A Geometric Interpretation of R-precision and Its Correlation with Average Precision

Javed A. Aslam*, Emine Yilmaz, Virgiliu Pavlu
College of Computer and Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115
{jaa,emine,vip}@ccs.neu.edu

ABSTRACT

We consider two of the most commonly cited measures of retrieval performance: average precision and R-precision. It is well known that average precision and R-precision are highly correlated and similarly robust measures of performance, though the reasons for this are not entirely clear. In this paper, we give a geometric argument which shows that under a very reasonable set of assumptions, average precision and R-precision both approximate the area under the precision-recall curve, thus explaining their high correlation. We further demonstrate through the use of TREC data that the similarity or difference between average precision and R-precision is largely governed by the adherence to, or violation of, these reasonable assumptions.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Performance evaluation*

General Terms

Theory, Measurement, Experimentation

Keywords

Average Precision, R-precision, Precision-recall Curve

1. INTRODUCTION

Given a ranked list of documents returned in response to a query, the average precision of this list is the average of the precisions at all relevant documents,[1] which is approximately the area under the precision-recall curve, while the R-precision of this list is the precision at rank R, where R is the number of documents relevant to the query. It has been shown that average precision and R-precision are highly correlated [3, 2] and have similar stability in terms of comparing systems using different queries [1]. The correlation between average precision and R-precision has been considered quite surprising given the fact that R-precision considers only a single precision point while average precision evaluates the area under the entire precision-recall curve [1]. In this paper, we provide a geometric interpretation of R-precision which shows that under a very reasonable set of assumptions, average precision and R-precision both approximate the area under the precision-recall curve, thus explaining their correlation.

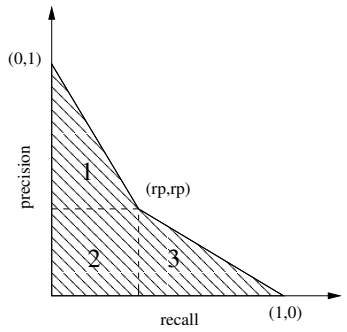

Figure 1: Precision-recall curve obtained by connecting points $(0,1)$, (rp,rp), $(1,0)$ with straight lines.

Given a query with R relevant documents, consider the list of documents returned by a retrieval system in response to this query, and let tot_rel(i) be the total number of relevant documents retrieved up to and including rank i. By definition, the R-precision rp of this list is the *precision* at rank R, $rp = \text{tot\_rel}(R)/R$. Furthermore, note that the *recall* at rank R is also tot_rel(R)/R. Thus, at rank R, the list has both precision and recall equal to rp, and assuming a continuous precision-recall curve, this curve would pass through the point (rp, rp).

Now assume that the precision-recall curve starts at the point $(0,1)$ and ends at the point $(1,0)$. Note that this assumption is often approximately true: retrieval systems tend to have high precisions at low recall levels and low precisions at high recall levels. Finally, assume that the precision-recall curve can be approximated in a piecewise-linear fashion by connecting these three points with straight lines (see Figure 1).

Now consider the area under this piecewise-linear approximation of the actual precision-recall curve, i.e., the shaded area in Figure 1. It can easily be shown that this area is rp by calculating the areas associated with the square (2) and the triangles (1) and (3). Thus, given the facts that

*We gratefully acknowledge the support provided by NSF grant CCF-0418390.

[1]The precision at an unretrieved relevant document is assumed to be zero.

SIGIR'05 August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

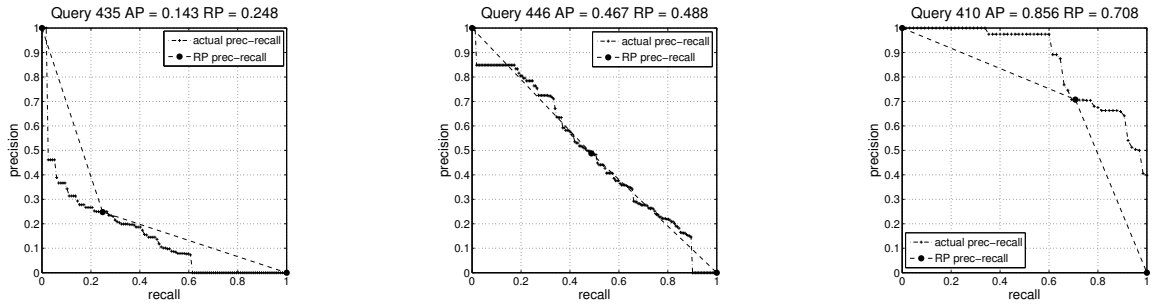

Figure 3: Actual precision-recall curves versus piecewise-linear approximations passing through points $(0,1)$, (rp,rp) and $(1,0)$ for system fub99a in TREC8.

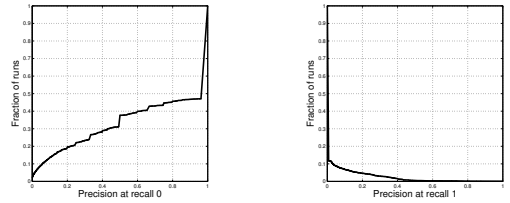

Figure 2: Cumulative distributions of runs in TREC8 based on precision at recall 0 (left) and precision at recall 1 (right).

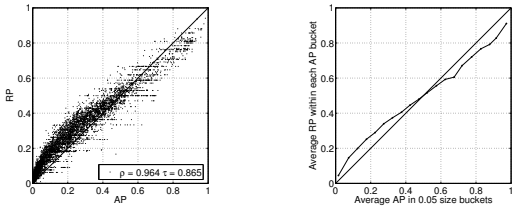

Figure 4: TREC8 average precision versus R-precision for each run (left) and for AP buckets of size 0.05 (right).

(1) average precision is approximately the area under the precision-recall curve, (2) under the assumptions stated, the precision-recall curve can be approximated by a piecewise-linear fit to the points $\{(0,1), (rp,rp), (1,0)\}$, and (3) the area under this piecewise-linear approximation is rp, we have that R-precision is approximately average precision.

2. EXPERIMENTAL RESULTS

We tested our hypotheses and assumptions using data from the TREC8 collection. First, we show that our assumption that the actual precision-recall curves start at the point $(0,1)$ and end a the point $(1,0)$ is satisfied by the majority of the runs. Figure 2 illustrates the cumulative distributions of runs corresponding to precisions at recalls 0 and 1, respectively.

In the left plot, for a precision value p on the x-axis, we calculate the fraction of runs whose precision at recall 0 is between 0 and p. In the right plot, we calculate the fraction of runs whose precision at recall 1 is between p and 1. As indicated in these plots, over 50% of the runs have precision 1 at recall 0, and nearly 90% of the runs have precision 0 at recall 1. (We note that this latter fact is partly a function of the truncated lists submitted to TREC.)

We next consider the assumption that precision-recall curves are piecewise-linear. In reality, precision-recall curves tend not to have a sharp change in slope at the point (rp,rp); rather, they tend to be "smoother" and concave-up for values of $rp < 1/2$ and concave-down for values of $rp > 1/2$ (see Figure 3). Thus we expect that R-precision, the area under the piecewise-linear approximation of the actual precision-recall curve, will tend to *overestimate* average precision when $rp < 1/2$, and it will tend to *underestimate* average precision when $rp > 1/2$. This fact is also illustrated in Figure 3, where the actual precision-recall curves of the system fub99a in TREC8 are compared with piecewise-linear approximations for three different queries, one each where $rp < 1/2$, $rp \approx 1/2$, and $rp > 1/2$. This phenomenon is fairly consistent across all the runs in TREC8. In Figure 4 (left), we plot the average precisions and R-precisions for each run submitted to the conference, together with the line $y = x$ for comparison. Note that when R-precision is less than $1/2$, it tends to overestimate average precision, and when R-precision is greater than $1/2$, it tends to underestimate average precision. For clarity, in the right plot these same runs are divided into average precision ranges of size 0.05, and for the runs in each such average precision "bucket," the average R-precisions and average precisions are calculated and plotted.

3. CONCLUSIONS

We have shown that under a reasonable set of assumptions, R-precision approximates the area under the precision-recall curve, thus explaining its high correlation with average precision. Furthermore, our geometric interpretation of R-precision implies that R-precision should overestimate average precision at values less than $1/2$ and underestimate average precision at values greater than $1/2$, and this fact is borne out in practice as well.

4. REFERENCES

[1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40. ACM Press, 2000.

[2] J. Tague-Sutcliffe and J. Blustein. A statistical analysis of the TREC-3 data. In *Proceedings of the Third Text Retrieval Conference (TREC-3)*, pages 385–398, 1995.

[3] E. M. Voorhees and D. Harman. Overview of the seventh text retrieval conference (TREC-7). In *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, pages 1–23, 1999.