A Theory of Term Weighting Based on Exploratory Data Analysis

Warren R. Greiff
Computer Science Department
University of Massachusetts, Amherst
www.cs.umass.edu/~greiff/

Abstract Techniques of exploratory data analysis are used to study the weight of evidence that the occurrence of a query term provides in support of the hypothesis that a document is relevant to an information need. In particular, the relationship between the document frequency and the weight of evidence is investigated. A correlation between document frequency normalized by collection size and the mutual information between relevance and term occurrence is uncovered. This correlation is found to be robust across a variety of query sets and document collections. Based on this relationship, a theoretical explanation of the efficacy of inverse document frequency for term weighting is developed which differs in both style and content from theories previously put forth. The theory predicts that a "flattening" of *idf* at both low and high frequency should result in improved retrieval performance. This altered *idf* formulation is tested on all TREC query sets. Retrieval results corroborate the prediction of improved retrieval performance. In conclusion, we argue that exploratory data analysis can be a valuable tool for research whose goal is the development of an explanatory theory of information retrieval.

1 Introduction

In 1972, Spark Jones demonstrated that document frequency can be used effectively for the weighting of query terms [23]. Ever since, formulations of *inverse document frequency* have played a key role in information retrieval research. In this paper a theory of why inverse document frequency has been so effective is developed. Both the approach taken and the conclusions drawn differ from theories previously put forth. Employing techniques of *exploratory data analysis* EDA, the *weight of evidence* WOE in favor of relevance offered by query term occurrence is studied. The result is an explanatory theory of inverse document frequency, *idf*, derived from observed statistical regularities of extensive retrieval data.

The work reported here is the first phase of a larger research project whose goal is the development of a retrieval formula that: 1) is explanatory, in that each component of the formula has a direct interpretation in terms of measurable statistical characteristics of identifiable retrieval objects (query terms, documents, etc.); 2) is supported by the careful observation and study of empirical data; and 3) yields retrieval performance comparable, if not superior, to current state of the art retrieval systems.

The goal of this work is not primarily the production of an improved retrieval technique. The principal objective is rather a retrieval procedure derived from an explanatory theoretical model supported by hard empirical evidence. With this in mind, the first phase of the study has centered on the relation between the frequency with which a term occurs in a collection and its ability to discriminate between relevant and non-relevant documents. In this phase, document frequency is analyzed independent of other considerations, such as term frequency and document length, despite the fact that these are known to be significant factors in the construction of effective term weights.

We begin, in the following section, with a brief review of the concepts of *weight of evidence* and *exploratory data analysis* in the context of this article and how they pertain to the research reported here. We continue with a review of research specifically related to the utilization of document frequency for the weighting of query terms. Particular emphasis is placed on the *combination match* model, proposed initially by Croft and Harper, and work by Salton, *et al.* on *term precision*.

We go on to present an empirical study of retrieval data. Analysis of this data leads us to propose *mutual information* between term occurrence and relevance as a natural and useful measure of query term quality. We conclude that this measure is correlated with document frequency and use this to derive a theoretical explanation in support of *idf* weighting which is different from theories that have previously been proposed. The theory developed, in conjunction with the empirical evidence, predicts that a modification of the *idf* formula should produce improved performance. In Section 6, we present experiments that corroborate this prediction. In conclusion we return to the research summarized in Section 3, and compare and contrast it with the work presented here.

2 Weight of Evidence & EDA

Weight of Evidence I. J. Good formally defines the weight in favor of a hypothesis, h, provided by evidence, e, as [12, 11]:

$$woe(h:e) = \log \frac{O(h|e)}{O(h)} \tag{1}$$

which he thinks is a concept "almost as important as that of probability itself" [11, p. 249]. Good elucidates simple, natural desiderata for the formalization of the notion of weight of evidence, including an "additive property":

$$woe(h:e_1 \wedge e_2) = woe(h:e_1) + woe(h:e_2 \mid e_1) \tag{2}$$

This property states that the weight in favor of a hypothesis provided by two pieces of evidence is equal to the weight provided by the first piece of evidence, plus the weight provided by the second piece of evidence, conditioned on our having previously observed the first (i.e. $woe(h : e_2)$, calculated on the subspace corresponding to e_1). Starting from these desiderata, Good is able to show that, up to a constant factor, weight of evidence must take the form given in eq. 1. From (1), it follows directly that:

$$\log O(h|e) = \log O(h) + woe(h : e)$$

That is, if we are disposed to think on a log-odds scale, our final belief in a hypothesis (e.g. relevance of a document) is equal to our initial belief plus the weight of whatever evidence we are presented with. Log-odds is an attractive scale because weights accumulate additively; also because the entire range from $-\infty$ to $+\infty$ is used.

There is nothing new about using either log-odds or weight of evidence in information retrieval. The Robertson/Sparck Jones term weight discussed in the next section is motivated by the desire to determine the log-odds of relevance conditioned on the term occurrence pattern of a document. The weight, w_{rsj} of eq. 4, can be viewed as the difference between the weights of evidence in favor of relevance provided by the occurrence and non-occurrence of the term. Also, the focus of statistical inference based on logistic regression is the probability of the event of interest transformed by the logit function; that is, the log-odds.

Exploratory Data Analysis Hartwig and Dearing define exploratory data analysis as "a state of mind, a way of thinking about data analysis – and also a way of doing it" [16, p. 9]. They advance adherence to two principles. First, that one should be skeptical of data summaries which may disguise the most enlightening characteristics of the phenomenon being investigated. Second, that one must be open to unanticipated patterns in the data, because uncovering such patterns can be, and often is, the most eventful outcome of the analysis.

The emphasis in exploratory data analysis is on making the most of graphical displays of the data. The human mind is far better at uncovering patterns in visual input than in lists or tables of numbers. Depending solely on the reduction of large quantities of data to a few summary statistics erases most of the message the data have for us. EDA embodies a set of useful methods and strategies, fomented primarily by John W. Tukey [24]. For example, techniques for data smoothing and re-expression of variables have been used in the study presented in this article.

3 Related Work

In 1972, Sparck Jones, convincingly demonstrated that the weighting of query terms can significantly improve retrieval performance compared to unweighted *coordination match* ranking [23]. The weighting formula she proposed was an approximation of:

$$w_{sj} = \log \frac{N}{n} \tag{3}$$

where n is the *document frequency* of the term (the number of documents in which the term appears); and N is the number of documents in the entire collection.

3.1 Probabilistic Explanations

In a letter to the Journal of Documentation, Robertson pointed out that, viewed as a function of the probability of term occurrence, the sum of weights could be interpreted as the probability of mutual occurrence of multiple query terms [17]; thus providing theoretical arguments for the use of w_{sj}. Together, in 1976, Robertson and Sparck Jones presented the Binary Independence Model [18], in which terms are weighted by:

$$w_{rsj} = \log \frac{p(occ|rel) \cdot (1 - p(occ|\overline{rel}))}{(1 - p(occ|rel)) \cdot p(occ|\overline{rel})} \tag{4}$$

where $p(occ|rel)$ is the probability of the term occurring in relevant documents[1], and $p(occ|\overline{rel})$ is the corresponding probability for non-relevant documents. Use of the model depends on the availability of relevance feedback information, on which estimates of the two conditional probabilities can be based.

Applying the probabilistic approach of Robertson and Sparck Jones, Croft and Harper [5] work with an equivalent formulation of w_{rsj}:

$$w_{rsj} = \log \frac{p(occ|rel)}{1 - p(occ|rel)} - \log \frac{p(occ|\overline{rel})}{1 - p(occ|\overline{rel})} \tag{5}$$

Their goal is the development of a probabilistically justified weighting formula that can be used in a retrieval setting in the absence of, or prior to, relevance feedback. They make two assumptions: 1) there "is no information about the relevant documents and we could therefore assume that all the query terms had equal probabilities of occurring in the relevant documents" [5, p. 287]; and 2) the probability, $p(occ|\overline{rel})$, of a term occurring in a non-relevant document can be estimated by $\frac{n}{N}$, the proportion of documents that contain the term in the entire collection. With these two assumptions, the *combination match* formula:

$$w_{ch} = k + \log \frac{N - n}{n} \tag{6}$$

is derived. In this formula, k is an experimentally determined constant, corresponding to the log-odds of a term occurring in a relevant document. The second component is essentially equivalent to (3) for all but very high frequency terms.

Robertson and Walker [19] have recently looked anew at the combination match weight, w_{ch}. They point out two "anomalies" of the Croft/Harper weights. One is that the probability of a term occurring in a relevant document must go to zero as the probability of a term occurring in the collection as a whole goes to zero. More important, they state, is that the weight, w_{ch} of (6), will assume negative values for high frequency terms. These anomalies cause them to modify the assumption of equal probability of occurrence in relative documents, in favor of an assumption that this probability "increases from a non-zero starting point to reach unity" [19, p. 19] for a term that appears in all documents. In this paper, we shall examine in greater detail both of the assumptions leading to the combination match weighting formula, as well as their modification as proposed by Robertson and Walker.

[1] For the purposes of exposition, the original notation used by the authors discussed in this section has been replaced by the notation used in this paper. In this way, we hope to facilitate comparison among the different approaches, including the approach presented here.

3.2 Term Precision

In a series of papers in this same period, Salton and coworkers reported both theoretical and empirical work on a ranking formula based on what they called *term precision*. In earlier papers, term precision was defined as [20]:

$$w_{tp} = \frac{p(occ|rel)}{1 - p(occ|rel)} \; / \; \frac{p(occ|\overline{rel})}{1 - p(occ|\overline{rel})}$$

Later, term precision was defined as the log of this quantity [21, 27], yielding the same weight as given by Robertson and Sparck Jones (eq. 4). The form they adopt for what amounts to $p(occ|rel)$ differs from that of both Croft/Harper and Robertson/Walker. The term precision model assumes a two-piece linear function with: $p(occ|rel) = 0$ at $p(occ) = 0$; $p(occ|rel) = 1$ at $p(occ) = 1$; and a change in slope at $p(occ) = p(rel)$. This function is chosen based on the assumption that "the user will pick terms with properties somewhere between those obtaining for the random and perfect terms" [27, p. 159], sustained by solid theoretical arguments as to what the probability of occurrence conditioned on relevance must be for both perfect and random terms as a function of document frequency.

3.3 Regression

Regression strategies (explicitly or implicitly) assume a parameterized model and apply statistical techniques to fit the model to available data. In 1983, Fox used multiple regression analysis to derive an equation for predicting the probability that a document will be judged relevant to a query [3]. In [28], Yu and Mizuno use linear regression to determine parameter settings for both a binary and non-binary model. Fuhr and Buckley [9, 8] have used a least-square error criterion to determine coefficients for a polynomial weighting function of term-document pair descriptor variables. The group at Berkeley has conducted extensive research into the use of logistic regression [10, 4]. Logistic regression is generally considered a natural approach for estimating a probability. The $[0, 1]$ range that can be assumed by a probability does not correspond to other regression models, but is accounted for in logistic regression. Also, normality assumptions which are often behind the statistical inference techniques used in standard regression analysis are inappropriate for a dichotomous response variable (such as relevance). We shall return to discuss the benefits of applying exploratory data analysis in the selection of potential statistical models to which regression techniques can be applied.

3.4 Other Weighting Approaches

Not all work in term weighting has centered on the probability of relevance. Prior to the term precision model, Salton and others experimented with the *discrimination value* of a term [20]. This is a measure of how important a term is in distinguishing documents of the collection from each other. Information theoretic considerations have also been used. In early work, information theory was used to derive a weight based on signal-noise ratio [22]. In [26], Wong and Yao develop a term weighting theory based on the entropy of a term's distribution in the collection. They show that *idf* weighting is easily derived as a special case of their more general weighting scheme. In this paper, however, we restrict our attention to term weighting based on the probability of relevance.

4 TREC Data Analysis

In this section we present a retrospective analysis of information retrieval data. This analysis was undertaken for the specific purpose of gaining insight into the relationship between the document frequency of a query term and the expected value of the term as evidence in favor of relevance. The study involved data from queries 051-100 from the first Text REtrieval Conference (TREC) and the Associated Press (AP) documents from TREC volume 1 [13]. Each data point corresponds to one query term. The query terms were taken from the concepts field of the TREC 1 topics. For the purposes of uncovering underlying statistical regularities, we wanted a set of quality query terms which would keep to a minimum the "noise" in the data to be analyzed.

4.1 Binning the Data

Initially, we had planned for all query terms to be plotted. Two problems immediately presented themselves. First, rare terms are likely to have zero counts. For variables that are functions of log odds, a zero count translates to a (positive or negative) infinite value. One way around the problem is to add a small value to each of the counts of interest, as done for instance in [18], where for the purpose of estimating w_{rsj}, 0.5 is added to each count.

The choice of constant, however, is to a large degree arbitrary. Two slightly different choices for the constant value can give a very different overall picture of the data when they are plotted, particularly at the low frequency end. Since our objective is precisely to infer the "true shape" of the data, this approach is inadequate to our needs.

A second problem, is that the variance of the variables we are interested in is large, relative to the effects we hope to uncover. This can be seen clearly, for example, at the left of figure 4 where $p(occ|rel)$ is plotted against $\log O(occ)$ for all terms for which it has a finite value.

In order to confront both of these problems, data points were grouped together in bins. Each bin was then converted into a single *pseudo-term* by averaging all counts. Calculations of probabilities, weights, etc. were done on the pseudo-terms and these results were then plotted. A bin size of $k = 20$ was found to be best for our purposes. The plot of binned pseudo-terms corresponding to the left of figure 4 is shown at the right of the same figure. Although we will focus on the binned plots in this paper, each of these plots will be displayed alongside its unbinned version, in order that the reader may get a feel for the raw data. It should be kept in mind, however, that points with zero counts are not represented in the unbinned versions.

4.2 Plotting the Data

Taking a lead from the Croft and Harper formulation of eq. 5, our data analysis begins by focusing on the components, $p(occ|rel)$ and $p(occ|\overline{rel})$, and how these components correlate with document frequency. Because we hope to compare various document sets of differing sizes, we prefer not to plot our data in terms of absolute document frequencies. Instead, we plot against $p(occ) = \frac{df}{N}$,

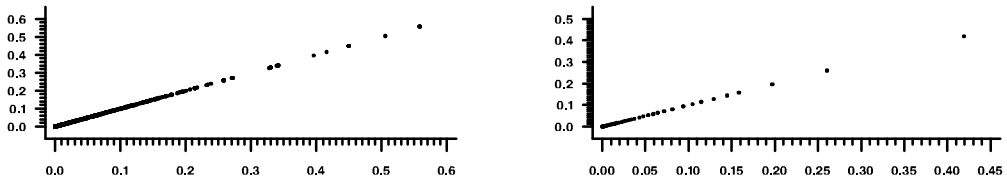

Figure 1: $p(occ|\overline{rel})$ as function of $p(occ)$

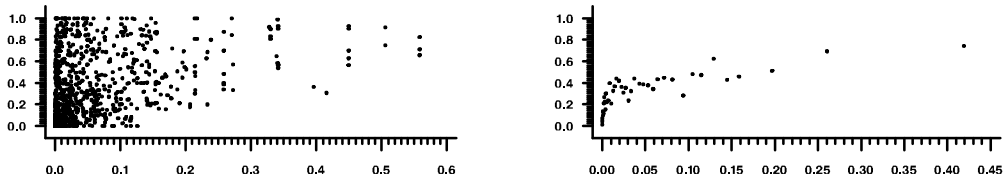

Figure 2: $p(occ|rel)$ as function of $p(occ)$

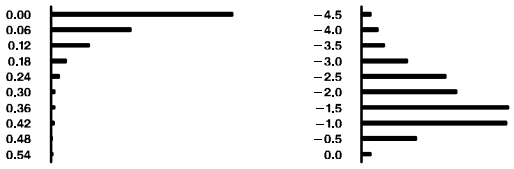

Figure 3: histograms for $p(occ)$ and $\log O(occ)$

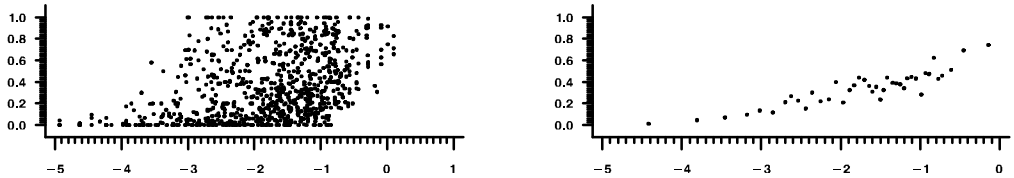

Figure 4: $p(occ|rel)$ as function of $\log O(occ)$

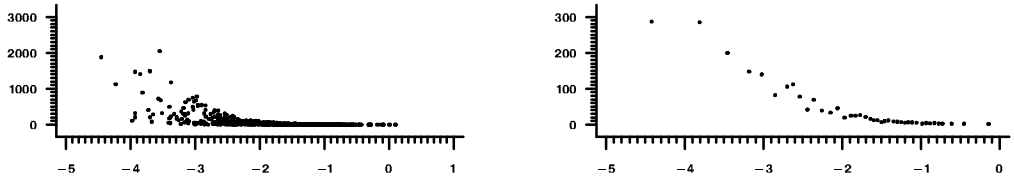

Figure 5: $\frac{p(occ|rel)}{p(occ)}$ as function of $\log O(occ)$

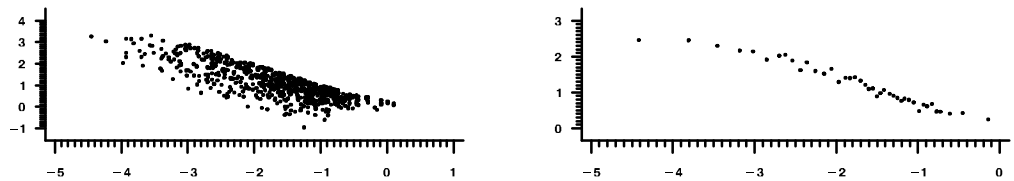

Figure 6: $\log \frac{p(occ|rel)}{p(occ)}$ as function of $\log O(occ)$

the probability that the term will occur in a document chosen randomly from the collection.

Occurrence in Non-relevant Documents With respect to $p(occ|\overline{rel})$, we see in figure 1 that it is well approximated by $p(occ)$. We will return to analyze the variable, $p(occ|\overline{rel})$, in more detail.

Occurrence in Relevant Documents More interesting is figure 2, which shows a plot of $p(occ|rel)$ as a function of $p(occ)$. We see from this scatter plot that as the document frequency (equivalently, probability of term occurrence) gets small, the probability of the term occurring in a relevant document gets small as well.

This plot of $p(occ|rel)$ vs. $p(occ)$ gives us reason to question the advisability of the assumption of equal probability of term occurrence in the relevant documents, used in [5] as the basis of eq. 6. This also puts into question the assumptions made in [19]. Both the assumption that $p(occ|rel)$ increases from a non-zero starting point and that the increase is linear with increasing $p(occ)$ contradict the evidence provided by figure 4. We return to discuss these points further in Section 7.

A glance at this graph suggests that a re-expression of variables may be indicated. The histogram shown at the left in figure 3 confirms that the distribution of document frequencies is highly skewed. With this type of skew, a logarithmic[2] transformation is often found to be beneficial [24]. In this paper, we go one step further and re-express the variable as $\log O(occ) = \log \frac{p(occ)}{1-p(occ)}$. For practical purposes, given typical document frequencies for query terms, the difference between $\log p(occ)$ and $\log O(occ)$ is negligible. For the development of a general theory, $\log O(occ)$ tends to be a preferable scale on which to work, due to the symmetric treatment it gives to probabilities below and above .5. The histogram at the right in figure 3 shows the distribution of the variable after it has been re-expressed as $\log O(occ)$. Of course, our interest in $\log p(occ)$ or $\log O(occ)$ is further motivated by the knowledge that this statistic is, in fact, known to be a useful indicator of term value.

The variable, $p(occ|rel)$, is re-plotted as a function of $\log O(occ)$ in figure 4. The plot against the log-odds shows that the decrease in $p(occ|rel)$ continues as document frequency get smaller and smaller, a fact that was obscured by the bunching together of points below $p(occ) \approx 0.01$ in the original plot (figure 2).

p(occ|rel) Relative to p(occ) Despite the transformation of the independent variable, looking at $p(occ|rel)$ directly makes it hard to appreciate the phenomenon of interest. The conditional probability of occurrence is higher for high frequency terms. But, high frequency terms are more likely to appear in documents, in general. It comes as no great surprise, then, that they are more likely to occur in relevant documents. This is particularly obvious for very high frequency terms as compared to very low frequency terms.

What may be of more interest to us, then, is how much more likely it is for a term to occur in the relevant documents compared to its being found in an arbitrary document of the collection as a whole. Figure 5 shows a plot of the ratio $\frac{p(occ|rel)}{p(occ)}$ as a function of $\log O(occ)$. We observe here a clear non-linear increase in this ratio as document frequency decreases. From this plot it is evident that, in general: 1) query terms are more likely to appear in relevant documents than in the collection as a whole, and 2) how much more likely their appearance in relevant documents is correlates inversely with document frequency. The apparent exponential nature of this correlation calls out for the logarithm of $\frac{p(occ|rel)}{p(occ)}$ to be investigated.

Log of the Ratio of p(occ|rel) to p(occ) In figure 6 the log of the ratio $\frac{p(occ|rel)}{p(occ)}$ is plotted against the logarithm of the odds of occurrence in the collection. In the plot, we observe:

- a roughly linear overall increase in $\log \frac{p(occ|rel)}{p(occ)}$ with decreasing $\log O(occ)$;
- a stronger linear relationship apparent in the midrange of document frequencies on the log scale;
- an apparent flattening of this growth at both high and low frequencies.

A number of comments are in order. First, a clear pattern has emerged that is difficult to attribute to chance. Furthermore, the "reality" of this regularity is corroborated by our inspection of data from other collections included in TREC volumes 1 and 2. To the author's knowledge, this relationship has not been previously reported in the information retrieval literature.

Second, the apparent flattening of the curve at the two extremes is supported by theoretical considerations. At the low-frequency end, we note that:

$$\frac{p(occ|rel)}{p(occ)} = \frac{p(occ \wedge rel)}{p(rel)p(occ)} = \frac{p(rel|occ)}{p(rel)} \leq \frac{1}{p(rel)} \qquad (7)$$

If we assume that, for a given query, the probability of relevance across the entire collection is approximately one in a thousand[3], then $\log \frac{p(occ|rel)}{p(occ)}$ must be below 3.0. We can conclude that, on average, the growth of the log ratio observed between $\log O(occ) = -1.0$ and $\log O(occ) = -3.0$ cannot be sustained for very small document frequencies. It is reasonable to assume that this growth should begin to taper off as $\log \frac{p(occ|rel)}{p(occ)}$ approaches $-\log p(rel)$.

The argument is similar at the high frequency end. We can safely assume that, on average, a query term, even a very high frequency query term, is more likely to appear in a relevant document than it is to appear in an arbitrary document of the collection. Hence, the ratio, $\frac{p(occ|rel)}{p(occ)}$, is greater than 1, and its logarithm greater than 0. Since we conclude that $\log \frac{p(occ|rel)}{p(occ)}$ can be expected to be positive at all document frequencies, its rate of descent must taper off at same point before reaching 0. Presumably it approaches 0 asymptotically as the log odds of occurrence goes to ∞ (i.e. term occurs in all documents). It is reasonable to entertain the hypothesis that this leveling off is what we are observing with the rightmost points in figure 6 (and have observed in plots for other collections as well). We must be cautious, however. The leveling off may, in truth, occur at higher frequencies; the flattening suggested by the few points in question attributable to chance happening.

[2] As an aid to intuitive comprehension, all logarithms in this paper are logarithms to the base 10.

[3] For the AP collection, the average probability of relevance over the 50 queries is .00085.

Finally, we note that the quantity:

$$\log \frac{p(occ|rel)}{p(occ)} = \log \frac{p(occ \wedge rel)}{p(rel) \cdot p(occ)} = \log \frac{p(rel|occ)}{p(rel)}$$

has connections to information theory. Often referred to as *mutual information*, it has been used as a measure of variable dependence in both information retrieval [25, 6] and computational linguistics [2]. In a very important sense, it can be taken as a measure of the information about one event provided by the occurrence of another [7]. In our context, it can be taken as a measure the information about relevance provided by the occurrence of a query term. In what follows, we shall adopt the notation,

$$\mathrm{MI}(occ, rel)$$

for this quantity, which we believe to be an object worthy of attention as a measure of term value in IR research. It will be the main focus in the analysis that follows.

5 Mutual Information and *idf*

In this section, we show how the general relationship observed between $\mathrm{MI}(occ, rel) = \log \frac{p(occ|rel)}{p(occ)}$ and *df* can be used to explain why inverse document frequency should be expected to produce good retrieval performance when used for term weighting.

5.1 Δwoe

Our interest in this paper is in modeling the weight of evidence in favor of relevance provided by the occurrence or non-occurrence of a query term. Presumably, the occurrence of a query term provides positive evidence and its absence is negative evidence. If we will assign a non-zero score only to those terms that appear in a document, this score should be, $woe(rel : occ) - woe(rel : \overline{occ})$. This quantity, which we shall denote by Δwoe, measures how much more evidence we have in favor of relevance when the term occurs in a document than we do when it is absent. Based on the formal definition of weight of evidence (1), together with that for mutual information, it is not difficult to show that Δwoe can be expressed as:

$$\begin{aligned}\Delta woe = \mathrm{MI}(occ, rel) + \log p(occ) - \log p(\overline{occ}|rel) \\ - \log p(occ|\overline{rel}) + \log p(\overline{occ}|\overline{rel}) \qquad (8)\end{aligned}$$

5.2 Δwoe ≈ MI(occ,rel)

We will now argue that:

1) $\log p(occ|\overline{rel}) \approx \log p(occ)$
2) $-\log p(\overline{occ}|rel)$ is not too big;
3) $\log p(\overline{occ}|\overline{rel}) \approx 0$

This done, we will be able to reduce eq. 8 to the following approximation for Δwoe:

$$\Delta woe \approx \mathrm{MI}(occ, rel) + \log p(occ) \qquad (9)$$

This approximation, together with assumptions concerning the form of $\mathrm{MI}(occ, rel)$ based on our data analysis, will lead us to an understanding of *idf* weighting.

Although we have made every effort to maintain an appropriate level of rigor in what follows, the arguments below do not attempt to be precise. We speak in terms of "not too large", "approximately the same", "not much greater than". The goal is to explain why, based on our analysis, *idf* in the form of $-\log O(occ)$ can, in general, be expected to perform well. We do not conclude that *idf* is optimal as a term weight; nor do we make any attempt at a precise estimate of how far from optimal it may be. Figure 6, and similar plots for other collections we have studied, lead us to expect values of $\mathrm{MI}(occ, rel)$ in the approximate range of 0.5 to 2.5 for the vast majority of query terms. This should be kept in mind. Quantities of the order of magnitude of 0.1 may then be considered negligible when the goal is to show that Δwoe is roughly approximated by $\mathrm{MI}(occ, rel)$.

1) $\log p(occ|\overline{rel}) \approx \log p(occ)$ We assume $p(rel)$ is small. If the probability of occurrence is fairly large relative to the probability of relevance, the order of magnitude of $p(occ|\overline{rel})$ will, of necessity, be the same as that of $p(occ)$, since it can be shown that:

$$\frac{p(occ) - p(rel)}{1 - p(rel)} \leq p(occ|\overline{rel}) \leq \frac{p(occ)}{1 - p(rel)} \qquad (10)$$

Therefore, $\log p(occ|\overline{rel})$ will be close to $\log p(occ)$.

When $p(occ)$ is not large relative to $p(rel)$, we can also conclude that $\log p(occ|\overline{rel}) \approx \log p(occ)$, but the reasoning requires knowledge of the relationship between $p(occ|rel)$ and $p(occ)$. For small $p(rel)$,

$$p(occ|\overline{rel}) \approx p(occ) \cdot \left(1 - \frac{p(occ|rel)}{p(occ)} \cdot p(rel)\right)$$

For a probability of occurrence greater than 1 in 10,000, the data indicate (figure 5) that $\frac{p(occ|rel)}{p(occ)}$ can be expected to be no more than 300. Given typical values for $p(rel)$, $(1 - \frac{p(occ|rel)}{p(occ)} \cdot p(rel))$ can still be expected to be less than one half of the value of $p(occ)$, and $\log_{10} p(occ|\overline{rel})$ can be expected to be not too different from $\log_{10} p(occ)$.

Concluding that $p(occ|\overline{rel}) \approx p(occ)$ may be problematic for query terms whose probability of occurrence is much less than .0001. On the other hand, query terms of this sort appear to be few and far between. Even when such a rare query term appears, its scarcity in the collection as a whole implies that its precise term weight is unlikely to have a major impact on overall retrieval performance.

The derivation of the combination match weighting formula (equation 6) in [5] also depends on $p(occ|\overline{rel})$ being well approximated by $p(occ)$. We emphasize, however, that for low frequency query terms, the argument given here depends heavily on the value of $\frac{p(occ|rel)}{p(occ)}$ relative to $p(rel)$. In theory, at least, $p(occ|\overline{rel})$ could be an arbitrarily small fraction of $p(occ)$. If this were the case, $\log O(occ|\overline{rel})$ would then be very different from $\log O(occ)$. Knowledge of the behavior of $\frac{p(occ|rel)}{p(occ)}$ supports a conclusion, $p(occ|\overline{rel}) \approx p(occ)$, which cannot be rigorously maintained in its absence.

2) $-\log p(\overline{occ}|rel)$ is not too big Though it may not be negligible, $-\log p(\overline{occ}|rel)$ cannot be too big. The data show that $p(\overline{occ}|rel) > .1$, even for the most frequent pseudoterm. Therefore, $0 < -\log p(\overline{occ}|rel) < 1$.

A component of the derived weight that approaches 1.0 is not insignificant. We believe that an *idf* formulation that takes this factor into consideration should perform better than one that does not. Nonetheless, a value

close to 1 for -log $p(\overline{occ}|rel)$ is achieved by only a small percentage of query terms – those which appear in more than 25% of all documents. Also, log $p(\overline{occ}|rel)$, falls off rapidly with decreasing document frequency. For the AP data, it is already less than 0.5 for the second bin of 20 data points. In and of itself, the effect of ignoring the contribution of log $p(\overline{occ}|rel)$ should not overwhelm the overall effect of the more important component, MI(occ, rel), of Δwoe given in eq. 8

3) log p($\overline{\mathbf{occ}}|\overline{\mathbf{rel}}$)≈0 Presumably, a query term is more likely to occur in the relevant documents than in the collection as a whole. Hence, it is more likely not to be present in the non-relevant documents than in a random document of the entire collection. That is, $p(\overline{occ}|\overline{rel}) > p(\overline{occ})$. In this study, $p(\overline{occ})$ is found to be greater than .7 for all pseudo-terms. Equivalently, $0 > \log p(\overline{occ}|\overline{rel}) > -0.15$. This component too, has a minimal effect on Δwoe.

5.3 *idf* approximates Δwoe

There is little question about our ability to infer from the available data that MI(occ, rel) increases with decreasing document frequency. To a first order approximation, we can say that this increase is roughly linear with respect to log $p(occ)$.

$$\Delta woe \approx \text{MI}(occ, rel) + \log O(occ) \qquad (11)$$

But, k_2 can be ignored. By casual inspection of figure 6, we see that any reasonable linear approximation of the plot of log $\frac{p(occ|rel)}{p(occ)}$ as a function of log $p(occ)$ will have an intercept value relatively close to 0. Once the constant k_2 has been eliminated, the remaining constant, k_1, becomes irrelevant for the purposes of ranking. And so we conclude that the *idf* formulation,

$$idf = -\log O(occ) = \log \frac{N - n_i}{n_i} \qquad (12)$$

should produce good retrieval performance.

6 Improving on IDF

We have shown that by accepting some, empirically motivated, assumptions concerning query terms the quantity Δwoe can be approximated by MI(occ, rel). By further assuming that MI(occ, rel) is roughly linear in log $O(occ)$, we showed that traditional *idf* formulations should perform well. We also argued in Section 4, however, that both theoretical and empirical considerations give reason to assume a flattening of MI(occ, rel) at both ends of the practical spectrum of document frequencies.

If we assume that the "true" form of the function that maps log $O(occ)$ to MI(occ, rel) involves flattening at the extremes, the map to Δwoe will exhibit similar shape. If we accept the hypothesis that the plot of figure 6 is representative of the general behavior of query terms for the types of queries and collections we study, we should expect improved retrieval performance from a term weighting formula that accounts for the observed flattening.

To test this prediction, we compared retrieval performance of two versions of the INQUERY IR system [1] on each of the ad-hoc tasks for TREC 1 through TREC 6 [14]. Queries were formed by taking all words from both the title and description. All stopwords were removed, as were all duplicates. The baseline system used pure *idf* term weighting with $idf = -\log O(occ)$[4]. The test system used a flattened version of *idf*. For this version, weights were kept at 0 for all values of -log $O(occ)$ below 1.0; increased at the same rate as -log $O(occ)$ from $-\log O(occ) = 1.0$ to $-\log O(occ) = 3.0$; and maintained at a constant value for all terms for which -log $O(occ)$ exceeded 3.0.

An alternative to this 3-piece piecewise-linear function, would have been to do either a (non-linear) regression to fit a curve to the pseudo-term data of figure 6; or a (non-linear) logistic regression to derive a function for log $O(occ|rel)$ vs. log $O(occ)$ using the 0/1 (rel/non-rel) values for the original, unbinned data. Although we plan to take just this approach eventually, we believe that it is premature at this stage of the research.

We do not feel that precise curve fitting is appropriate when important interactions are not yet being accounted for. In this case, the problem is dependencies that are known to exist among the query terms. If two query terms are not independent, the weight of evidence provided by the second term must be conditioned on whether the first term occurred or not. If the terms are correlated, $woe(rel : occ_2)$ will be greater than $woe(rel : occ_2 \mid occ_1)$.

It is generally accepted that interdependence of query terms has a noticeable impact on the effectiveness of term weighting [15, 25, 10]. Since, to date, we have made no attempt to model the influence of term dependence, determination of a precise function for estimation of $woe(rel : occ)$ is not indicated. What we look for, instead, is to test a general conclusion that the weights for terms at the low frequency extreme should be approximately equal, and the same for terms at the high frequency extreme.

The results of these tests are summarized in Table 1. The test version outperforms the baseline system in terms of average precision, on all six query sets, substantially on five of the six. The test system also outperforms the baseline system on a majority of queries on each of the six query sets. The "-/+" column gives the number of queries for which the test system performed below/above baseline. The column labeled "sign" gives the results of the sign test for each query set. Each value indicates the probability of the test version outperforming the baseline on as many of the queries as it did were each system equally likely to outperform the other. The column labeled "wilcoxon" gives the analogous probability according to the wilcoxon test, taking into account the size of the differences in average precision for each of the queries. Improvement was found at all (11) levels of recall on TREC's 2 through 5; all but the 50% recall level on TREC 1 and all but the 80% recall level on TREC 6.

7 Discussion

We have shown strong empirical support for concluding that MI(occ, rel) as a function of log $O(occ)$ is roughly linear, with a slope of the order of magnitude of $\frac{1}{2}$; and that this can be used to explain why inverse document frequency has been found to be so useful for term weighting. Previous probabilistic explanations have started from plausible a priori assumptions, in particular assumptions concerning the probability of a query term occurring in a relevant document. In this section, we

[4] Tests with $idf = -\log p(occ)$ were also run. For all test sets, performance differences were small, with -log $O(occ)$ outperforming -log $p(occ)$ on all 6 of the test sets.

	avg. prec.					
	baseline	test	% diff	- / +	sign	wilcoxon
TREC 1	0.1216	0.1312	7.88	18/32	0.0325	0.0201
TREC 2	0.0693	0.1021	47.36	10/40	0.0000	0.0000
TREC 3	0.0676	0.1257	86.03	4/46	0.0000	0.0000
TREC 4	0.0680	0.1002	47.42	15/34	0.0047	0.0006
TREC 5	0.0466	0.0688	47.63	17/32	0.0222	0.0006
TREC 6	0.1185	0.1422	20.01	12/37	0.0002	0.0000

Table 1: 3-piece piecewise-linear vs. linear versions of *idf*

review these earlier efforts in light of the results reported here.

Central to the combination match model of Croft and Harper is the assumption of constant $p(occ|rel)$ for all terms. In the absence of any pertinent prior knowledge concerning these terms, this is a quite reasonable assumption; essentially an application of the Laplacian "law of insufficient reason". However, with the availability of large numbers of conscientiously formulated queries, systematically judged against diverse, voluminous document collections, pertinent information becomes accessible. Inspection of this data supplies us with sufficient reason for assigning unequal probabilities for $p(occ|rel)$ based on a term's document frequency.

The probabilities suggested by the data vary over a wide range. The value of the first term, $\log O(occ|rel)$, in eq. 5, ranges from approximately 0.0 to 2.0. This value, which is treated as constant in the model, varies over almost half the range of the second term, $\log O(occ|\overline{rel})$, which stays between 0.0 and 4.0 for virtually all of the terms of our study. Also, the second term, $\log O(occ|\overline{rel})$, cannot be presumed to be approximated by $\log O(occ)$, *a priori*. This puts the theoretical foundation of the combination match model in question.

The Robertson/Walker adjustment of the combination match formula allows for an increase in $p(occ|rel)$ that grows linearly with $p(occ)$. This is intuitively appealing at the same time that it resolves an anomaly in the combination match model. What's more, the data confirm that $p(occ|rel)$ does rise monotonically with $p(occ)$. However, the increase is not at all linear, at least not for the greater portion of the document frequency range. Also, the data indicate that the positive value that should be assumed for $p(occ|rel)$ for the lowest frequency terms must be very very small. Unfortunately they find themselves restricted to values above 0.5. Figure 2 shows that only fairly high frequency terms can be expected to appear in as many as half of the relevant documents.

The term precision model comes closest to being validated by the empirical data. The overall shape of the curve for $p(occ|rel)$ predicted by the model comes closest to approximating the plot shown in figure 2. But, the 2-piece piecewise-linear function of the term precision model derives from the assumption that the query term of a given document frequency will have a probability, $p(occ|rel)$, that is a linear combination of the best possible query term and a randomly chosen query term at that document frequency. Again, a quite reasonable assumption in the absence of any pertinent knowledge, appears to be contradicted by the data.

All of these models have resulted from what can be considered *a priori* reasoning. While the conceptualization involved is insightful and to a large degree forced on earlier researchers due to the paucity of hard data, the availability of extensive retrieval data is, we believe, an invaluable asset which should not be ignored. This extends as well to research that seeks to apply statistical techniques such as regression analysis to the IR task. This research does have a more empirical flavor. Data is used so that parameters can be estimated. That is to say, so that a member of a family of functions can be chosen. The *a priori* aspect, though more subtle, is still present, however. From which family of functions, is the "best" member to be selected from? There is typically little reason, *a priori*, to believe that the relationship of interest is well modeled by, for example, a polynomial function; or that the log-odds of some event is linear as a function of the proposed "explanatory" variables. Exploratory analysis could be part of an initial phase, during which the researcher becomes acquainted with data in order to determine what would be a reasonable family of functions on which to base regression techniques.

As mentioned in the introduction, the work reported here represents step 1 of a more extensive research agenda. Major objectives that lie ahead include the analysis and modeling of:

- the overall effect of query term dependencies on the total weight of evidence. As explained in Section 6 we have yet to account for what we know to be a systematic overestimate of Δwoe.
- within-document term frequency as a source of evidence;
- document length as a source of evidence;
- distinctions as sources of evidence between different classes of terms, such as: phrases vs. simple words, capitalized words vs. lower case words; query expansion terms vs. terms of original user query.

We are optimistic that these investigations will prove fruitful.

8 Acknowledgments

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, and also supported in part by United States Patent and Trademark Office and Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract number F19628-95-C-0235.

Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsors.

References

[1] J. P. Callan, W. B. Croft, and S. M. Harding. The inquery retrieval system. In *Proceedings of the 3rd In-*

ternational Conference on Database and Expert Systems Applications, pages 78–83, 1992.

[2] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164, Hillsdale, NJ, 1991. Lawrence Erlbaum Associates.

[3] W. S. Cooper, D. Dabney, and F. Gey. Probabilistic retrieval based on staged logistic regression. In Nicholas Belkin, Peter Ingwersen, and Annelise Mark Mejtersen, editors, *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 198–210, Copenhagen, Denmark, June 1992.

[4] Wm. S. Cooper, Aitao Chen, and Fredric C. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtreival Conference (TREC-2)*, pages 57–66, Gaithersburg, Md., March 1994. NIST Special Publication 500-215.

[5] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, December 1979.

[6] W. B. Croft and Jinxi Xu. Corpus-specific stemming using word form co-occurence. In *Proceedings for the Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 147–159, Las Vegas, Nevada, April 1995.

[7] Robert M. Fano. *Transmission of Information; a Statistical Theory of Communications*. MIT Press, Cambridge, MA, 1961.

[8] N. Fuhr. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems*, 7(3):183–204, 1989.

[9] Norbert Fuhr and Chris Buckley. Probabilistic document indexing from relevance feedback data. *ACM Transactions on Information Systems*, 9(2):45–61, 1991.

[10] Fredric C. Gey. Inferring probability of relevance using the method of logistic regression. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 222–231, Dublin, Ireland, July 1994.

[11] I. J. Good. *Probability and the Weighing of Evidence*. Charles Griffin, London, 1950.

[12] I. J. Good. Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*, pages 249–269. North-Holland, Amsterdam, 1983.

[13] Donna Harman. Overview of the first Text REtrieval Conference (TREC-1). In D. K. Harman, editor, *The First Text REtrieval Conference (TREC1)*, pages 1–20, Gaithersburg, Md., February 1993. NIST Special Publication 500-207.

[14] Donna Harman. Overview of the fifth Text REtrieval Conference (TREC-5). In E. M. Voorhees and D. K. Harman, editors, *The Fifth Text REtreival Conference (TREC-5)*, pages 1–28, Gaithersburg, Md. 500-238, November 1997. NIST Special Publication 500-238.

[15] D. J. Harper and C. J. van Rijsbergen. An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3):189–216, September 1978.

[16] Frederick Hartwig and Brian E. Dearing. *Exploratory Data Analysis*. Sage Publications, 1979.

[17] S. E. Robertson. Term specificity. *Journal of Documentation*, 28(2):164–165, 1972. Letter to the editor, with response by K. Sparck Jones.

[18] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1977.

[19] S. E. Robertson and S. Walker. On relevance weights with little relevance information. In Nicholas J. Belkin, A. Desai Narasimhalu, and Peter Willett, editors, *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 16–24, Philadelphia, Pennsylvania, July 1997.

[20] G. Salton, A. Wong, and C. T. Yu. Automatic indexing using term discrimination and term precision measurements. *Information Processing & Management*, 12:43–51, 1976.

[21] G. Salton, H. Wu, and C. Y. Yu. The measurement of term importance in automatic indexing. *Journal of the American Society for Information Science*, 32:175–186, 1981.

[22] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[23] K. Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[24] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, MA, 1977.

[25] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2 edition, 1979.

[26] S. K. M. Wong and Y. Y. Yao. An information-theoretic measure of term specificity. *Journal of the American Society for Information Science*, 43(1):54–61, 1992.

[27] C. T. Yu, K. Lam, and G. Salton. Term weighting in information retrieval using the term precision model. *Journal of the ACM*, 29(1):152–170, January 1982.

[28] Clement T. Yu and Ilirotaka Mizuno. Two learning schemes in information retrieval. In Yves Chiaramella, editor, *Proceedings of the 11th International Conference on Research and Development in Information Retrieval*, pages 201–215, Grenoble, France, June 1988.