6 THE QUALITY OF A RANKING METHOD

There is [a]... great cause of the little advancement of the sciences, which is this: it is impossible to advance properly in the course when the goal is not properly fixed.
—Francis Bacon, 1620

6.1 INTRODUCTION

The most commonly used performance measures represent the characteristics of a search up to a certain point in the retrieval process. These measures include precision, recall, and Van Rijsbergen's and Shaw's E measure. Swets' E measure, on the other hand, addresses the performance of a particular ranking based on the characteristics of the distributions for relevant and non-relevant documents. It is therefore independent of the stopping point chosen in a particular search. The relationship between ordered sequences, such as lists of ranked documents, may be studied using a number of techniques (Levenstein, 1966; Myers & Miller, 1988), ranging from sophisticated cost-based methods to counting the number of swaps necessary in sorting one sequence so that it becomes the other sequence. How best to study the relationships between sequences is an open question.

Below, we take a different approach, *studying ranking formulae themselves* to determine when and how they act optimally. This form of measure may be used to compare retrieval performance for different ranking methods under a variety of different conditions. Most importantly for our purposes, it may also be used in the analytic determination of the ASL for a given set of conditions.

Below, we develop a measure of the degree of optimality of a ranking procedure, whether a simple ranking formula or a complex formula representing a mixture of other, simpler formulae. This technique is then applied to a number of term weight-

ing systems for the case of single terms. By examining the circumstances under which a ranking algorithm is optimal, combined with considering how knowledge, ignorance, and misinformation affect performance, we will be able to make general statements about filtering performance. When these techniques are combined to produce methods for computing the ASL for optimal and sub-optimal ranking in Chapter 6, numeric performance figures are generated without using the simulations that dominate traditional information retrieval research. In addition, using these measures and methods will lead to the additional understanding of *why* one method is better than another.

Optimal Rankings

Assume that the set of available documents $\mathbf{v}$ is weakly ordered by a ranking algorithm $\mathcal{R}_x$ and there are $N = |\mathbf{v}|$ elements in the set of documents $\mathbf{v}$. The ranking $\mathcal{R}_x$ produces an ordering of the documents $\mathbf{v}_y = d_1 \succeq d_2 \succeq \cdots \succeq d_N$, that is, d_1 is weakly ordered before d_2 which is weakly ordered before d_3, and so forth to d_N. When studying the ranking of a set of documents and the resulting performance, it may be necessary to compute $\Pr(\mathbf{v}_y|\mathcal{R}_x)$ the probability that one will obtain a particular document ordering $\mathbf{v}_y$ when ranking procedure $\mathcal{R}_x$ is used.

Definition 6.1 (Probability of the optimal ranking) *One ranked set of documents is the best, or optimal, denoted as* $\mathbf{v}_o$*, and the probability that we will have this when the optimal ranking* ($\mathcal{R}_o$) *procedure is used is* $\Pr(\mathbf{v}_o|\mathcal{R}_o) = 1$.

The probability $\Pr(\mathbf{v}_y)$ may be understood as the proportion of the series of queries meeting the assumptions of the model that have ordering y, given the probabilities associated with the query. The probability may represent the subjective probability that an individual believes that the ranking will be as given.

Measuring the Degree of Optimality

Sometimes it may be helpful to measure the degree of optimality for a ranking procedure $\mathcal{R}$, the probability the ranking is the same as that provided by optimal ranking ($\mathcal{R}_o$). This, combined with other performance factors, can allow us to compute expected performance in the single term case.

Definition 6.2 (Degree of optimality $\mathcal{Q}$) *The probability that ranking methods* $\mathcal{R}_x$ *and* $\mathcal{R}_o$ *produce the same ranking, computed over the set of rankings, is*

$$\mathcal{Q}_x = \Pr(\mathcal{R}_o, \mathcal{R}_x) = \sum_{i \in Perm(\mathbf{v})} \Pr(\mathbf{v}_i, \mathcal{R}_x, \mathcal{R}_o), \tag{6.1}$$

where $Perm(\mathbf{v})$ *is the set of distinct weak orderings possible, rearrangements of the document profiles in vector* $\mathbf{v}$.

In the case of using a single term in queries and documents, two possible orders exist: placing documents with the feature first and the documents without the feature second, or the reverse. Note that for the case where there is more than one term or more than two possible orders, $\mathcal{Q}_x$ denotes the probability of obtaining document ordering x.

We consider here cases where there are documents with identical profiles. $\mathcal{Q}$ is normed and has the value of 1 when ranking is optimal and 0 in the worst-case. It

is used to average joint probabilities over the two sets of ordered documents, where, $1 \succ 0$ and $0 \succeq 1$.

The simplest environment in which we may examine the degree of optimality exists when there is only a single term in the query. When two documents have the same profile, and all characteristics being considered by the ranking procedure are the same, any two deterministic methods will provide the same ranking. When there is a difference between the documents, the ranking procedure selected determines which ranking is to be preferred. To compute $\mathcal{Q}$ for this single term case, consider the probability that the ranking of documents for binary feature d is the same for two ranking functions (in this case the ranking function $\mathcal{R}_x$ and the optimal ranking function $\mathcal{R}_o$), in the case where the document with a 1 is ranked ahead of or the same as a document with profile 0 (denoted here as $1 \succeq 0$) and the case where a document with a 0 is ranked ahead of a document with a 1 (denoted here as $1 \prec 0$). For purposes of notational simplicity, we consider the documents to have been re-parameterized so that the feature value 1 is assigned to the feature characteristic that produces the highest discrimination value under optimal ranking.

Theorem 6.1 (Degree of optimality for a term) *When there is a single term in the query, the probability that ranking method $\mathcal{R}_x$ is optimal is,*

$$\mathcal{Q}_x = \Pr(\mathcal{R}_x, \mathcal{R}_o, 1 \succ 0) + \Pr(\mathcal{R}_x, \mathcal{R}_o, 1 \preceq 0), \tag{6.2}$$

the probability that the ranking method in question and the optimal ranking have the same ordering.

When there is only a single, binary feature being considered (which we assume below unless specified otherwise), there are only two orderings possible: documents with the feature ordered before those without the feature ($1 \succ 0$) and documents without the feature ordered before or the same as documents with the feature ($1 \preceq 0$). Because $\mathcal{Q}_x$ is the degree of overlap between ranking $\mathcal{R}_x$ and the optimal ranking $\mathcal{R}_o$, $\mathcal{Q}_x = \Pr(\mathcal{R}_o, \mathcal{R}_x)$.

Parameter Orderings

Computing $\mathcal{Q}$ for different ranking methods requires in some cases that estimates be made of the probability that one parameter exceeds or is less than a second parameter. The probability that one parameter is greater than another, or that it is within a certain range, may be computed from historical data. For example, the data in Table 6.1 has $p \geq t$ for three of the four queries, thus $\Pr(p \geq t) = 3/4$, while $p < .5$ one quarter of the time (query 4,) yielding a probability $\Pr(p < .5) = 1/4$. Given historical data, and knowing the formula for $\mathcal{Q}$ for a specific ranking method, one can determine which ranking method will perform best for a given set of parameters and thus a set of database and query pairs.

This approach to computing $\mathcal{Q}$ is used when one desires to historically analyze a set of executed queries. A Bayesian method can predict $\mathcal{Q}$ for the future, using the distribution of values for a given parameter. This model allows us to make claims about future performance, given all available current knowledge.

Table 6.1. Computing the probabilities for $\mathcal{Q}$ for 4 single term queries and 6 documents. Here, $p = \Pr(d = 1|rel)$ and $t = \Pr(d = 1)$.

Document	*Relevance for Queries*			
Profiles	1	2	3	4
0	N	N	N	N
0	Y	Y	N	N
0	N	Y	Y	Y
1	N	N	Y	N
1	Y	Y	Y	N
1	Y	Y	Y	N
p:	2/3	1/2	3/4	0
t:	1/2	1/2	1/2	1/2

6.2 DEGREE OF OPTIMALITY FOR SPECIFIC MODELS

The retrieval status value for a document in which the single term has weight or discrimination value w and term frequency d is denoted as the product dw. We find that for binary term frequencies, documents with feature frequency 1 are ordered before documents with a 0, that is, $(1 \succ 0)$, if and only if $1w > 0w$, that is, when w is greater than zero. One can similarly compute the probability that $(1 \preceq 0)$ as the probability that $1w \leq 0w$, which is the probability that w is less than or equal to 0.

In several sections below, specific retrieval algorithms will be studied to determine their degree of optimality. These results can be used in several different ways. One is as a criteria for determining when ranking is optimal. We will find, for example, that IDF ranking, for a single term, produces optimal ranking when $p > t$. One can also compute $\mathcal{Q}$ as representing the average degree of optimality, given a large number of queries. Thus, the IDF weighting is optimal in the those cases where $p > t$ and the probability it is optimal is $\Pr(p > t)$. A third use of $\mathcal{Q}$ addresses our subjective belief about the degree of optimality for a single query, given available knowledge about the parameters of the model. For example, a professional, based on their own expertise, may believe that for an individual search using the IDF weighting, the probability that the search will be optimal is 3/4, that is, that p will be greater than t about 3/4 of the time for queries such as the one in question.

Best-case Ranking

The optimal ranking of documents is obtained by ordering documents by decreasing expected precision, which is the same as ranking them by the probability that they are relevant, $\Pr(rel|d)$. Ranking documents by their probability of relevance, given that they have certain features, places documents with higher probabilities of relevance ahead of documents with lower probabilities of relevance. We assume for the next several sections that the probabilities being considered are point probabilities.

This best-case ranking is the same as the ranking obtained with

$$\Pr(rel|d) = \frac{\Pr(d|rel)\Pr(rel)}{\Pr(d)}.$$

Because the $\Pr(rel)$ component is constant for each document, it may be dropped without affecting ranking; thus, ranking documents by $\Pr(d|rel)/\Pr(d)$ produces the same ranking. Given binary features, the weight for a feature is

$$w = \log\left(\frac{p/(1-p)}{t/(1-t)}\right), \tag{6.3}$$

where $p = \Pr(d = 1|rel)$ and $t = \Pr(d = 1)$.

To apply Equation 6.1 we must compute the probability that a document with a feature value of 1 is ranked ahead of or equal to a feature with value 0:

$$\begin{aligned} 1w &> 0w \\ \log\left(\frac{p/(1-p)}{t/(1-t)}\right) &> 0. \end{aligned} \tag{6.4}$$

This is the case only when $p > t$. When this condition is met, the number in the left hand side of Equation 6.4 is positive.

Similarly, we compute the probability that a document with a feature value of 0 is ranked ahead of a feature with value 1 as

$$\begin{aligned} 1w &\leq 0w \\ \log\left(\frac{p/(1-p)}{t/(1-t)}\right) &\leq 0. \end{aligned}$$

We find that this is the case when $p \leq t$.

The probability that optimal ranking is obtained when the ranking function $\mathcal{R}$ is $\mathcal{R}_o$ is

$$\begin{aligned} \mathcal{Q}_B(p,t) &= \Pr(p > t, p > t) + \Pr(p \leq t, p \leq t) \\ &= 1. \end{aligned} \tag{6.5}$$

Worst-case Performance

The worst-case ordering is that obtained when documents with different term frequencies are always ranked the opposite by $\mathcal{R}$ than by the optimal ranking $\mathcal{R}_o$. This requires that we compute the following probabilities: $\Pr(\mathcal{R}_{o,1\succ 0}, \mathcal{R}_{x,1\preceq 0})$ and $\Pr(\mathcal{R}_{o,1\preceq 0}, \mathcal{R}_{x,1\succ 0})$.

Using the results derived in Equation 6.5 for the probability of optimality for the best case ranking, this can be seen to be

$$\begin{aligned} \mathcal{Q}_W(p,t) &= \Pr(p > t, p \leq t) + \Pr(p \leq t, p > t) \\ &= 0. \end{aligned} \tag{6.6}$$

This problem may be viewed as suggesting the ranking of documents by $1-\Pr(rel|d) = \neg\Pr(rel|d)$, where $\Pr(rel|d)$ is the optimal weighting, if we wish to achieve worst-case

performance. The probabilities obtained in Equation 6.6 may be treated as

$$\begin{aligned} \mathcal{Q}_W(p,t) &= \Pr(p > t, \neg(p > t)) + \Pr(p \leq t, \neg(p \leq t)) \\ &= 0. \end{aligned}$$

For notational purposes, the joint probabilities used in describing the $\mathcal{Q}$ values will always list the optimal condition first, followed by the condition for the model whose probability of optimality is being computed.

Random Ranking

The random ranking of documents could be expected to give optimal ranking half of the time and worst-case ranking the other half. Using Equation 6.2, one can compute

$$\mathcal{Q}_R(p,t) = \Pr\left(p > t, \frac{p > t}{2}\right) + \Pr\left(p \leq t, \frac{p \leq t}{2}\right) = \frac{1}{2} \tag{6.7}$$

where the fractions $\Pr((p > t)/2)$ and $\Pr((p \leq t)/2)$ represent the portion (one half) of each probabilistic distribution that is to be combined in the joint probability. Exactly which documents are used (from which this random half is taken) is not important for purposes here.

IDF

Inverse document frequency (IDF) weighting is an effective feature weighting system when no information about relevant documents is available, other than the fact that a term is or is not in the query (Sparck Jones, 1972; Yu & Salton, 1977; Croft & Harper, 1979). Using as a weight

$$w = -\log t,$$

we find that the weight increases as t decreases.

One finds that $1 \succ 0$ if and only if $-\log t > 0$, that is, when $t > 0$, which is the case for all valid probabilities, all values of t. For the other ordering, $1 \preceq 0$ is found if and only if $-\log t \leq 0$, which never occurs, for any valid probability t.

Equation 6.2 may be used to compute

$$\begin{aligned} \mathcal{Q}_I(p,t) &= \Pr(p > t, t > 0) + \Pr(p \leq t, t \leq 0) \\ &= \Pr(p > t). \end{aligned} \tag{6.8}$$

P-Weighting

For illustrative purposes, let us consider weighting documents by

$$d \log \frac{p}{1-p}.$$

We will call this *p-weighting* and propose it for illustrative purposes. Documents are ranked so that $1 \succ 0$ when $\log((p/(1-p))/(t/(1-t))) > 0$ (from the best-case ordering) and $\log(p/(1-p)) > 0$ (from the p-weighting). These inequalities are met when $p > t$ and $p > .5$. The ordering $1 \preceq 0$ is obtained when $\log((p/(1-p))/(t/(1-t))) \leq 0$ and

$\log(p/(1-p)) \leq 0$, occurring when $p \leq t$ and $p \leq .5$. We can thus derive the $\mathcal{Q}$ value for p-weighting as

$$\mathcal{Q}_P = \Pr(p > t, p > .5) + \Pr(p \leq t, p \leq .5). \tag{6.9}$$

Example. Consider the set of $\{p, t\}$ pairs $\{.8, .2\}, \{.6, .3\}, \{.4, .45\}$, and$\{.6, .7\}$, characterizing 4 different searches. We note that for the first two parameter pairs, both $p > t$ and $p > .5$. For the third pair, $p \leq t$ and $p \leq .5$. The last $\{p, t\}$ pair clearly doesn't fall into either of these two categories. Since 3 of the 4 pairs would be included in one of the two probabilities summing to produce $\mathcal{Q}_P$ in Equation 6.9, we compute $\mathcal{Q}_P = 3/4$.

Decision Theoretic Weighting

The weighting provided by Equation 5.12 is possibly the most widely supported weighting for binary independent features. We find for ordering $1 \succ 0$ that

$$\log\left(\frac{p/(1-p)}{q/(1-q)}\right) > 0$$

only when $p > q$.

We find for ordering $1 \preceq 0$ that

$$\log\left(\frac{p/(1-p)}{q/(1-q)}\right) \leq 0$$

only when $p \leq q$.

We can thus calculate

$$\mathcal{Q}_D(p, t, q) \quad = \quad \Pr(p > t, p > q) + \Pr(p \leq t, p \leq q). \tag{6.10}$$

The variable t will always be a value between p and q, that is, either $p \geq t \geq q$ or $p \leq t \leq q$. The $\mathcal{Q}_D(p, t, q)$ value can be interpreted as 1 minus the probability of being in the gap between t and q. This gap will be very small for large datasets where almost all documents are non-relevant and the value of t approaches the value of q.

Coordination Level Matching

Coordination level matching (Van Rijsbergen, 1986; Losee, 1987) suggests the retrieval of documents arranged in decreasing order of the number of terms in common with the query, referred to as the *coordination level*. Used in the early Cranfield indexing and retrieval tests as a weighting method, it computes the document weight as the summation of the binary term frequency times a weight constant for all terms, or

$$\sum_{i=1}^{n} d_i c. \tag{6.11}$$

When this is implemented as a series of Boolean expressions, first retrieving documents with all terms, then retrieving documents with all but one term, and so on until

Table 6.2. Comparing ranking methods.

$\mathcal{R}$		$\mathcal{Q}$
$\mathcal{R}_o$ = Best-case	$\mathcal{Q}_B$ =	$\Pr(p > t, p > t) + \Pr(p \leq t, p \leq t)$
	=	1
Random	$\mathcal{Q}_R$ =	$(\Pr(p > t) + \Pr(p \leq t))/2$
	=	$1/2$
Worst-case	$\mathcal{Q}_W$ =	$\Pr(p > t, p \leq t) + \Pr(p \leq t, p > t)$
	=	0
Dec Theo	$\mathcal{Q}_D$ =	$\Pr(p > t, p > q) + \Pr(p \leq t, p \leq q)$
	=	$(\Pr(p > max(t, q)) + \Pr(p \leq min(t, q)))$
IDF	$\mathcal{Q}_I$ =	$\Pr(p > t, t > 0) + \Pr(p \leq t, t \leq 0)$
	=	$\Pr(p > t)$
CLM	$\mathcal{Q}_C$ =	$\Pr(p > t)$

documents with no terms are retrieved, it is referred to as quorum level searching (Salton, Wong, & Yu, 1976).

A document with frequency 1 is retrieved ahead of a document with feature frequency 0 if, and only if, $c > 0$, that is, if c is positive. Similarly, we have the ordering $1 \preceq 0$ if and only if $c \leq 0$. If we assume that c is always positive for query terms,

$$\begin{aligned} \mathcal{Q}_C(p, t) &= \Pr(p > t, c > 0) + \Pr(p \leq t, c \leq 0) \\ &= \Pr(p > t). \end{aligned} \tag{6.12}$$

EXERCISE 6.1. [*Moderate*] How does one compute $\mathcal{Q}_D$, the degree of optimality under the decision theoretic model (Equation 6.10) assuming that perfect knowledge about p is beta distributed and that q is beta distributed?

Comparing Ranking Methods

Table 6.2 summarizes the $\mathcal{Q}$ values for several filtering models. Clearly the optimal ranking is best, with $\mathcal{Q}_B = 1$, and the worst is $\mathcal{Q}_W = 0$. In cases where the query term is a positive discriminator, which it usually is, IDF, CLM, and the decision theoretic models perform better than random but worse than optimal retrieval. One can see that as the difference between q and t approaches 0, $\mathcal{Q}_D$ approaches 1.

EXERCISE 6.2. [*Moderate*] What would be $\mathcal{Q}$ for the ranking of documents by $d \log q/(1-q)$? How does this compare to what is obtained for the IDF measure?

EXERCISE 6.3. [*Moderate*] Under what conditions would $\mathcal{Q}_D$ be lowest? What value would $\mathcal{Q}_D$ approach under these circumstances?

EXERCISE 6.4. [*Moderate*] Given the ranking formula $d \log(p - t)$, what is this ranking method's $\mathcal{Q}$ value?

6.3 RELEVANCE FEEDBACK

Changing the values of parameter estimates changes the degree of optimality for a ranking procedure. Incorporating relevance feedback essentially changes the operat-

ing characteristics of the ranking algorithm $\mathcal{R}$ by modifying the parameters of the prior distribution describing our knowledge about the parameters describing term frequencies in documents.

Relevance feedback is user supplied information about the relevance of documents retrieved by the system. Some assumptions are made below, including that relevance is binary, that is, a document is relevant or it is not relevant, and that the judgments that are provided are binary, that is, they are not probabilistic judgments, such as believing that there is a thirty percent chance that a document is relevant and, conversely, a seventy percent chance that it is not relevant. The requirement that relevance and relevance judgments are binary is dropped near the end of this chapter.

To estimate $\mathcal{Q}$ it is necessary to estimate probabilities for p, t, and q. While these probabilities may be point probabilities such that either $p \geq t \geq q$ or the reverse holds, the probabilities may also be estimated as distributions. In some situations, our knowledge about these parameters for binary distributions may best be represented by beta distributions, describing the probability that the parameter p of the binary distribution has a certain value. We often can treat t as a point probability, since we can always know precisely what percent of documents have a particular term. The parameters p and q are not known before hand, although for large databases we may often make the approximation $q \approx t$ and thus treat q as a point probability.

The parameter values for the beta distribution may be learned through relevance feedback provided by the searcher. As the user provides judgments of "relevant" or "non-relevant" about retrieved documents, the parameters increase in value and the variance of the distribution decreases, increasing the accuracy of the estimate provided by the distribution.

When computing $\mathcal{Q}$ for different models, the probability that one parameter was above or below another was encountered. This may be implemented by using a form of the incomplete beta distribution to represent our knowledge about a parameter such as p:

$$\beta_x^y(a,b) = \frac{\Gamma(a+b)}{\Gamma(b)\Gamma(a)} \int_x^y p^{a-1}(1-p)^{b-1}\, dp.$$

This allows us to compute the portion of a distribution within a certain range, the probability that one parameter, for example, is greater than another. In this case, the distribution represents the probability that the random variable lies within the range from x to y.

As an example, consider the application of limited knowledge to p-weighting where our knowledge about p is represented by a beta distribution:

$$\begin{aligned} \mathcal{Q}_P &= \Pr(p > t, p > .5) + \Pr(p \leq t, p \leq .5) \\ &= min\left(\beta_t^1(\mu \to p, \sigma^2 \to 0), \beta_{.5}^1(a,b)\right) \\ &\quad + min\left(\beta_0^t(\mu \to p, \sigma^2 \to 0), \beta_0^{.5}(a,b)\right). \end{aligned} \tag{6.13}$$

We treat optimal ranking as though p were described by a beta distribution here with a mean approaching the value p ($\mu \to p$) (if p is a point probability and relevance feedback improves performance) and with the variance approaching 0 ($\sigma^2 \to 0$). This approximates a point estimate at p. The second beta distribution, with parameters a and b, describes the estimate for p held by the user of the model.

The degree of optimality $\mathcal{Q}_P$ is at its maximum, that is, p-weighting approaches optimal ranking, when the beta distribution describing the user's knowledge matches the distribution describing the p for optimal ranking. This is the case where the user has perfect knowledge.

The minimum value for $\mathcal{Q}_P$ occurs when the mean for the beta distribution describing p in the optimal model (the true value of p) is within the gap between t and .5 and the distribution peaks within this area. This peaking occurs when there is a great deal of misleading relevance feedback about the characteristics of relevant documents.

6.4 COMPARING RANKING UNDER DIFFERENT LEVELS OF KNOWLEDGE

The differences in ranking quality may be computed by examining the difference between the $\mathcal{Q}$ values. The difference between methods i and j may be computed as $\Delta\mathcal{Q}_{i,j} = \mathcal{Q}_i - \mathcal{Q}_j$.

The difference between $\mathcal{Q}$ for decision theoretic and IDF ranking may be computed as

$$\Delta\mathcal{Q}_{DI} = \Pr(p > t, p > q) + \Pr(p \leq t, p \leq q) - \Pr(p > t).$$

Remember that $\mathcal{Q}_D$ is the area in the distribution describing p outside the range from q to t, and p is described by a distribution.

If the term is a positive or neutral discriminator, then it will be the case that $p \geq t \geq q$. If the knowledge about p is described by a distribution then $\Pr(p > t)$ is the cumulation of the function (distribution) from t up to 1 while $\Pr(p > q)$ is the cumulation of the function from q up to 1. Since $t \geq q$, the joint probability $\Pr(p > t, p > q)$ is the cumulation from t up to 1. Using similar techniques, $\Pr(p \leq t, p \leq q)$ is the cumulation from the 0 point in the function up to q. Similar techniques are used when the term is a negative discriminator, that is, $p < t < q$. When the term is a positive discriminator, then $\mathcal{Q}_I$ is the area from the high end of the distribution down to t. In this case, $\Delta\mathcal{Q}_{DI}$ will be the area remaining, from below q to the low end of the distribution.

When the term is a negative discriminator, then $\mathcal{Q}_I$ is the area from the high end of the distribution down to t, which will be lower than q. Thus, the area will include both from the high end of the distribution down to q and the gap from q down to t. In this case, $\Delta\mathcal{Q}_{DI}$ will be the area below t minus the gap between q and t. If the area in the gap is larger than the area below t, then one would expect IDF weighting to outperform decision theoretic weighting, i.e., $\Delta\mathcal{Q}_{DI} < 0$. If the distribution is concentrated below t, however, the area under the distribution below t will be larger than that in the gap and decision theoretic methods will be expected to outperform IDF methods, i.e., $\Delta\mathcal{Q}_{DI} > 0$.

Example: Using $\mathcal{Q}$ to Evaluate Parameter Estimates

One application of $\mathcal{Q}$ is in the evaluation of parameter estimation techniques. Robertson and Sparck-Jones (1976) have suggested that a suitable estimate for p values when

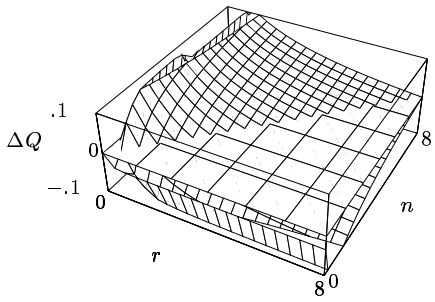

Figure 6.1. The difference in $\mathcal{Q}$ values when using the Shaw minus the Robertson-Sparck-Jones estimates. Positive values occur in the region where Shaw's estimate is superior (the smaller grid squares are above the large squared zero plane,) while negative values are regions where RSJ is superior (the larger grid squares are above the small squares). Shaw's method is superior in the region where the number of non-relevant documents retrieved exceeds the number of relevant documents retrieved.

no other information is available is

$$p = \frac{r + \frac{1}{2}}{r + n + 1}, \tag{6.14}$$

where r is the number of relevant documents retrieved with the term and n is the number of relevant documents retrieved without the term. This form of estimation has a long statistical history in a variety of applications. It can be seen as consistent with LaPlace's principle of succession (Equation 2.12) and the Bayesian models developed in Chapter 2 and 3, where the $1/2$ and 1 represent parameters consistent with non-informative prior information.

Shaw (1995) has suggested a second initial estimate for p. For most values, $p = r/(n + r)$. When r would be 1, p is computed as $p = 1 - 1/N^2$ where N may be the number of documents in the database. When r would be 0, p is set to $1/N^2$.

These specific formulae for estimating parameters may be compared by computing the $\mathcal{Q}$ value for each one and then comparing them. This is done in Figure 6.1 for a range of r and n values. Here, N is set to 100.

This analysis assumes that the prior information about p is beta distributed and that ordering is otherwise optimal. In many situations these assumptions are not met, but making simplifying assumptions allows us to roughly compare the two estimation techniques. If these assumptions hold, it is clear from the figure that the Shaw method is superior as an initial estimate over one range of data values and the Robertson-Sparck-Jones method is better in another region. This is not to argue that one of these estimates is universally or conclusively better or worse than another method. Instead, we have suggested that a problem such as this can be addressed analytically. However, there are still assumptions in the analysis that need to be evaluated through data collection and improved modeling.

EXERCISE 6.5. [*Research Problem*] Under what conditions is the RSJ (Robertson & Sparck Jones, 1976) method for estimating parameters, Equation 6.14, better than random?

6.5 A GENERAL MODEL OF RANKING PERFORMANCE

A more general model of $\mathcal{Q}$ may be developed beginning with the work above. We denote the distribution describing perfect knowledge about p as $g(x)$, where x is a probabilistic parameter such as p.

Parameter p as a Point Probability

We may define $\mathcal{Q}$ when our knowledge of p is treated as a point probability, as:

$$\mathcal{Q} = \Pr(\mu_r > \mu, X) + \Pr(\mu_r \leq \mu, Y), \tag{6.15}$$

where X and Y represent ranking specific conditions, and μ, μ_r, and μ_n represent the expected frequency for all documents, relevant documents, and non-relevant documents respectively. Under the binary model, $\mu = t$, $\mu_r = p$, and $\mu_n = q$. As an example of ranking specific conditions, under best-case ranking $X = \mu_r > \mu$ and $Y = \mu_r \leq \mu$. For worst-case ranking, $X = \mu_r < \mu$ and $Y = \mu_r \geq \mu$.

Parameter p is Described by a Distribution

When our knowledge about p is perfect and is represented by a distribution g rather than a point probability p, then

$$\begin{aligned} \mathcal{Q} &= min\left(\int_{\mu}^{\infty} g(z)\,dz, \int_{x}^{\infty} g(z)\,dz\right) \\ &\quad + min\left(\int_{0}^{\mu} g(z)\,dz, \int_{0}^{y} g(z)\,dz\right) \\ &= 1 - \int_{min(\mu,x)}^{max(\mu,y)} g(z)\,dz, \end{aligned} \tag{6.16}$$

where x and y represent values specific to the model being considered. For example, the optimal (best-case) binary retrieval model may be implemented in the general model of $\mathcal{Q}$ by setting g to the beta distribution, with $x = \mu$ and $y = \mu$, while with the decision theoretic model, $x = \mu_n$ and $y = \mu_n$ ($\mu_n = q$).

An example of the application of this method to computing $\mathcal{Q}_D$ is given in Figure 6.2. As the amount of knowledge about p increases, $\mathcal{Q}_D$ varies. We set $q = .3$, and t is varied to be consistent with the corresponding value for p.

Inaccurate or Incomplete Knowledge about p

When our knowledge about p is imperfect and is represented by a distribution $h(x)$, and p is exactly distributed as described by the distribution $g(x)$, then we may

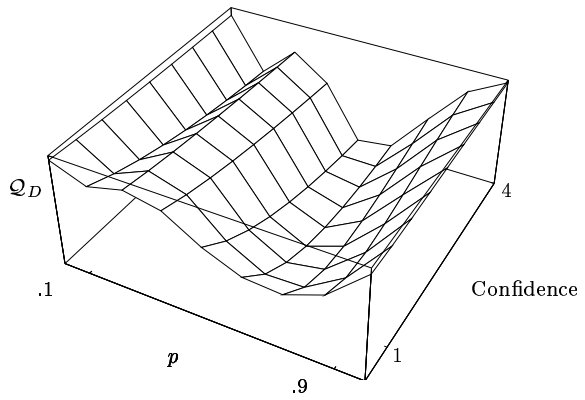

Figure 6.2. $\mathcal{Q}_D$ as p varies over the range of possible values and as the amount of knowledge increases (represented by an increased value on the bottom right axis). Confidence represents the conceptual number of documents retrieved.

compute:

$$\begin{aligned} \mathcal{Q} &= min\left(\int_{\mu}^{\infty} g(z)\,dz, \int_{x}^{\infty} h(z)\,dz\right) \\ &\quad + min\left(\int_{0}^{\mu} g(z)\,dz, \int_{0}^{y} h(z)\,dz\right). \end{aligned} \tag{6.17}$$

This can be useful when we wish to evaluate the effect of assuming one distribution about p, $h(x)$, when another distribution is the correct one, $g(x)$. When $g(x) = h(x)$ for all x, then Equation 6.17 will be equivalent to Equation 6.16.

The beta distribution is often used to model binary parameters such as p but it is usually conceded that this is a simplifying approach. When the true distribution is an irregular distribution, perhaps not being easily described by any of the traditional parametric distributions, $g(x)$ would be this empirical distribution and $h(x)$ would be the estimate provided by a simplifying distribution used by the model, such as a beta or normal distribution.

Continuous Relevance

Using the methods developed earlier, a continuous form of relevance may be incorporated into the analytic model of text filtering performance. We assume that the user wants to retrieve documents with a relevance value of x or greater. Instead of using a function $g(z)$ to describe our knowledge about the distribution of p, we use the function $g(z, r)$ to describe knowledge about p_x, where r is a relevance value that has as its lowest value x and as its highest value 1, $0 \leq x \leq 1$. When one is computing the component $\int_{\mu}^{\infty} g(z)\,dz$ in Equation 6.17 (with a similar analysis for the other integrals over $g(z)$), continuous relevance may be incorporated through the use of

expressions such as

$$\int_x^1 \frac{\Pr(r)}{\int_x^1 \Pr(r')\,dr'} \int_\mu^\infty g(z,r)\,dz\,dr.$$

The computation is similar to Equation 6.17 but the integral incorporating $g()$ function is averaged over the relevance values (r) from x up to 1. Similar modifications can be made to other formulae computing $\mathcal{Q}$ to make them consistent with the assumption of continuous relevance. Full development of this model is left as an exercise below.

EXERCISE 6.6. [*Moderate*] Design an experiment to determine whether the mean term frequencies for the different relevance values are linear with regard to the level of relevance, or whether the nature of the relationship between relevance and term frequencies is non-linear.

EXERCISE 6.7. [*Research Problem*] Assume that term frequencies are binary and that ranking uses decision theoretic weighting. If underlying relevance is continuous but it is treated as binary, what is the probability that ranking is optimal? What would this be generally for any weighting formulae, not just for decision theoretic weighting?

EXERCISE 6.8. [*Research Problem*] How would one compute $\mathcal{Q}$ for the situation where term frequencies are Poisson distributed but those terms with frequency greater than 0 are treated as having the binary frequency 1 and those with term frequency 0 are treated as having binary term frequency 0?

6.6 EXISTENCE AND CONSTRUCTION OF RANKING PROCEDURES

While earlier developments started from specific document ranking systems and moved toward a more general model of $\mathcal{Q}$, generality can be obtained by beginning with $\mathcal{Q}$ and moving toward a ranking procedure, constructing at least one ranking procedure for every $\mathcal{Q}$.

Theorem 6.2 (Existence of ranking procedures for any $\mathcal{Q}$) *There is at least one single term ranking procedure that has a degree of quality $\mathcal{Q}$, where $0 \leq \mathcal{Q} \leq 1$. Thus, there are an infinite number of ranking procedures, given the infinite number of values between 0 and 1.*

Let us assume that p is distributed as described by continuous distribution $g(z)$. The probability that a ranking is optimal may be computed when $g(z)$ describes the distribution of knowledge about the distribution of term frequencies in relevant documents as Equation 6.16:

$$\mathcal{Q} = 1 - \int_{min(\mu,x)}^{max(\mu,y)} g(z)\,dz.$$

When x approaches the low feature frequency value (i.e., 0 for the binary term distribution) and y approaches the high feature frequency value (i.e. 1 for the binary term distribution), we find that $\mathcal{Q}$ approaches 0. Conversely, when x and y approach μ then $\mathcal{Q}$ approaches 1. Clearly, by choosing appropriate values for x and y, one can obtain any desired value for $\mathcal{Q}$

Algorithm to construct a single-term ranking algorithm. A single binary term ranking procedure parameterized to produce $\mathcal{Q}$ (assuming it represents a positive discriminator) as $\mathcal{Q} = \Pr(p > t, a > b) + \Pr(p \leq t, a \leq b)$ may be produced with the weighting

$$w = \log \frac{a/(1-a)}{b/(1-b)}, \tag{6.18}$$

where $0 \leq a \leq 1$ and $0 \leq b \leq 1$. We assume that a is described by a distribution, as is p. Arbitrarily set point value b to t, and assume that the term is parameterized to be a positive discriminator. Assume that $\Pr(a > b) = 1 - \Pr(a \leq b)$ and that $\Pr(p > t) = 1 - \Pr(p \leq t)$. When $\mathcal{Q}$ is better than random, set distribution variable a numerically so that

$$\Pr(a > b) = \Pr(p > t) - \mathcal{Q}. \tag{6.19}$$

Note that

$$\Pr(a \leq b) \geq \Pr(p \leq t). \tag{6.20}$$

Consider the case where the mean estimate for a is lower than for p, leading to sub-optimality and $\mathcal{Q} < 1$. In this case, the condition in Equation 6.20 is met for the distributions considered here. That portion of $\mathcal{Q}$ computed from components less that t is thus $\Pr(p \leq t)$. Because of Equations 6.19 and 6.20, $\mathcal{Q} = \Pr(a > t) + \Pr(p \leq t)$. Similar procedures can be constructed for other point or distributional assumptions about $\mathcal{Q}$, p, and t.

EXERCISE 6.9. [*Difficult*] Starting with the procedure above, construct a ranking formula with $\mathcal{Q} = 3/4$. Using a method of your own design, produce a ranking formula with $\mathcal{Q} = 3/4$. How do they differ? Why might one choose one over the other for a practical system?

6.7 SUMMARY

The ranking performance of text filtering and retrieval systems can be estimated from the parameters of term frequency distributions. The measures considered in this chapter may be used to either describe (after the fact) or to predict (before the fact) some of the characteristics and causes of the performance of document ranking systems. The use of these measures in predicting future performance will be described in the next chapter for the relatively simple case of single terms in queries. Later chapters will then examine the more complex problems associated with predicting the performance of multiple term queries and natural language phrases.