Empirical Studies of Query/Document Characteristics as Evidence in Favor of Relevance

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

Abstract Query/document characteristics known to be useful for information retrieval are analyzed for a specific collection/query-set pair. These features are analyzed in terms of the *weight of evidence* in favor of relevance provided by values assumed by the feature variables. Weight of evidence, a measure of how much more likely a hypothesis is believed to hold after evidence is considered than before it is observed, is formally defined; and a technique for the analysis of weight of evidence as a function of features of interest is presented. The method is exemplified by showing how it has been applied to analyze evidence in the form of: the *coordination level*, and the *inverse document frequencies* and *term frequencies* for all of the query terms. The result of data analysis is a model of weight of evidence that can be used as the foundation of a retrieval ranking formula. Results of preliminary evaluation of the derived formula are presented and discussed.

1 Introduction

This paper presents an analysis of the weight of evidence in favor of relevance offered by query/document features traditionally used for ranking in information retrieval. The predominate objective is to obtain a more precise and rigorous understanding of the relationship these retrieval characteristics have to the probability that a document will be judged relevant. The long-term goal of this analysis is the development of a retrieval formula, the components of which can be understood in terms of statistical regularities observed in the class of retrieval situations of interest.

In earlier work [13], inverse document frequency (*idf*) was studied as a source of information concerning the weight of evidence in favor of relevance indicated by the occurrence of a query term in an arbitrary document. In this paper, a more general formal modeling framework is developed. Based on it, a methodology is presented for the analysis of the relationship between query/document characteristics and the probability that a document will be judged relevant to the query. Application of the proposed methodology to a homogeneous collection of documents – 1988 news articles from the associated press (AP88), taken from volume 2 of the TREC data, evaluated for queries 151-200 from TREC 3 [14] – will serve as the vehicle for exposition of the principle techniques involved. The characteristics that have been studied, and will be discussed here are:

coordination level: the number of query terms that occur (one time or more) in the document;

inverse document frequency: for each of the query terms, $-\log(df/N)$, where df is the number of documents containing the term and N is the size of the collection;

term frequency: for each of the query terms, the number of the times the term occurs in the document.

Although we focus on these particular features, the approach is general, and can in principle be applied to any feature set deemed to be of interest to the researcher or system designer.

The development of the model is in terms of the conditional *weight of evidence* in favor of relevance provided by the values of query/document features. I.J. Good [11, 12] has shown that from 3 intuitively plausible desiderata for a concept of weight of evidence, it follow that:

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

This is interpreted as the weight of evidence in favor of the hypothesis, $H = h$, provided by the evidence associated with observing that the event, $E = e$, has occurred. Good has shown that, to within a constant factor, this must be equal to the *odds* that the hypothesis holds after the evidence is considered, relative to the odds of it being true prior to the evidence being observed, where the odds of the hypothesis $O(H = h)$, is equal to how much more likely it is to be true than not true, $p(H = h)/p(H \neq h)$. The ratio is measured on a log scale. The constant factor can be absorbed in the base of the logarithm, which for this purposes of this paper, will be taken as 10.

More generally, the weight of evidence of one piece of evidence, $E_2 = e_2$, can be conditioned on having previously observed other evidence, $E_1 = e_1$. The more generally applicable form of eq. 1 is then:

$$\begin{aligned} &woe(H = h : E_2 = e_2 \mid E_1 = e_1) \\ &= \quad \log \frac{O(H=h|E_2=e_2,E_1=e_1)}{O(H=h|E_1=e_1)} \end{aligned} \tag{2}$$

It is not hard to show that from eq. 1 and eq. 2, it follows that[1]:

$$\begin{aligned} &woe(H = h : E_1 = e_1, E_2 = e_2) = \\ &\quad woe(H = h : E_1 = e_1) \\ &+ \quad woe(H = h : E_2 = e_2 \mid E_1 = e_1) \end{aligned} \tag{3}$$

[1] Actually, eq. 3 is one of Good's desiderata, making it more accurate to say that eq. 1 and eq. 2 follow from it.

which will be important for the development of our model.

This article shows how query/document features can be studied, how a model in terms of this evidence can be formulated, and parameters for it can be determined. The resulting model can be used directly as a scoring mechanism for which the ranking status values (RSVs) that are produced have a precise probabilistic interpretation.

We will also discuss preliminary results suggesting that the modeling framework, and more important the general approach to the analysis of evidence, proposed here may lead to a ranking formula that performs as well as state-of-the-art retrieval formulas that have evolved over the years.

The following section will discuss related work, placing this paper in the contest of historical trends in IR research. In section 3, the analytic approach is explained by following the course of analysis for the AP88 data. Where section 3 focuses on exposition of the methodology, section 4 goes on to discuss in greater detail some of the issues involve, decisions taken, and alternatives considered

This article reports progress on a long-term research agenda: the development of a methodology for the application of empirical methods and probability theory to IR research and system design. Much work remains to be done. Section 5 discusses the next steps that will be taken in the pursuit of these goals.

2 Related Work

The study presented in this paper is founded on two key points: 1) probabilistic modeling in terms of weight of evidence, and 2) the use of exploratory data analysis [23, 15] to uncover and analyze statistical regularities that may be exploited for the purposes of ranked retrieval. Both of these aspects continue existing lines of information retrieval research.

In the early years of information retrieval, much research focused on automating the indexing process. Statistical approaches were taken in much of this work. This included the application of statistical techniques such as factor analysis, discriminant analysis, and latent class analysis [1]. In the mid 1970s, work by Salton and colleagues studied use of the *term discrimination model* to establish the value of an index term [19, 18]. Empirical methods played a central role in this research. In the same period, Bookstein and Swanson conjectured that the distribution of words occurring in a collection might be well modeled as a mixture of Poisson distributions [2]. Harter studied a collection of Freud's works to investigate this conjecture and used the method of moments to estimate parameters for a 2-Poisson model. This line of investigation was continued by Srinivasan, who considered both 2-Poisson and 3-Poisson models [22] , and later by Margulis, who permitted the number of distributions in the mixture to vary between 2 and 8 [16]. Extensive analysis of term distributions was a key component of this research.

Research with a very similar flavor to the work reported here was presented by Singhal *et al.* [21, 20]. They used empirical methods to study the relationship between document length and probability of relevance, and their use of binning predates the use of binning in our work.

For many years, weight of evidence has been an object of study in information retrieval. The Binary Independence Model is a formalization based on the log-odds of relevance [17]. The ranking formula derived from it is the weight of evidence in favor of relevance provided by the occurrence pattern of query terms. Cooper's observation that the Binary Independence Model depends on what he calls the linked dependence assumption [5], can be rephrased neatly in terms of independence of weights of evidence. Where the Binary Independence Model depends on relevance feedback, the Combination Match Model [7] extended application of the theory to retrieval in the absence of relevance judgments, and in this sense is more closely related to the research reported here. Salton, *et al.* also investigated weight of evidence of term occurrence, which they called *term precision*, and derived a theory of its relation to inverse document frequency.

The analyses to be discussed here rely heavily on the use of regression techniques. In 1983, Fox used multiple regression analysis [8]. Yu and Mizuno have used linear regression to set parameters for their models [24], and Fuhr and Buckley have used regression to fit polynomial curves [9]. Most closely related to the work presented here is research into the use of logistic regression at the University of California, Berkeley [4, 6, 10]. Both approaches make extensive use of regression techniques. Both approaches attempt to model the log-odds of relevance conditioned on a given set of predictor variables.

3 Modeling of Relevance

In this section, the use of exploratory data analysis to study evidence in favor of relevance is explained. Analysis of the relevance judgments for TREC queries 151-200 against the AP88 collection is used to exemplify the process. Queries were taken from the titles of the 50 TREC topics with stopwords removed and duplicate terms eliminated. The analysis centers on the modeling of log-odds of relevance conditioned on the evidence:

$$\log O(rel|e_1,\ldots,e_n) \quad = \quad \log \frac{p(rel \mid e_1,\ldots,e_n)}{p(\overline{rel} \mid e_1,\ldots,e_n)}$$

where $p(rel \mid e_1,\ldots,e_n)$ is the probability of relevance given all the evidence and $p(\overline{rel} \mid e_1,\ldots,e_n)$ is the conditional probability of non-relevance.

First we analyze evidence corresponding to coordination level. This results in the M_{CQ} model of relevance conditioned on the query being evaluated and the number of query terms occurring in the document. In section 3.2, we see that inverse document frequency is correlated with residual log-odds of relevance, relative to the M_{CQ} model. Extension of the model to include idf_i for each of the query terms, $i = 1, 2, \ldots$, produces the M_{ICQ} model. Finally, analysis of the role of term frequency results in the M_{TICQ} model. It is this model on which a ranking formula will be based.

3.1 Modeling of $woe(rel : co \mid q)$

In order to model the weight of evidence offered by coordination level, the data were first grouped by the value of the *Coord* variable into subsets, $C_1, C_2, \ldots$

$$C_i = \{(q,d) \in \mathcal{Q} \times \mathcal{D} \mid Coord(q,d) = i\}$$

where $\mathcal{Q}$ is the set of queries and $\mathcal{D}$ is the set of documents. For each of these subsets of query/document

pairs, an *expected number of relevant documents* was computed. The calculation is based on the estimated probabilities of relevance, $\hat{p}(rel|Qry = q)$, for each query.

$$\hat{r}_i = \sum_{Qry=q} n_{i,q} \cdot \hat{p}(rel|Qry = q) \quad (4)$$

where $n_{i,q}$ is the number of documents that contain i terms from query q. The probability, $\hat{p}(rel|Qry = q)$, is estimated by counting the fraction of documents relevant to the query. The product, $n_{i,q} \cdot \hat{p}(rel|Qry = q)$, is an estimate of the number of these $n_{i,q}$ documents that can be expected to be relevant. The sum of these over all queries is then an estimate of the number of relevant documents in the set C_i. Although somewhat less intuitive, the summation given in eq. 4 can also be expressed as:

$$\hat{r}_i = \sum_{(q,d):Coord(q,d)=i} \hat{p}(rel|Qry = q) \quad (5)$$

This formulation will be seen to be more useful as this technique is extended to the analysis of *idf* and *tf* as sources of evidence.

Accompanying the calculation of $\hat{r}_i$, the *actual* number of documents in C_i that are relevant, r_i, can be counted. Both of these can be transformed into log-odds: $\log \frac{\hat{r}_i}{n_i - \hat{r}_i}$ and $\log \frac{r_i}{n_i - r_i}$, where $n_i = |C_i|$ is the number of query/document pairs for which the coordination level is i. The difference between the two:

$$res_i = \log \frac{\hat{r}_i}{n_i - \hat{r}_i} - \log \frac{r_i}{n_i - r_i}$$

can be viewed as the *residual* log-odds of relevance; the difference between the *observed* log-odds of relevance and the log-odds that would be *predicted* by a model that only uses information about which query is being evaluated. After the residuals were calculated for each subset of query/document pairs, C_i, a residual plot was produced. Figure 1(a) shows the scatter plot of residuals against coordination level. The lightly shaded bars in the background give a cumulative histogram. The height of a bar at $Coord = i$ indicates the fraction of query/document pairs under consideration for which $Coord(q, d) \leq i$. The small circle along the bottom of the graph at $Coord = 7$ indicates that there were 0 relative documents in subset C_7. Since $\hat{r}_i$ is undefined for 0 relevant documents (i.e. $\hat{r}_i = -\infty$), and hence res_i is undefined, the circle serves to remind us that a point is missing from the plot at this value for the predictor variable.

Figure 1(b) gives a *smoothed* version of the same residual plot. The smoothing mechanism that was used will be explained with the discussion of the *idf* variable, where the role of smoothing is more important and the effect, being more dramatic, can be more easily appreciated.

Overlaid on figure 1(b) is a regression line:

$$res = \beta_0^{\mathrm{CO}} + \beta_1^{\mathrm{CO}} \cdot co$$

After the smoothed version of the residual points was produced, a linear regression was performed, motivated by the linearity suggested by figure 1(a). The line produced is the line that best fits the data in that it minimizes the mean square error. For this purpose each point in figure 1(b) is considered as n_i points. This regression line can be used to form a model, M_{CQ}, that takes into account both the query being evaluated and the coordination level. The points marked by small x's at each value, $co = 1, 2, \ldots$, that lie about the line, $res = 0$, show the difference between the log-odds predicted by the model, M_{CQ}, and the log-odds actually observed – their proximity to the line, $res = 0$, demonstrating that the model provides a close fit to the data.

The log-odds difference given by the model is equivalent to the weight of evidence in favor of relevance provided by the coordination level, conditioned on the query.

$$\begin{aligned} & woe(rel : co \mid q) \\ &= \log \tfrac{O(rel|co,q)}{O(rel|q)} \\ &= \log O(rel|co, q) - \log O(rel|q) \end{aligned}$$

This M_{CQ} model will be the central component of the complete model. The next step will be to use it as a basis for the analysis of the evidence provided by the rarity of a term.

3.2 Modeling of $woe(rel : idf \mid co, q)$

The analyses for both *idf* and *tf* as sources of evidence also depend on the study of residual log-odds of relevance. Whereas *Coord* is a feature of query/document pairs, for both *idf* and *tf*, data points involve individual query terms as well. For the analysis of coordination level, each query/document pair corresponded to only one data point. For the analysis of *idf* and *tf*, each query/document pair corresponds to multiple data points – one for each term appearing in the document. Since the same relevance judgment applies to each of these points, each point will be considered *weighted* by $w(q, d, t) = 1/Coord(q, d)$.

In this way, the 5 points corresponding to a relevant query/document pair with a coordination level of 5 will each receive a weight of 1/5 – i.e. each will be considered as 1/5 of a relevant document; 2 points corresponding to a non-relevant document with a coordination level of 2 will be considered as 1/2 of a non-relevant document; in total, 1 relevant and 1 non-relevant document.

With this in mind, the method of analysis for *idf* is a straightforward extension of that for coordination level. For each query term, t, an expected number of relevant documents is computed.

First, the data are grouped into subsets,

$$I_i = \{(q, d, t) \in \mathcal{Q} \times \mathcal{D} \times \mathcal{T} \mid Qry(t) = q, t = i\}$$

($\mathcal{T}$, the set of query terms), with one subset for each query term. (Occurrences of the same word used in 2 or more different queries are considered different terms.) Here again, the actual number of relevant documents can be counted and the fraction of documents that are relevant can be compared against the expected fraction for each subset. It must be kept in mind that both the observed and expected values are based on counts of entries weighted by the inverse of the coordination level. More precisely,

$$r_i = \sum_{(q,d,t):t=i, Rel(q,d)=1} w(q, d, t)$$

The calculation of the expected number of relevant documents for each subset is analogous to that given in eq. 5:

$$\hat{r}_i = \sum_{(q,d,t):t=i} \hat{p}(rel|Coord = co, Qry = q) \cdot w(q, d, t)$$

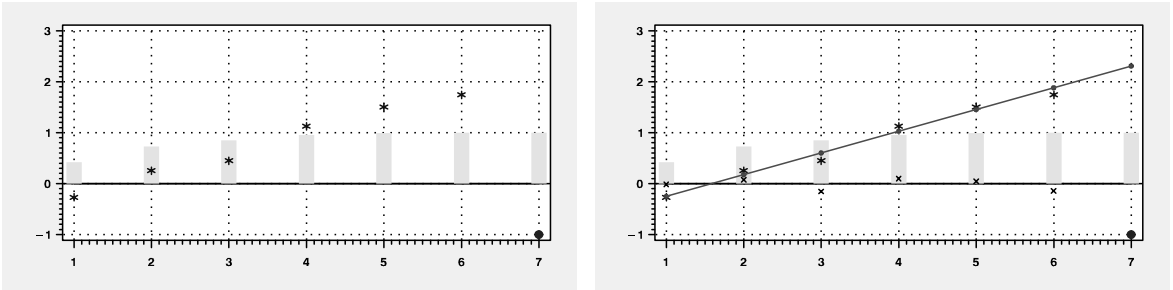

Figure 1: Residual log-odds as function of coordination level: a) unsmoothed b) smoothed with regression

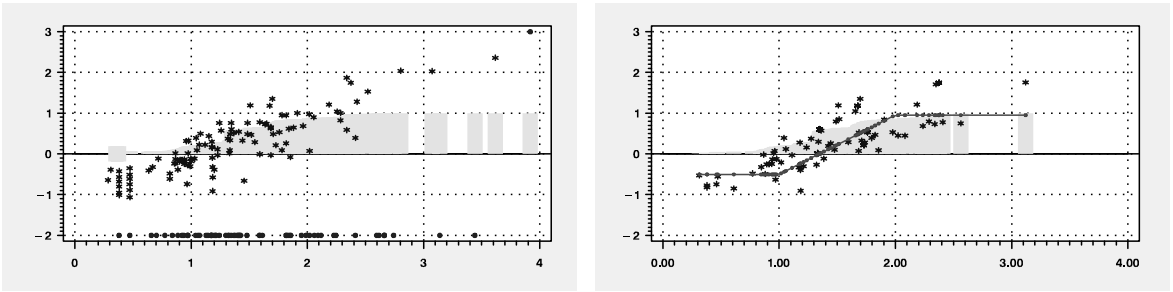

Figure 2: Residual log-odds as function of *idf*: a) unsmoothed b) smoothed with regression

where the estimated probability is calculated from the estimated log-odds of relevance:

$$\log O(rel|co) \quad = \quad \log O(rel|q) + woe(rel : co \mid q)$$

with the second term being the weight of evidence predicted by the M_{CQ} model; i.e. given by the regression line shown in figure 1(b).

Figure 2(a) shows a scatterplot of these residuals against *idf* value. Again, small circles are shown at the bottom of the graph for each residual that is undefined because the corresponding term did not appear in any relevant documents. The vertical bars in the background give a cumulative histogram for the (weighted) data points.

Figure 2(b) shows a smoothed version of these same data. To produce the smoothed plot, a number of *bins*, b, is fixed – with $b = 100$ for this graph. The points of figure 2(a) are sorted by *idf* and assigned to bins as follows. Starting with low-*idf* terms, data points are assigned one-by-one to the first bin, until 2% (= 1/50) of the relevant documents have been accumulated. When a term is reached that will not fit in the first bin, it is partitioned in two parts. The division is such that the first part can be assigned to the first bin, completing the alloted 2% of relevant documents. The second part is then distributed to the second bin, and the process continues. Three counts – observed relevant documents, expected number of relevant documents, and total documents – are distributed proportionally when data for a term must be partitioned across two bins. For each bin, the three counts are summed over all terms assigned to the bin. At the same time, an *idf* value is assigned to the bin by taking a weighted average of the idf values for the terms of the bin. The weighting is based on the number of documents associated with each term (keeping in mind, again, that a term occurrence is only counted as $1/Coord(q,d)$ data points).

In this way 50 (idf_i, res_i) pairs are generated. None of these points will correspond to 0 relevant documents. Also, by choosing a reasonable bin size, the possibility of the fraction of relevant documents reaching 100% (which would also yield an undefined $\hat{r}_i$, and hence, res_i) can be effectively eliminated.

Previous investigation of the weight of evidence provided by term *idf* values, as reported in [13], suggested that the weight of evidence provided by *idf* is well-modeled by a 3-piece linear function. Review of the general form of residual plots, such as that shown in figure 2(b), generated at various level of smoothing, tended to corroborate these earlier findings. Together, these two factors motivated the attempt to model $woe(rel : idf \mid co, q)$ as a 3-piece linear function.

In order to realize this, a linear regression was performed to determine parameters for the following linear model:

$$res = \beta_0^{\text{IDF}} + \beta_1^{\text{IDF}} x_1$$

$$\text{where } x_1 = \begin{cases} 0 & \text{if } idf < 1 \\ idf - 1 & \text{if } 1 \leq idf \leq 2 \\ 1 & \text{if } idf > 2 \end{cases}$$

The resulting estimates for the parameters, β_0^{IDF} and β_1^{IDF}, yield the model, M_{ICQ}, that minimizes the mean square error of all those models for which the expected value, $E[r_i]$, of the residual is a 3-piece linear function of *idf* with flat segments at the two extremes, and *elbows* at $idf = 1.0$ and $idf = 2.0$. Regressions were also run with a 4-parameter function, allowing for a general 3-piece linear model (one without the flat-segments restriction). These regressions showed no statistical evidence of non-zero slope in either of the tails, an indication that might have justified consideration of a more general model. Regressions were also run for other settings for the elbows; with values close to 1.0 and 2.0 resulting in the best fit. The 3-piece linear curve shown in figure 2(b) shows the resulting model imposed on the scatterplot of smoothed residual values.

3.3 Modeling of $woe(rel : tf \mid idf, co, q)$

Analysis of the evidence provided by term frequency proceeds along the same lines as that of inverse document frequency. Data points are grouped into subsets

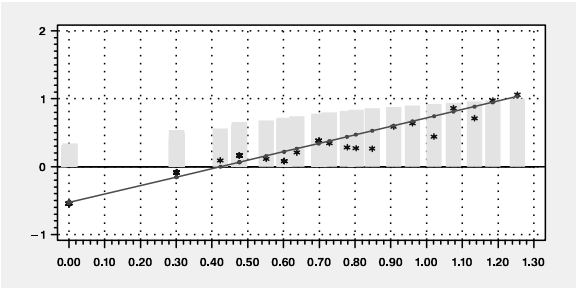

Figure 4: Residual log-odds as function of log(tf): smoothed with regression

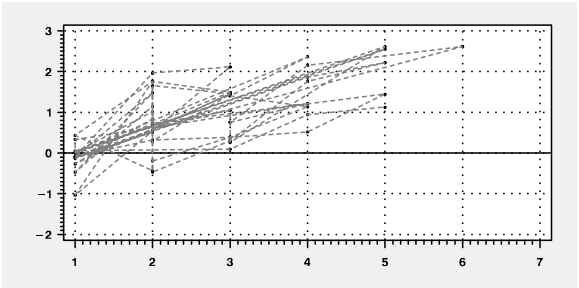

Figure 5: Residual log-odds as function of coordination level for individual queries

$TF_1, TF_2, \ldots$, according to the number of occurrences of the query term in the document, $Tf\ (q,d,t)$. For each subset, TF_i, the observed number of relevant documents is determined, and the expected number of relevant documents is calculated as:

$$\hat{r}_i \quad = \sum_{(q,d,t):Tf(q,d,t)=i} \hat{p}(rel|idf, co, q)$$

where $\hat{p}(rel|idf, co, q)$ is calculated from the log-odds of relevance according to model, M_{ICQ}:

$$\begin{aligned} \log \hat{O}(rel|idf, co, q) \quad &= \quad \log \hat{O}(rel|q) \\ &+ \quad w\hat{o}e(rel : co \mid q) \\ &+ \quad w\hat{o}e(rel : idf \mid co, q) \end{aligned}$$

A scatterplot of the resulting residuals is shown in figure 3(a). A number of transformations of the variables involved were tried. Figure 4 shows a plot of the residuals with $\log(tf)$ as predictor variable, smoothed to 50 bins. (Some of the original points, in particular the point for $tf = 1$, i.e. $\log(tf) = 0$, are spread over a number of bins, accounting for several points with the same coordinates, overlapping one another, on the smoothed version of the graph.) The apparent linearity motivated the application of a simple linear regression, the result of which is overlaid on the smoothed scatterplot of figure 4. The fit of the curve to the smoothed data on the more natural, unlogged tf scale can be seen in figure 3(b). For the resulting model, M_{TICQ}, weight of evidence in favor of relevance provided by the coordination level and the idf and tf values of each of the query terms is given by:

$$\begin{aligned} &\log \hat{O}(rel|q,d) = \log \hat{O}(rel|q) \\ &\quad + \quad w\hat{o}e(rel : tf_1, \ldots, tf_n, idf_1, \ldots, idf_n, co \mid q) \end{aligned}$$

with

$$\begin{aligned} &w\hat{o}e(rel : tf_1, \ldots, tf_n, idf_1, \ldots, idf_n, co \mid q) = \\ &\qquad w\hat{o}e(rel : co \mid q) \\ &\quad + \sum_{i=1}^{n} \quad w\hat{o}e(rel : idf_i \mid co, q) \\ &\qquad\qquad + \quad w\hat{o}e(rel : tf_i \mid idf_i, co, q) \end{aligned} \tag{6}$$

where

$$\begin{aligned} w\hat{o}e(rel : co \mid q) \quad &= \quad \beta_0^{\mathrm{CO}} + \beta_1^{\mathrm{CO}} co \\ w\hat{o}e(rel : idf_i \mid co, q) \quad &= \\ \beta_0^{\mathrm{IDF}} + &\begin{cases} 0 & \text{if } idf_i < 1 \\ \beta_1^{\mathrm{IDF}}(idf_i - 1) & \text{if } 1 \le idf_i \le 2 \\ \beta_1^{\mathrm{IDF}} & \text{if } idf_i > 2 \end{cases} \\ w\hat{o}e(rel : tf_i \mid idf_i, co, q) \quad &= \quad \beta_0^{\mathrm{TF}} + \beta_1^{\mathrm{TF}} \log(tf) \end{aligned}$$

4 Discussion

In this section, we review some of the important issues involved in the analysis presented in the previous section.

4.1 Coordination level as evidence

The study of coordination level gives convincing evidence that the weight of evidence provided by the number of query terms appearing in the document, $Coord = co$, is well modeled as a linear function of co. This conclusion is supported by analysis of individual queries, over a number of different data sets. Figure 5 shows plots of $w\hat{o}e(rel : co \mid q)$ as a function of coordination level, co, for all of the TREC 3 queries for which at least one document of AP88 was judged relevant. Remarkable regularity is evidenced by this plot. This is especially true when we take into account that the number of relevant documents corresponding to many of the queries is small, which would cause us to expect substantial variability in the data. Similar regularity was observed for other data sets that were examined.

4.2 Inverse document frequency as evidence

The modeling of idf is more problematic in two respects. First, the variance of residual log-odds across terms with similar idf values is large. The trend of increasing weight of evidence for increasing idf in figure 2(a) is clear, although the number of points that are undefined (circles along the bottom of the graph) must not be forgotten. The trend is also evident in the smoothed version of the plot (figure 2(b)) where all query terms, including those that do not appear in relevant documents, contribute to the visual effect. Even in the smoothed version, large variance is in display. Review of more coarsely smoothed plots has not helped much. This large variance makes it difficult to have confidence in the modeling decisions.

Second, although the general trend is quite robust, the magnitude of the effect and the exact form of the increase, seem to vary considerably across varying collections and query sets. More study will be required to arrive at a more thorough understanding of the nature of the evidence provided by the value of the idf feature.

4.3 Term frequency as evidence

There are two reasons that speak in favor of a log transformation of the term frequency variable. The shape of the curve shown in figure 3 strongly suggests a sharp decrease in the weight of evidence provided by one additional occurrence of a query term as term frequency

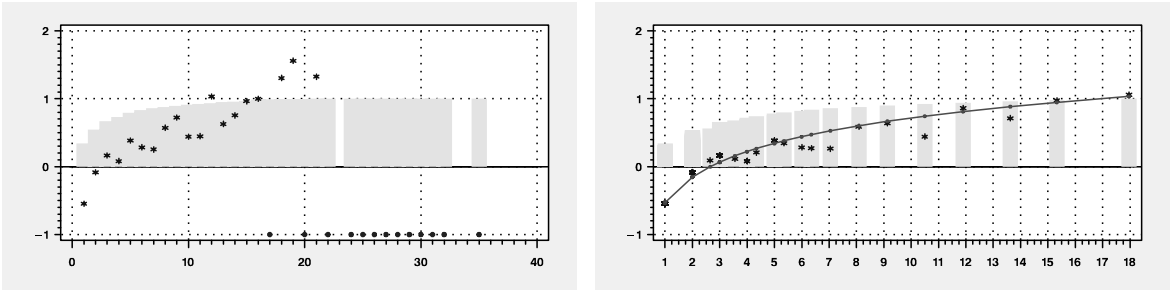

Figure 3: Residual log-odds as function of *tf*: a) unsmoothed b) smoothed with curve fit

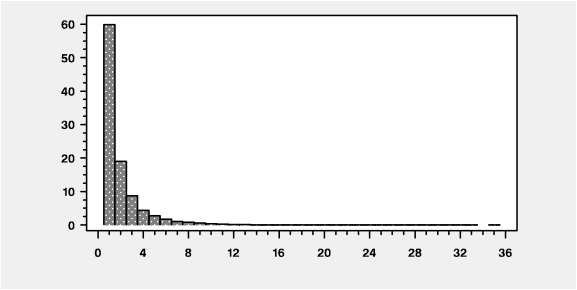

Figure 6: Histogram of *tf* values

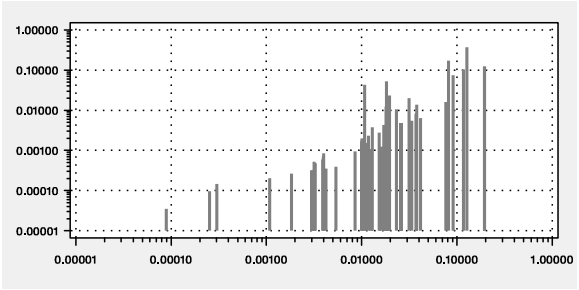

Figure 7: observed $p(rel)$ vs. expected $p(rel)$ for 50 bins

increases. All data sets studied exhibited this same behavior. Intuitively, this is what one would expect, and this intuition has been the inspiration for a number of ranking formulas that are used in IR research. One approach to the modeling of this effect that was tried, was treating the residual log-odds as a function of *tf* whose distance from a fixed maximum value decreases exponentially:

$$res \quad = \quad \beta_0 - \beta_1 \cdot e^{-\beta_2 \cdot tf}$$

By fixing β_0, and transforming the response variable, this exponential model could be converted to an equivalent linear model:

$$\beta_0 - \log(res) \quad = \quad \beta_1' - \beta_2 \cdot tf$$

The problem with this is that it requires an estimation of the asymptote, β_0. Examination of other data sets revealed a notable robustness in the general shape of the curve, but also unfortunate variability in the apparent value of β_0.

In retrospect, it becomes clear that a log transformation of *tf* should have been applied before even considering the shape of the approximating function. This is because of the highly skewed distribution of the Tf variable.

Both intuition and the histogram shown in figure 6. suggest that the difference between $Tf = 1$ and $Tf = 2$ should not be treated as equal to the difference between $Tf = 21$ and $Tf = 22$. This is an indication that a log transformation is likely to provide a more appropriate scale on which to analyze the data.

4.4 Probabilistic interpretation of RSV

One advantage of a probabilistic retrieval model is that the ranking status value will have a precise interpretation. This interpretation enables us to analyze the behavior of a system in a way that is different from methods traditionally used to evaluate system performance. The RSV produced by the M_{TICQ} model can be interpreted as a weight of evidence; specifically the conditional weight of evidence favoring the hypothesis that the document is relevant to the query. Adding that weight of evidence to the observed log-odds of relevance for the query gives a probability of relevance for the document. After sorting the query/document pairs by this probability of relevance, we can perform the same binning operation that was used for smoothing in the analysis of evidence. For each bin, we can calculate the fraction of documents in the bin that can be expected to be relevant based on the probability of relevance associated with each of the documents. Figure 7 shows a plot of the fraction of documents that are relevant for each bin against the fraction predicted for that bin. The plot is produced on log-log scale, since the fractions involved are small and span over three orders of magnitude.

The ability to produce a plot such as this is a valuable tool for information retrieval research. For example, from figure 7 we see that, while the model is doing well overall at predicting probabilities, there appears to be a tendency to underestimate probabilities at the low end of the scale. This says something about the way the model is behaving and gives direction to investigations into how modeling might be improved.

4.5 Preliminary performance evaluation

The primary goal of this research is to acquire a better understanding of the relation between query/document features and the probability of relevance. Subordinate to this objective, at the stage of the research, is the development of a ranking formula. Nonetheless, it is useful even at this point, to get a feel for how a ranking formula resulting from the analysis might perform. To test the M_{TICQ} model, the Inquery information retrieval system [3] was modified to apply a formula based on eq. 6 for the RSV calculation. Performance was compared to an unmodified version of Inquery as a baseline.

A series of tests were run with various parameter set-

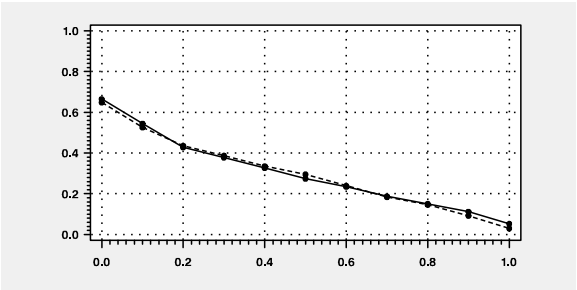

Figure 8: Recall-precision graph for TREC 3 queries on AP88

tings. Figure 8 shows an 11-point recall-precision graph for the M_{TICQ} model using the best parameter settings found. It is compared against the unmodified Inquery system. The test system is represented by the solid line, with a broken line used for the baseline. Performance is almost identical at all levels of recall. This is encouraging, giving reason to believe that the traditional multiplicative *tf-idf* formulation may ultimately yield to a probabilistic ranking formula that is founded on observable statistical regularities.

A number of caveats are in order. First, the test results shown in figure 8 correspond to testing the M_{TICQ} model on the same data set for which it was developed. However, test results on other data sets are promising.

Figure 9 shows results for two other document collections: ZIFF2 and WSJ89. In both cases, performance for the two systems were again quite comparable. The collections involved are similar to the AP88 collection whose analysis is discussed in this paper. The ZIFF2 test was run with the same TREC 3 queries that were used for analysis with the AP88 data. Here, the test system performs slightly better than the baseline system.

Interestingly, the test system performs well also on the WSJ89 test, even though both the collection and the query set, queries from TREC 1, were different from those used to develop the model. Individual collections of other types of articles have yet to be studied, but tests of the full TREC volumes 1 and 2 gave poor results. Thus, while preliminary testing can be said to give reason for hopefulness, research still remains to be done before a competitive ranking formula, robust over a wide range of document types, can be expected to emerge.

Second, it must be emphasized that traditional trial-and-error tuning methods were used for setting the parameters for these tests. The initial parameter values tried were those produced by the regressions. One by one each parameter was then allowed to vary and tests were run. In general, performance was found to be robust over a comfortable range of values, and the initial parameter settings were found to be reasonably close to optimal. The settings used for the tests reported here were those produced by the respective regressions for all parameters save one.

The one exception was the value for β_1^{CO}, the slope of the line modeling the weight of evidence, $woe(rel : co \mid q)$, provided by coordination level. In this case, the setting used was the one resulting from tuning. There has not been time, prior to the preparation of this article, to fully analyze the reasons for the poor performance of the regression setting for this parameter. It is on the top of the priority list for continued investigation.

Finally, the model has been developed for, and the tests have been run on, homogeneous collections of articles; whereas the baseline system has been designed to perform well over a heterogeneous mix of collections representing a wide range of document types. Presumably, state-of-the-art retrieval systems would perform better on news articles if they had been designed and tuned for performance on this more restrictive type of document. This should be kept in mind when comparing the performance of the test system to the baseline.

5 Future Work

The analysis of retrieval situations and experimentation with the resulting model give reason for a degree of confidence in the viability of the approach. Much work remains to be done if this potential is to be realized. In this section, we discuss the major steps that are planned for the continuation of this research.

5.1 Controlled retrieval environments

In this phase of our research, attention will continue to be focused on the modeling of a single class of documents – specifically news articles. Although there is a great need for retrieval systems that can be applied to a diverse mix of documents, there is also a need to be able to design systems for more restricted classes of documents.

More important is that research must be conducted in as controlled a setting as possible, if a greater understanding of the behavior of features used for retrieval is to be gained. We believe that extensive controlled studies, in restricted environments will be necessary, before insight into the factors involved in general purpose retrieval can be achieved.

5.2 Interaction effects

The M_{TICQ} model developed here treats evidence individually. To date, no attempt has been made to study possible interaction effects. For example, perhaps evidence provided by term frequency is different for rare and common terms. Initial analysis gives reason to believe that this is not the case, but more detailed study of this question, as well as the interaction effects that may exist among other predictor variables, need to be conducted.

5.3 Document length as a source of evidence

Many modern retrieval formulas take the length of a document into consideration. Normalization of term frequency based in some way on document length is common. Also there is evidence, at least for TREC data, that longer documents are more likely to be judged relevant [21, 20]. Both document length as a source of evidence by itself, and the possible interaction between term frequency and document length will be studied in the future.

5.4 Model fitting criteria

A major question mark in this research is how modeling decisions relate to traditional measures of retrieval performance when the model is the basis of a ranking formula. Minimizing mean square error of residual log-odds is a simple and intuitively satisfying approach to model fitting. It also permits the considerable machinery of regression analysis to be exploited. But, minimizing the mean square error does not translate necessarily to optimal performance when measured by average precision,

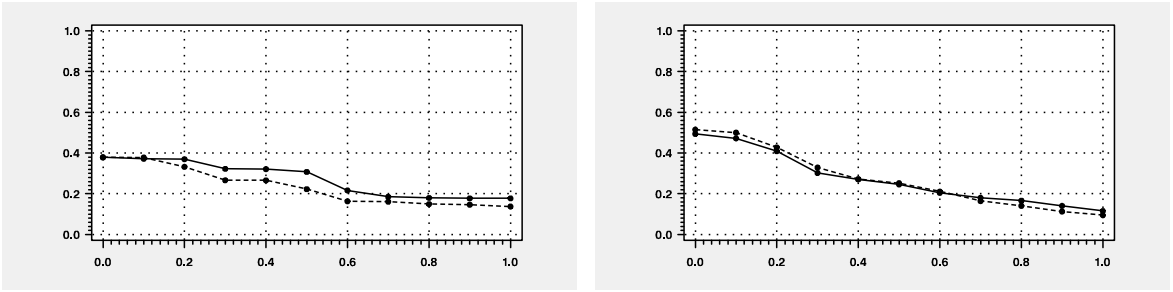

Figure 9: Recall-precision graphs: a) TREC 3 queries on ZIFF2 b) TREC 1 queries on WSJ89

for example. This issue is critical and will be a major focus of future research.

6 Conclusion

An approach to the analysis of query/document characteristics has been devised, and applied to TREC queries 151-200 and the AP88 collection. The analysis resulted in the development of a model of weight of evidence in favor of relevance provided by these features. The model was used as the basis of a ranking formula and preliminary retrieval experiments were run.

Initial indications suggest that retrieval strategies can be designed and evaluated based on the analysis of observed statistical regularities supported by a sound theoretical, probabilistic foundation. The model has also been seen to do a credible job of predicting the fraction of relevant documents that will be found to be relevant based on the ranking status values generated by a system based on the model.

Much work remains to be done with regard to: further theoretical analysis of the relation between weight of evidence and traditional measures of retrieval performance; more accurate modeling of conditional weight of evidence as a function of predictor variables; more detailed statistical analysis of these models (generation of confidence intervals, hypothesis testing, etc.); and more extensive experimentation.

7 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] C. D. Batty. The automatic generation of index languages. *Journal of Documentation*, 25(2):142–149, June 1969.

[2] A. Bookstein and D. R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 26(1):45–50, January-February 1975.

[3] J. P. Callan, W. B. Croft, and S. M. Harding. The inquery retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.

[4] 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.

[5] William S. Cooper. Some inconsistentencies and misnomers in probabilistic information retrieval. In A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan, editors, *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 57–61, Chicago, Illinois, USA, October 1991.

[6] 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.

[7] 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.

[8] Edward A. Fox. *Extending the Boolean and Vector Sapce Models of Information Retrieval with P-Norm Queries and Multiple Concept Types.* PhD thesis, Computer Science, Cornell University, 1983.

[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, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society – Series B.*, 22:319–331, 1960.

[13] Warren R. Greiff. A theory of term weighting based on exploratory data analysis. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998. ACM Press.

[14] Donna Harman. Overview of the first Text REtrieval Conference (TREC-3). In D. K. Harman, editor, *The Third Text REtreival Conference (TREC-3)*, pages 1–20, Gaithersburg, Md., April 1995. NIST Special Publication 500-225.

[15] Frederick Hartwig and Brian E. Dearing. *Exploratory Data Analysis*. Number 016 in 07. Sage Publications, Beverly Hills, 1979.

[16] Eugene L. Margulis. Modelling documents with multiple Poisson distributions. *Information Processing & Management*, 29(2):215–227, 1993.

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

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

[19] G. Salton and C. S. Yang. On the specification of term value in automatic indexing. *Journal of Documentation*, 29(4):351–371, 1973.

[20] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland, aug 1996. ACM Press.

[21] Amit Singhal, Gerard Salton, Mandar Mitra, and Chris Buckley. Document length normalization. *Information Processing & Management*, 32(5):619–633, September 1996.

[22] P. Srinivasan. On generalizing the two-Poisson model. *Journal of the American Society for Information Science*, 41(1):61–66, January 1990.

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

[24] 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.