[vR79] C. J. van Rijsbergen. *Information Retrieval.* Butterworth, 2nd edition, 1979.