7 PERFORMANCE WITH ONE TERM

The price of reliability is the pursuit of the utmost simplicity.
—C. A. R. Hoare, 1980 ACM Turing Award Lecture

7.1 INTRODUCTION

The performance of a retrieval system can be determined in a straightforward manner with analytic techniques if one limits oneself to using only a single term from the query. Limiting our discussion to one term allows us to fully understand many retrieval characteristics and options that are far more difficult to understand in a multi-term case.

We begin with the measure of the quality of a ranking method, $\mathcal{Q}$ (Equation 6.2). This is combined with the characteristics of both the best-case (optimal) and the worst-case ranking to compute the Average Search Length (ASL), the expected position of a relevant document in the ranked list of documents. Then, given the characteristics of the query, the database, and of the ranking algorithm itself, the expected performance of the system can be computed.

Documents may be presented to the user in one of two different orders. Documents with a binary query feature with frequency d may be followed by those with frequency $\overline{d} = 1 - d$, which we refer to as the *optimal ordering*. It is assumed that the term weight for d is greater than for $\overline{d}$; if this is not the case, the values may be switched (re-parameterized) so that the weight for d is greater than or equal to the weight for $\overline{d}$. Thus, we assume that the features are re-parameterized so that feature frequency d multiplied by the term weight has a higher value than feature frequency $\overline{d}$ multiplied by the term weight. This is best-case or optimal ordering. The worst-case ordering retrieves documents with feature frequency $\overline{d}$ before documents with frequency d.

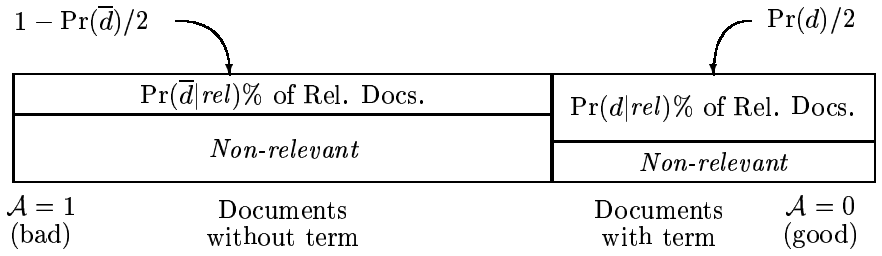

Figure 7.1. Documents ordered for retrieval so that $1 \succ 0$.

Definition 7.1 ($\mathcal{A}$) *The parameter $\mathcal{A}$ is the expected proportion of documents examined in an optimal ranking if one examines all the documents up to the document in the average position of a relevant document. It is the expected position of a relevant document, scaled from 0 to 1.*

The average location of a relevant document in the optimal ranking may be computed based on a scale from 0 to 1, a unit scale, with an $\mathcal{A}$ of 0 representing the average position of relevant documents being at the beginning of the search process and an $\mathcal{A}$ of 1 being at the end of the search process.

When viewing this in an information filtering context, where some documents are passed through the filter and others are not passed, one can interpret $\mathcal{A}$ as 1 minus the probability a document is retrieved when retrieving documents in ranked order up to the average position of a relevant document. Similarly, if documents aren't ranked, $\mathcal{A}$ may be understood as 1 minus the probability that a document will be retrieved given that only documents of the average frequency of the term in relevant documents or higher are retrieved.

Theorem 7.1 ($\mathcal{A}$ for single terms) *Given binary terms with probabilities $p = \Pr(d|rel)$ and $t = \Pr(d)$, then*

$$\mathcal{A} = \frac{1 - p + t}{2}. \tag{7.1}$$

The variable $\mathcal{A}$ is computed by noting that documents with feature frequency d are at the low end of the $\mathcal{A}$ spectrum and those with feature frequency $\overline{d}$ at the high end of the spectrum. The middle (average) position for each of the profiles when they are arranged in order is such that $\Pr(d)/2$ is the average position for documents with feature frequency d. The mean position is $1 - \Pr(\overline{d})/2$ for the documents with feature frequency $\overline{d}$, as in Figure 7.1. Documents of each feature frequency may be weighted by the probability of having the frequency of the feature given that they are relevant. This is the percent of relevant documents that have the particular frequency. Thus,

$$\mathcal{A} = \Pr(d|rel)\Pr(d)/2 + \Pr(\overline{d}|rel)(1 - \Pr(\overline{d})/2). \tag{7.2}$$

The worst-case ranking that is obtained when the ordering is the opposite, that is, the placement of documents without the feature ordered before documents with

the feature $(\overline{d} \succ d)$, is

$$\overline{\mathcal{A}} = \Pr(\overline{d}|rel)\Pr(\overline{d})/2 + \Pr(d|rel)(1 - \Pr(d)/2). \tag{7.3}$$

Here, $\overline{\mathcal{A}} = 1 - \mathcal{A}$.

Equation 7.2 may be simplified algebraically to:

$$\mathcal{A} = \frac{1 + \Pr(d) - \Pr(d|rel)}{2}.$$

Because the terms are binary, that is, have the values $d = 1$ or $\overline{d} = 0$, we find that

$$\mathcal{A} = \frac{1 - p + t}{2},$$

where $p = \Pr(d|rel)$ and $t = \Pr(d)$.

Example. Consider a sequence of ten documents, the first four having the feature in question (and three of these documents being relevant) and the remaining six documents not having the feature (and one of these being relevant). We find that $\Pr(d|rel) = 3/4$, $\Pr(\overline{d}|rel) = 1/4$, $\Pr(d) = 4/10$, and $\Pr(\overline{d}) = 6/10$. Applying Equation 7.2, we compute the weighted average position of a relevant document: $(3/4)(2/10) + (1/4)(7/10) = 13/40$. Alternatively, using Equation 7.1, we find that $(1 - p + t)/2 = (1 - 3/4 + 4/10)/2 = 13/40$. It is most helpful to view this process as computing the position of the average relevant document, normed to being between 1 and 0, or it is the proportion of documents in the dataset retrieved when retrieving the conceptual document at average position of a relevant document.

EXERCISE 7.1. [*Easy*] Consider a single term query. Five relevant documents have the term and five relevant documents don't have the term. One hundred documents have the term and one thousand documents don't have the term. Compute $\mathcal{A}$.

Best and Worst-Case $\mathcal{A}$

Knowledge of the best-case and worst-case values for $\mathcal{A}$ can allow us to predict best and worst outcomes, as well as to determine how close an empirical method is to optimal. We refer to the best-case as the upper bounds for performance and the worst-case as lower bounds, although the upper bounds for performance are actually the lowest numeric value for $\mathcal{A}$ and ASL, with a similar direction reversal from the traditional for worst-case and lower bounds of performance representing high $\mathcal{A}$ and ASL values.

Theorem 7.2 (Approximate best-case $\mathcal{A}$) *The best-case value for $\mathcal{A}$ may be approximated when $p = 1$ and $t = 0$. In this case $\mathcal{A} = 0$.*

Corollary 7.1 (Approximate worst-case $\mathcal{A}$) *The worst-case value for $\mathcal{A}$ may be approximated when $p = 0$ and $t = 1$. In this case $\mathcal{A} = 1$.*

Theorem 7.3 (Exact best-case $\mathcal{A}$) *The best-case value for $\mathcal{A}$ occurs when $p = 1$ and $t \to 0$ and the value obtained is $\mathcal{A} = (1 - 1 + g)/2 = g/2$, where g is the generality.*

When $p = 1$, the lowest value for t becomes $g = \Pr(rel)$, the generality. Note that t will have as its low value gp, the proportion of all documents that have frequency

1 and are relevant. If we compute $\mathcal{A} = (1 - p + gp)/2$, we find that as p increases, the $-p$ component will lower the $\mathcal{A}$ value faster than the gp value will raise it, for all $g < 1$. Thus, there is no intermediate value, $p < 1$, that will produce the best-case $\mathcal{A}$ before $p = 1$.

Corollary 7.2 (Exact worst-case $\mathcal{A}$) *The worst-case value for $\mathcal{A}$ occurs when $p = 0$ and $t \to 1$ and $A = (1 - 0 + (1 - g))/2 = 1 - g/2$.*

As with the best-case performance, setting $p = 0$ establishes a limit on the highest value that t can have. Clearly, if no relevant documents have the feature in question, then the highest probability of a document having the feature is $1 - g$.

7.2 COMPUTING THE ASL

Being able to predict retrieval performance based on the conditions that exist at the start of the retrieval process is the cornerstone of understanding text filtering and retrieval.

Theorem 7.4 (ASL for single terms) *Given N documents and $\mathcal{A}$ (Equation 7.1) and $\mathcal{Q}$ (Equation 6.2) as defined earlier, then the Average Search Length (ASL) when using a single query term is*

$$ASL = N\left(\mathcal{Q}\mathcal{A} + \overline{\mathcal{Q}}\overline{\mathcal{A}}\right) + 1/2.$$

To estimate the ASL it is necessary to compute the the weighted average of the $\mathcal{A}$ and $\overline{\mathcal{A}}$ values. The weighting factors are the percent of orderings that are optimal and the percent of rankings that are worst-case. The probabilities or weightings used in computing the average are $\mathcal{Q}$ for $\mathcal{A}$ and $\overline{\mathcal{Q}}$ for $\overline{\mathcal{A}}$, where $\overline{\mathcal{Q}} = 1 - \mathcal{Q}$. We may compute

$$\begin{aligned} ASL = N\Big(&\mathcal{Q}\left[\Pr(d|rel)\Pr(d)/2 + \Pr(\overline{d}|rel)(1 - \Pr(\overline{d})/2)\right] \\ &+ \overline{\mathcal{Q}}\left[\Pr(\overline{d}|rel)\Pr(\overline{d})/2 + \Pr(d|rel)(1 - \Pr(d)/2)\right]\Big) \\ &+ 1/2 \end{aligned} \tag{7.4}$$

$$ASL = N\left(\mathcal{Q}\mathcal{A} + \overline{\mathcal{Q}}\overline{\mathcal{A}}\right) + 1/2. \tag{7.5}$$

The ASL may be estimated more simply if optimal retrieval is assumed. We find that the ASL for optimal ranking, that is, where $\mathcal{R}_o$ is the ranking method used, is

$$\begin{aligned} ASL_O &= N\frac{1 - p + t}{2} + \frac{1}{2} \\ &= N\mathcal{A} + \frac{1}{2}. \end{aligned} \tag{7.6}$$

We may denote the ASL performance $\mathcal{P}$ given p, t, and a database of size N as

$$\begin{aligned} \mathcal{P}(p_e, p, t_e, t, q_e, q, N) \quad = \quad & N[\mathcal{Q}(p_e, p, t_e, t, q_e, q)\mathcal{A}(p, t, q) \\ & + \overline{\mathcal{Q}}(p_e, p, t_e, t, q_e, q)\overline{\mathcal{A}}(p, t, q)] + \frac{1}{2} \end{aligned} \tag{7.7}$$

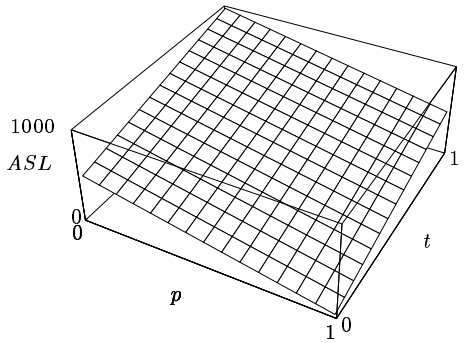

Figure 7.2. The performance as $p = \Pr(d|rel)$ and $t = \Pr(d)$ vary, decreasing the ASL (increasing the performance) as p increases and t decreases, for $N = 1000$ and $\mathcal{Q} = 1$.

where the parameters subscripted with e are estimates and those without subscripts are the correct or actual values.

The theorems here assume that $\mathcal{Q} = 1$.

Theorem 7.5 (Approximate best-case ASL) *If* $p = 1$ *and* $t = 0$, *then* $\mathcal{A} = 0$ *and* $ASL = 1/2$.

Theorem 7.6 (Exact best-case ASL) *When* $p = 1$ *and* t *approaches* 0, *then* $\mathcal{A} = g/2$ *and* $ASL = Ng/2 + 1/2$, *where* g *is the generality.*

Corollary 7.3 (Approximate worst-case ASL) *If* $p = 0$ *and* $t = 1$, *then* $\mathcal{A} = 1$ *and* $ASL = N + 1/2$.

Corollary 7.4 (Exact worst-case ASL) *When* $p = 0$ *and* t *approaches* 1, *then* $\mathcal{A} = 1 - g/2$ *and* $ASL = N(1 - g/2) + 1/2$.

Example. A closer examination of Equations 7.5 to 7.7 helps show the relationships and functions of the variables. We assume that $N = 1000$ for this example. Figure 7.2 shows the ASL when retrieval is optimal, that is, when $\mathcal{Q} = 1$ and p and t are varied over the full range of possible values. Beginning with the leftmost corner in Figure 7.2 and moving clockwise, we can see that when $p = t = 0$, the $\mathcal{A}$ factor is simply the constant $1/2$, implying essentially random retrieval. Setting p and t to 0 implies that all the documents have feature value 0 and all the relevant documents have value 0; essentially, the relevant documents are not distinguishable from the other documents.

At the top of Figure 7.2, we find that $t = 1$ and $p = 0$, resulting in the weighting of $\mathcal{A} = 1/2 - 0/2 + 1/2 = 1$. Almost all the relevant documents have feature frequency 0 while almost all of the remaining documents have feature frequency 1, forcing the relevant documents to the end of the ordering, assuming the ordering $1 \succ 0$.

Continuing clockwise around Figure 7.2, the $\mathcal{A}$ when $p = t = 1$ is computed as $\mathcal{A} = 1/2 - 1/2 + 1/2 = 1/2$. This is similar to the situation that arises when $p = t = 0$ on the left hand side of the Figure; that is, the relevant documents are distributed in the same class of documents as are all the documents, and the retrieval is essentially random.

The best-case retrieval occurs at the bottom of the Figure when $p = 1$ and $t \to 0$. Here the $\mathcal{A}$ component is computed as $1/2 - 1/2 + 0/2 = 0$. There are few documents with the feature but most of the relevant document have the feature, making it easy to isolate the relevant documents.

We compute ASL as $N(\mathcal{QA} + \overline{\mathcal{QA}}) + 1/2$. One factor in decreasing the ASL is to decrease N. While this is not possible in some circumstances, it is obvious that one will not have to look at as many documents if there are fewer documents to select from, all other factors held equal. Keeping N constant, the ASL is minimized in two cases. The first is when the ranking procedure is optimal and $\mathcal{A}$ approaches 0 and the second case is when the ranking procedure is worst-case and $\mathcal{A}$ approaches 1.

Variation in ASL values

The variation in ASL with optimal ranking may be used to show how much variation might be expected, given the amount of knowledge available about the parameters. We assume here that t is fully and accurately known. The variation in ASL may be computed by determining the ASL for both very high levels of the prior distribution of p and for low levels of p. For example, if we compute the ASL using the value of p that occurs at the 2.5% point for the prior distribution of p (the point where the CDF for the prior distribution is .025) and also compute the ASL for the 97.5% point, the range between the two different ASLs represents the set of ASL values (or *confidence interval*) over which we have a 95% chance that the actual ASL will occur if we had perfect knowledge.

EXERCISE 7.2. [*Easy*] Assume that we have a database and query such that $N = 10,000,000$, $t = .5$, and that ranking is optimal. What is the ASL when $p = .5$ and what is the ASL at $p = .51$? What is the ASL at $p = .9999$?

EXERCISE 7.3. [*Easy*] Do the preceding exercise assuming that ranking is random and $\mathcal{Q} = .5$.

EXERCISE 7.4. [*Moderate*] Assuming that $N = 10,\ 000,\ 000$, $p = .50$, and ranking is optimal, what will be the value of t if the ASL is to be improved so that $ASL = 10$, a reasonable ASL for many human searchers? How do you interpret and explain this result?

EXERCISE 7.5. [*Moderate*] Select 4 terms occurring in this book that you think might be reasonable search terms. Search for these terms on an existing retrieval system or network search engine where you can search for common terms, e.g. "the," to find about the approximate size of the database, or where you know the size of the database through other means. Assume that you are searching for the terms individually and that each query term is in *all* of the relevant documents; it can't get much better than this! Given the empirically determined t values and the assumed $p = 1$, what will be the $\mathcal{A}$ and ASL value for each term assuming optimal retrieval. What does this say about the likely search performance?

7.3 A GENERAL THEORY OF PERFORMANCE

A general theory of retrieval performance is developed below consistent with the earlier text filtering and retrieval performance model. The reader who has understood what has been presented so far should find it a logical and comfortable generalization. The general theory allows us to develop an expression relating performance and probabilistic parameters. It is expressed using general probabilities, rather than

specific distributions, making it more flexible than what was presented above. We first present the model for document features that are continuous, with an application to normal feature distributions. We then examine a model for discrete feature distributions, including the binary distribution. The Poisson distribution for term frequencies, which has been argued to closely approximate the distribution of terms in natural language, and the negative binomial distribution, which may be similarly used, can be incorporated into the general model.

The expected term frequency for a term in the relevant documents, $E(d|rel) = \mu_r$, is the relative average position (scaled from 0 to 1) for a document in the distribution of relevant documents. By looking at the documents with frequencies at or above this position in all the documents, the $\mathcal{A}$ component may be computed as follows:

$$\mathcal{A} = \int_{\mu_r}^{\infty} \Pr(j)\, dj \tag{7.8}$$

where $\Pr(j)$ is the probability of a document having feature frequency j.

We use two distributions in developing this model further. We refer to f_{all}, the distribution of feature frequencies in all documents, while f_{rel} refers to the distribution of feature frequencies in relevant documents. The mean of the distribution f_{rel} is μ_r and the mean of the distribution of all documents, f_{all}, is μ, denoted without a subscript.

Definition 7.2 (Survival function) *We denote the sum of the probabilities above frequency x for the distribution of all terms as the* survival function, *$\mathcal{S}(x, all)$.*

The survival function is one minus the cumulative distribution function for the appropriate distribution.

Theorem 7.7 ($\mathcal{A}$ for binary relevance) *For continuous term distributions and binary relevance,*

$$\mathcal{A} = \mathcal{S}(\mu_r, all). \tag{7.9}$$

This follows from the definitions of $\mathcal{A}$ and the survival function.

Continuous Relevance

In some circumstances, it may be beneficial to consider relevance as a continuous variable, with those documents the user wishes to see as the set of documents with associated relevance values at or above a certain point on the relevance scale.

Theorem 7.8 ($\mathcal{A}$ for continuous relevance) *For continuous relevance and relevance scaled from 0 (no relevance) to 1 (complete relevance), then for documents of a relevance value r of x or above, $0 \leq x \leq 1$, $x \leq r \leq 1$, with the average frequency for documents of relevance r denoted as μ_r, the average value of $\mathcal{A}$, as computed in Equation 7.9, is*

$$E_x(\mathcal{A}) = \int_x^1 \frac{\Pr(r)}{\int_x^1 \Pr(r')\, dr'} \quad \mathcal{S}(\mu_r, all)\, dr.$$

The nature of the distribution of μ_r over the set of relevance values is an open and important question.

7.4 CONTINUOUS FEATURE DISTRIBUTIONS

The nature of $\mathcal{A}$ is best understood by examining continuous feature distribution models because of their relative simplicity. We assume binary relevance below unless stated otherwise.

Normal Distribution

The model for $\mathcal{A}$ that is consistent with normally distributed term frequencies may be applied to a variety of feature distributions when used as an approximation. The mean value for the distribution of the features in relevant documents is μ_r, μ is the mean for the distribution of the feature in all documents, and we assume there are equal variances σ^2 for both distributions. We find that

$$\mathcal{S}(\mu_r, all) = 1 - \frac{1}{2}\left(1 + Erf\frac{\mu_r - \mu}{\sqrt{2}\sigma^2}\right), \tag{7.10}$$

where $Erf(x)$ is the standard error function,

$$Erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,dt.$$

For small increments, the increase or decrease in performance is proportional to the factor $(\mu_r - \mu)/\sigma^2$, although the performance does not improve linearly as the factor $(\mu_r - \mu)/\sigma^2$ increases. The reader might note the similarity in structure of this factor and the E measures proposed by Swets (Equation 4.2) and by Brookes (Equation 4.3). This active component in retrieval performance is such that when the difference between μ_r and μ increases, performance improves. As the mean μ_r for the distribution of features in relevant documents increases beyond the mean μ for all documents, presumably most of them non-relevant, the performance improves because of the increased separation between the feature frequencies for the two categories of documents.

When the variance increases for a fixed set of differences, the performance decreases. This may be best understood by imagining the relevant and all the documents having their means at some fixed difference. If the distributions are narrowed, then there is a less overlap between the two categories of documents and a greater degree of separation. Conversely, if both distributions are nearly flat, that is, both distributions have a high variance, then there is a great deal of overlap between the two distributions, making them harder to separate, resulting in decreased performance.

Similar symbolic and numerical techniques may be used to compute performance values for other continuous distributions, such as the gamma distribution or, more generally, for the exponential family of distributions.

Example. Assume that we are searching using a feature that is normally distributed with $\mu_r = 20$ and $\mu = 10$, with $\sigma^2 = 10$. Using Equation 7.10, $\mathcal{A} = .159$. If $\mu_r = 15$ and the feature is less discriminating, then we find that $\mathcal{A} = .309$, while if μ_r is further reduced to 10, so that $\mu_r = \mu$, we find that $\mathcal{A} = .5$, which is what one expects from performance that is equivalent to random.

EXERCISE 7.6. [*Moderate*] Develop an equation similar to Equation 7.10 that doesn't assume equal variances. Give an example of when this might prove useful.

7.5 DISCRETE TERM FREQUENCIES

When computing $\mathcal{A}$ using this general approach consistent with discrete term frequencies, a similar discrete technique may be developed with sums rather than integrals. A satisfactory model for the discrete case is similar to the original binary model. This general model will also prove useful later as the basis for multivariate models. $\mathcal{A}$ is computed as

$$\mathcal{A} = 1 - \sum_{i=0}^{\infty} \left(\mathcal{C}_i(D) - \frac{\Pr(d_i)}{2} \right) \Pr(d_i|rel), \tag{7.11}$$

where

$$\mathcal{C}_n(D) \quad = \quad \sum_{i=0}^{n} \Pr(d_i). \tag{7.12}$$

The function $\mathcal{C}_n(D)$ represents the cumulation of probabilities up to the value for n for the random variable D. Using this, Equation 7.11 computes the average position of relevant documents. Note that the majority of Equation 7.11 is subtracted from 1 because it computes an average such that higher values are associated with higher term frequencies, which are, in turn, associated with the lower ASL values.

The continuous analog of Equation 7.11 is used when features d are continuously distributed:

$$A = 1 - \int_{-\infty}^{\infty} \mathcal{C}_i(D) \Pr(d_i|rel)\, di$$

where the $\mathcal{C}_i(X)$ function is the cumulative distribution function $\int_{-\infty}^{i} \Pr(X = x_j)\, dj$.

Binary Model Parameters

Using Equation 7.11, and using the notation $p = \Pr(d = 1|rel)$ and $t = \Pr(d = 1)$, we compute $\mathcal{A}$ for binary terms as

$$\begin{aligned} \mathcal{A} &= 1 - \left[\left((1-t) - \frac{1-t}{2} \right) (1-p) + \left(1 - \frac{t}{2} \right) p \right] \\ &= \frac{1-p+t}{2}, \end{aligned}$$

the result that was computed earlier as Equation 7.1.

Poisson Model Parameters

Estimating the performance of retrieval may assume that term frequencies are distributed as described by the Poisson model. Terms are distributed this way if each occurrence of a term in a document depends only on the instantaneous rate of occurrence of terms, and not on when the previous term occurred. Applying Equation 7.11

to the Poisson distribution results in

$$\mathcal{A} = 1 - \sum_{i=0}^{\infty} \left(\sum_{j=0}^{i} \frac{e^{-\lambda}\lambda^j}{j!} - \frac{e^{-\lambda}\lambda^i}{2(i!)} \right) \frac{e^{-\lambda_r}\lambda_r^i}{i!}. \tag{7.13}$$

Computing the survival function from point x upward for the Poisson distribution may be simplified by using the lower incomplete gamma function, $\gamma(x,n) = \int_0^n e^{-t}t^{x-1}\,dt$. The survival function for the Poisson distribution from x upwards is then $\gamma(x,\lambda)/\Gamma(x)$ (Haight, 1967).

Example. Applying Equation 7.13 to the situation where $\lambda_r = \lambda$ results in $\mathcal{A} = .5$, that is, when the relevant documents are distributed the same as the rest of the documents, then $\mathcal{A}$ reflects the fact that the average relevant document is to be found in the middle of the documents. When $\lambda_r = 4$ and $\lambda = 3$, $\mathcal{A}$ decreases to .355. As the gap between λ_r and λ increases, $\mathcal{A}$ continues to decrease.

EXERCISE 7.7. [*Research Problem*] Assume that term frequencies are truly Poisson-distributed and that ranking is optimal. Systems may produce binary valued index terms by assigning a binary 1 when the term is present in the document and a 0 when the term is absent. Compute the decrease in ASL due to binary indexing, when compared to using the full Poisson-distributed term frequencies. Note that term frequencies of 0 are the same for both models.

7.6 $\mathcal{A}$, ASL, AND TRADITIONAL PERFORMANCE MEASURES

The ASL measure is computed from the mean location of relevant documents. Other locations could be used, such as the median, the seventy-fifth percentile, or other arbitrary points in the distribution of f_{rel}. One can average the 25%, 50%, and 75% recall points and one can similarly average $\mathcal{A}$ values computed at these points rather than at μ_r.

Given a point x in the distribution of relevant documents, we find that

$$\mathcal{A}(x, all) = \int_x^{\infty} \Pr(y)\,dy,$$

where Y is the random variable describing term frequencies in all documents. In the case where ASL is to measured, $x = \mu_r$, the mean for the distribution of relevant documents. Other stopping points may represent different stopping rules used by the searcher (Kraft & Lee, 1979). Stopping at μ_r can be interpreted as *satiation* with the relevant documents when reaching this point.

Traditional Performance Measures

The survival function may be used in the definition of traditional performance measures.

Theorem 7.9 (Recall) *The recall at point x, where x is a feature value, may be computed as*

$$recall_x = \mathcal{S}(x, rel).$$

This is the percent of the relevant documents retrieved up to point x. If f_{rel} is a symmetrical distribution, $\mathcal{S}(\mu_r, rel) = .5$

The precision of a retrieval system at a given recall level x is proportional to the ratio of the number of relevant documents retrieved at that point to the total number of documents retrieved.

Theorem 7.10 (Precision) *The precision at point x is*

$$precision_x = \frac{\mathcal{S}(x, rel)}{\mathcal{S}(x, all)} g. \tag{7.14}$$

This is the expected number of relevant documents retrieved up to point x divided by the expected number of documents retrieved up to point x. The variable g is the generality measure, $||\{rel\}||/||\{all\}||$, Equation 4.1. Generality is used here to compensate for unequal numbers of relevant and all documents.

Corollary 7.5 (Precision as a function of recall)

$$precision_x = \frac{recall_x}{\mathcal{S}(x, all)} g.$$

This shows the functional relationship between precision and recall.

Note that one can plot a precision-recall graph based on the above formulae to produce a set of (recall, precision) points for each x value, given the generality g and the two distribution functions f_{rel} and f_{all}. One merely moves from $x = 1$ down to $x = 0$, computing the precision and recall values at points sufficiently close together for the level of accuracy desired. One may also numerically solve for one measure from the other by noting that the inverse function of the CDF is the quantile function (found in many computer packages having the CDF), and the inverse of the survival function is one minus the quantile function.

E Measure

Using these measures of precision and recall, one may compute E at point x for optimal ordering thus:

$$\begin{aligned} E &= 1 - \frac{2}{\frac{1}{P} + \frac{1}{R}} \\ &= 1 - \frac{2}{\frac{1}{g}\frac{\mathcal{S}(x, all)}{\mathcal{S}(x, rel)} + \frac{1}{\mathcal{S}(x, rel)}} \\ &= 1 - \frac{2}{\frac{1 + \mathcal{S}(x, all)/g}{\mathcal{S}(x, rel)}} \\ &= 1 - \frac{2\mathcal{S}(x, rel)}{1 + \mathcal{S}(x, all)/g}. \end{aligned} \tag{7.15}$$

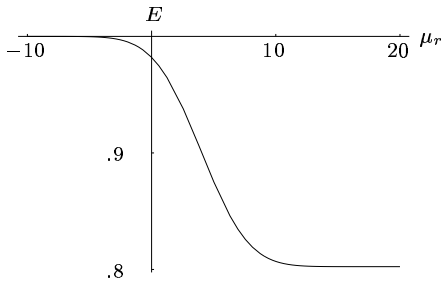

Figure 7.3. E performance.

One can compute E as 1 minus the ratio of twice the expected number of relevant documents retrieved up to that point in a search to the sum of the total number of relevant documents and the expected number of documents retrieved up to that point. The effect of generality on the E measure may also be understood using this equation, with an increase in generality decreasing (improving) the E value, with other factors being held constant. Somewhat easier to interpret is the F measure, $F = 1 - E$, which has higher values representing better retrieval results.

Optimal E

The optimal value of x can be computed analytically for some distributions and numerically for most distributions by taking the derivative of Equation 7.15, solving this for 0, and then solving for the other parameters.

One way of studying filtering performance is to examine how the E measure is affected by features distributed in a manner described by the normal distribution. Figure 7.3 illustrates the values of E when the variance of both distributions is set to 3, $\mu = 0$, $x = 4$, and $g = .01$. When the discrimination power is increased, by increasing μ_r (moving to the right on the graph), the E value drops. Interestingly, the performance described by the S shaped curve appears to level off at a point, and there is essentially a floor below which E does not drop while increasing the discrimination of the term, holding the other parameters constant.

Holding the variance at 3 and setting $\mu_r = 3$ and $\mu = 0$, more complex relationships producing the E value can be studied. Using numerical techniques, for example, for $g = .01$, the optimal E of .9026 occurs when filtering documents with feature frequency of 6.25. Allowing the generality to vary, along with the various document points, one can see in Figure 7.4 how the E values vary as generality and x points vary, and one can see the optimal values for E. Interestingly E seems to drop sharply when x approaches its optimal value from above, but E doesn't increase much as x decreases beyond this point. Generality clearly has a large impact on E, with higher values for generality producing lower E values than are obtained at the optimal x value for lower values of generality.

EXERCISE 7.8. [*Difficult*] What is the expected value of the E measure?

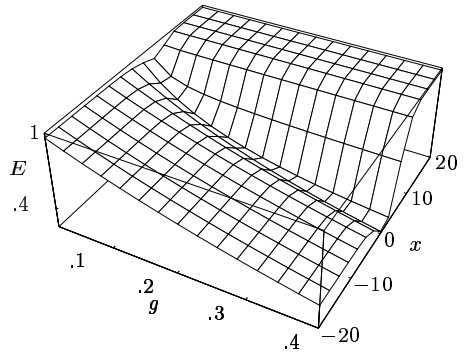

Figure 7.4. E as g and x vary, including the optimal E and its neighboring E values.

EXERCISE 7.9. [*Research Problem*] What are the variances of the precision and recall measures?

EXERCISE 7.10. [*Research Problem*] What is the variance of the E measure? What is the variance of the optimal E measure?

EXERCISE 7.11. [*Research Problem*] Blair and Maron (1985) note that "the amount of search effort required to obtain the same recall level increases as the database increases, often at a faster rate than the increase in database size." Assuming a single term query, a fixed number of relevant documents, and a fixed p, and for an arbitrary recall level, analytically determine some circumstances under which this does or doesn't hold. Treat ASL as linearly related to "search effort."

7.7 THE EFFECT OF PARAMETER VALUES

When computing the ASL using the methods developed above, each particular set of parameter values will have its own associated ASL value. In our model of text filtering performance, the $\mathcal{A}$ component should always use the exact values that exist in the system, that is, they should reflect perfect knowledge. On the other hand, $\mathcal{Q}$ reflects the quality of the ranking algorithm and incorporates the knowledge held by the user and the ranking algorithm as well as the true values. The parameter $\mathcal{Q}$ captures the relationship between the performance with perfect knowledge, provided by the optimal ranking procedure $\mathcal{R}_o$, and also the sub-optimal knowledge held by the user.

The effect of different parameter values may be seen in Figure 7.5, which shows the ASL for the situation illustrated earlier in Figure 6.2. With parameter q fixed at .3, the ASL varies as parameter p and the weight of knowledge, or confidence in our knowledge about p, is allowed to change. The confidence is computed as the conceptual number of documents retrieved. Parameter $\mathcal{Q}$ is computed as in Equation 6.16. Note that as the confidence about p increases, the ASL improves (decreases).

Ignorance may be reflected in the choice of a sub-optimal ranking algorithm, such as IDF weighting. This is not meant to criticize the use of IDF weighting, which

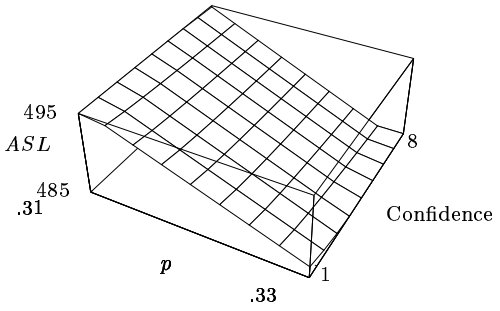

Figure 7.5. The ASL with $q = .3$ as p and the conceptual strength of our knowledge about p varies.

is appropriate in circumstances where the user's lack of knowledge about some parameter values would produce worse performance using other models. Ignorance of parameter values is the second way in which sub-optimal knowledge enters $\mathcal{Q}$.

The Effect of Feedback

The difference between performance with two different levels of relevance feedback may be modeled as $\Delta ASL = ASL_j - ASL_i$, the difference between the performance produced given two different knowledge sets, i and j, representing two different stages in the search, with j assumed to be a later stage than i. The value $\mathcal{A}$ is constant regardless of the human knowledge incorporated into $\mathcal{Q}$.

Theorem 7.11 (Condition for the effectiveness of relevance feedback) *When $\mathcal{A} \leq 1/2$, relevance feedback that changes the state of knowledge from i to j improves performance as measured by ASL only when $\mathcal{Q}_i < \mathcal{Q}_j$.*

Using either i or j as a subscript to indicate whether the $\mathcal{Q}$ is from the i or the j state of nature, we find that

$$\begin{aligned}
\Delta ASL &= ASL_j - ASL_i \\
&= N\left[\mathcal{Q}_j\mathcal{A} + \overline{\mathcal{Q}}_j\overline{\mathcal{A}} - \mathcal{Q}_i\mathcal{A} - \overline{\mathcal{Q}}_i\overline{\mathcal{A}}\right] \\
&= (\mathcal{Q}_j - \mathcal{Q}_i)\mathcal{A} + (\overline{\mathcal{Q}}_j - \overline{\mathcal{Q}}_i)\overline{\mathcal{A}} \\
&= (\mathcal{Q}_j - \mathcal{Q}_i)2(\mathcal{A} - \frac{1}{2}).
\end{aligned}$$

Feedback will improve performance (decrease ASL) when ΔASL is negative, that is, the ASL changes by the amount ΔASL. Performance will always increase when

1. $\mathcal{Q}_j > \mathcal{Q}_i$ when $\mathcal{A} < 1/2$, or
2. $\mathcal{Q}_j < \mathcal{Q}_i$ when $\mathcal{A} > 1/2$.

Clearly, the first condition is what is found when the term is a positive discriminator and feedback increases $\mathcal{Q}$.

EXERCISE 7.12. [*Research Problem*] The rate of ASL increase as $\mathcal{Q}$ increases is $N(2\mathcal{A}-1)$. Thus, for $\mathcal{A} < 1/2$, ASL decreases as $\mathcal{Q}$ increases. For a binary model, where knowledge about p for relevant documents is beta distributed, what is the rate of increase in the ASL as the sample size increases?

7.8 SUMMARY

Text filtering and retrieval performance may be computed after documents are retrieved, based on the results obtained from actual searches, or performance may be predicted beforehand based on probabilistic considerations. Simulations in the case of single term queries are unnecessary; the expected results may be computed analytically using simple methods. To use these techniques, more research is needed on the exact form of probabilistic distributions describing term frequencies in documents. Inconsistencies in experimental results, where the best method with one database is not the best with another, is due in part to the differences in these distributions. The design of a system should incorporate what is known about the distribution of term frequencies when designing the ranking algorithm and when predicting the performance.

In the following chapters, we examine the multivariate techniques necessary to understand multi-term queries and more complicated relationships, including the role of natural language processing in text retrieval and filtering.