MAXIMUM ENTROPY, WEIGHT OF EVIDENCE AND INFORMATION RETRIEVAL

A Dissertation Presented

by

WARREN R. GREIFF

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September, 1999

Department of Computer Science

MAXIMUM ENTROPY, WEIGHT OF EVIDENCE AND INFORMATION RETRIEVAL

A Dissertation Presented

by

WARREN R. GREIFF

Approved as to style and content by:

W. Bruce Croft, Chair

Jamie Callan, Member

Andrew G. Barto, Member

David Hosmer, Member

James F. Kurose, Department Chair
Department of Computer Science

To Paty and Sami:

gracias por estar a mi lado.

EPIGRAPH

The theory of probability is no more than a calculus of good sense. By this theory, we learn to appreciate precisely what a a sound mind feels through a kind of intuition often without realizing it. The theory leaves nothing arbitrary in choosing opinions or in making decisions, and we can always select, with the help of this theory, the most advantageous choice on our own. It is a refreshing supplement to the ignorance and feebleness of the human mind.

P. S. Laplace (1902), *A Philosophical Essay on Probabilities*. John Wiley & Sons, New York. Translated from the 6th French Edition, 1812.

ACKNOWLEDGMENTS

I would like to express my appreciation to Professor Bruce Croft for his support and encouragement. I am grateful, in particular, to have been able to work in an environment where I was free to explore areas of interest to me and pursue research questions in response to my personal perspective on the information retrieval problem.

I would also like to thank Professors Andy Barto, Jamie Callan, David Hosmer and James Allan for time they have so willingly spent with me discussing issues in Information Retrieval, Statistics and related areas.

Finally, I would like to thank my fellow students in the Information Retrieval Laboratory for sharing these last few years with me. I would like to give special thanks to Jay Ponte for the many engaging hours spent with me in front of a dusty old blackboard in a dark and dreary corner of the third floor.

ABSTRACT

MAXIMUM ENTROPY, WEIGHT OF EVIDENCE AND INFORMATION RETRIEVAL

SEPTEMBER, 1999

WARREN R. GREIFF

B.Sc., CASE INSTITUTE OF TECHNOLOGY

M.Sc., UNIVERSITY OF PENNSYLVANIA

M.Ed., ANTIOCH COLLEGE

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor W. Bruce Croft

The central theme of this dissertation is the statistical analysis of retrieval data. Features commonly used in modern retrieval systems are studied and modeled. The product of this analysis is a methodology for the study of retrieval data and the construction of probabilistic retrieval models. Model building is based on the formal concept of *weight of evidence*, which is a measure of how much our belief in a hypothesis (such as the relevance of a document) is increased as a result of the observation of the value of a random variable (for example, the number of times a query term appears in the document). Application of the methodology results in the development of a probabilistic model from which a ranking formula is derived. The *ranking status value* assigned to each document is equal to the weight of the evidence due to

the combination of features that have been observed. The resulting formula has two important properties: 1) it is decomposable, with each component corresponding to observed statistical regularities of retrieval situations; and 2) the value produced has a precise, empirically verifiable probabilistic interpretation. Experimental evidence is reported indicating that the ranking formula derived from the data analysis is able to produce retrieval performance comparable to that of a state of the art IR system.

In conjunction with the study of empirical data, a formal framework is developed which supports the approach to modeling that is used. The formalism is founded on the Maximum Entropy Principle. This principle states that the probability distribution that we attribute to an unknown stochastic process should be that which assumes the least consonant with constraints embodying the knowledge we do possess. Guided by this principle, a theory of weight of evidence is developed. In this theory additivity of weight of evidence is proved to be a characteristic of the maximum entropy distribution under general conditions on the form of the constraints. As well as serving as a justification for the modeling strategy adopted in the dissertation, two classical probabilistic retrieval models are shown to follow from the theory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

> *... the only physical quantities are those which can be measured and for which a measuring operation can be stated.*
>
> L. Brillouin, considering a "viewpoint repeatedly emphasized by Bridgman", in *Observation, Information and Imagination*, [9, p. 11].

Information retrieval is a successful technology which is having an increasing impact on both our professional and personal lives. Today, commercial and research systems are used by librarians, legal scholars, and government intelligence personnel at the same time that they see use as common tools by a burgeoning number of individuals in their everyday activities.

Although information can be presented in diverse forms ranging from tabular numeric data to graphical displays to photographic images to human speech, all of which are available electronically in large quantities, the term *information retrieval* (IR) as used in this dissertation shall refer specifically to the retrieval of *textual* information.

The specific task of interest will be the return of documents from a previously created, large collection of (digitally encoded) texts. The goal of the IR system will be considered to be the ordered presentation of these documents in response to an expression of a user's information need. For our purposes, the user will be restricted to represent the information need as a *query* in the form of a set of *terms*. So, for example, someone interested in reviewing articles concerning the trial of the president

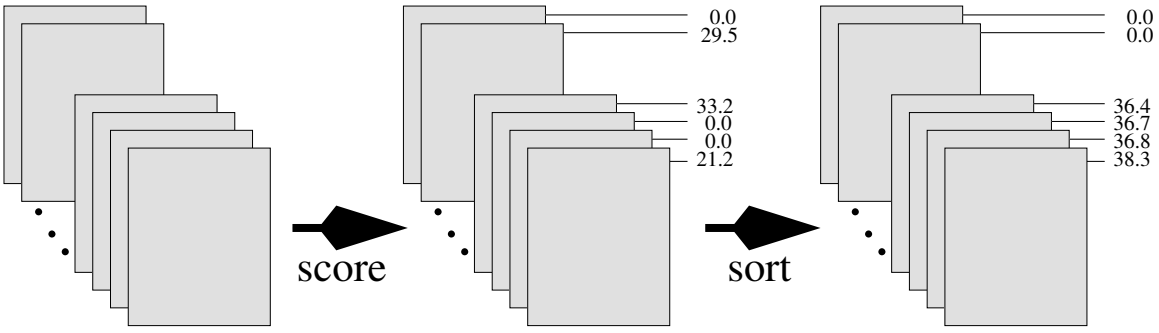

Figure 1.1. Scoring and ranking of documents

of the United States from a collection of 1999 news articles might present that need as:

Clinton impeachment proceedings

The specific task of interest in this research is *document ranking*, where the goal of the system is to order the documents of a collection. A successful system will, on average, do a good job of placing documents that the user judges to be *relevant* to the information need, high up in the *ranking*.

1.1 Basic Retrieval Strategies

In essence, the strategy employed by modern retrieval systems is straightforward. In response to an unstructured query such as *Clinton impeachment proceedings*, we can conceive of an IR system as assigning a score to each document, as shown in Figure 1.1. The score assigned to a document is often referred to as the *ranking status value* (RSV). In practical implementations, of course, only a small fraction of the documents are actually processed. The documents are then sorted according to the assigned scores, and presented to the user in order, highest scores first.

The score assigned to each document is some function of the query/document pair under consideration. Typically the score is a sum of values, with each value corresponding to one of the query terms appearing in the document. For the most part we may, for the purposes of this report treat *term* as synonymous with *word*. Here, *term*

shall be used because it encompasses something more general than simple words. For instance a term might be a phrase. If the representation of the information need were,

White House scandal

an IR system might treat *White House* as a single phrasal term rather than as two independent single-word terms. Presumably the user is interested in documents containing the phrase *White House* and not in documents containing both words *white* and *house* individually. The determination as to what is to be treated as a phrase can either be made by the user in the specification of the query, or automatically by the system based on some query modification algorithm.

By using a scoring function that adds values across all query terms appearing in a document, each query term that is present gets to contribute something to increasing the overall document score. Over the years, two things have become clear: 1) the contribution of a term that appears in the document being evaluated, but appears in few, if any, other documents of the collection should be greater than that of a term that is found in a large fraction of documents in a collection, and 2) the contribution of a term should be greater the more often it occurs in the document. That is, a contribution should increase inversely with document frequency and directly with term frequency. Here, *document frequency* (*df*) is understood as the number of documents in which the term occurs, and *term frequency* (*tf*) is the number of occurrences of the term in a given document.

1.2 Approaches to Information Retrieval

Information retrieval systems work because, for most information needs, they are able to order documents of a collection in such a way that documents relevant to the user's query appear, on average, much higher in the ranking than could be expected by chance. They work because they exploit statistical regularities of document collections and user queries.

In response to a query, the earliest systems returned the documents that contained the largest number of query terms. This was effective because a document that contains a query term is more likely to be relevant to the user. A document that contains many of the query terms is more likely to be relevant than a document that contains fewer of them. More modern systems also take into consideration the rarity of a query term and its frequency of occurrence in a document. All other things being equal, a document that contains rare query terms is more likely to be relevant than a document containing more common terms. In the same way, a document containing more occurrences of a query term is more likely to be relevant than a document with fewer occurrences of the same term. Out of four decades of research have evolved IR systems that take advantage of these statistical characteristics.

While modern retrieval systems have developed sophisticated means for exploiting statistical regularities, IR research has not, as a rule, tended to focus directly on the study of these regularities. Although research in IR is varied and a diversity of philosophies, strategies and techniques have been employed, two major trends may be discerned. For the purposes of this discussion, we may refer to these as the *engineering* and the *a priori modeling* approaches to IR research and system development.

In the engineering approach, intuitions concerning the nature of document collections and the behavior of users posing queries to an IR system are encoded in a (typically parameterized) retrieval algorithm. Experiments are run comparing a system incorporating this algorithm to a benchmark system; the result of varying parameter settings is studied; alternate versions of the algorithm are tried. If the research is successful, robust improvement to previous retrieval practice is realized and a better understanding of what is needed for effective retrieval performance results. Even in the absence of improved retrieval performance, new insight is often gained into the nature of the IR problem. Much of the progress of information retrieval is due to research of this nature.

The *a priori* modeling approach adopts a more theoretical, formal line of attack. In this case, the researcher attempts to formalize her intuitions in the form of *a priori* assumptions, and a theory, typically a probabilistic theory, of information retrieval is developed. From this theory an information retrieval strategy is usually derived. This strategy can then be implemented and tested. The proponents of this approach believe that the field of information retrieval is well served by the development of formal theoretical foundations. Formal theories promote precision in discourse and permit the application of deductive reasoning to the analysis of information retrieval questions. In Cooper's words, probabilistic theories "bring to bear ... a high degree of theoretical coherence and deductive power" [18, p. 242]. Cooper also emphasizes that a formal approach assists investigators to articulate their important underlying assumptions. "When the underlying theoretical postulates are known and clearly stated," he tell us, "their plausibility can be subjected to analysis" [p. 244].

In contrast to these, the research described in this dissertation may be considered a *data driven* approach. The goal of this research is the development of a model of IR document ranking based on observed statistical regularities of retrieval situations. It is similar to much work in probabilistic information retrieval in that the objective is to formally model the probability of a document being judged relevant to a query conditioned on the available evidence. It is significantly different from both the engineering and *a priori* modeling approaches in the emphasis that is placed on the study of existing retrieval data.

1.3 Research Statement

This dissertation research involves a detailed study of the statistical patterns on which modern information retrieval systems are based. The goal of this research is the development of a statistical model of IR document ranking. Techniques of *exploratory data analysis* (EDA) are employed to analyze how the probability of a document being

judged relevant is related to factors that are normally incorporated in the ranking formulas of modern IR systems. More formally, the goal will be to model the *weight of evidence* in favor of relevance given by the values of features associated with a query/document pair. Here the term "weight of evidence" is used in a formal sense which will be described fully in Section 3.1.

The significant research contributions that come out of this work are:

Data driven methodology: The principal contribution of this work is a data driven methodology for the analysis of evidence that may be considered for inclusion in a document ranking strategy. The methodology involves the adaptation of techniques of exploratory data analysis to the specific conditions encountered with regard to document ranking in information retrieval. These techniques include:

- the preparation of graphical displays
- the study of residuals
- smoothing of data
- transformation of variables

The data driven methodology centers on the analysis of the weights of evidence in favor of relevance given by various sources of evidence. The end product is a model for weight of evidence from which a ranking formula can be derived directly.

Formal theoretical framework: The data driven methodology and the modeling process are supported by a formal framework in terms of weight of evidence. The theory presented depends on a modeling principle based on the information theoretic notion of the entropy of a probability distribution. The keystone of this framework is a theorem in which it is proven that additivity of weight of

evidence follows from a general set of conditions respecting what is known about the sources of evidence.

Application of the methodology: The data driven methodology has been applied to the analysis of existing retrieval data and a probabilistic model has been produced.

Ranking formula: The result of the data analysis is a model of the weight of evidence in favor of relevance for the totality of evidence associated with a query/document pair being evaluated. This weight of evidence is converted directly into a formula for ranking. This formula has two significant properties:

Decomposable formula: The formula can be decomposed. Both the form of the decomposition and the details of the constituent parts correspond to observable statistical regularities.

Probabilistically interpretable RSV: The ranking status value has a precisely defined interpretation in terms of weight of evidence.

1.4 Organization of Dissertation

The remainder of the dissertation is organized as follows.

The following chapter reviews research spanning the last four decades that has focused on the probabilistic modeling of information retrieval. Emphasis is placed on efforts that are most closely related to the approach that is taken here. Chapter 3 gives brief descriptions of weight of evidence and exploratory data analysis, both of which are central to this research.

Chapter 4 presents an initial stage of the research in which attention was focused on the relationship between the document frequency for a term and the weight of evidence associated with its occurrence in a document. The course of the analysis is fully described. The result of the analysis is a novel theory for why inverse document

frequency weighting is useful for document ranking. Also resulting from this analysis is an hypothesis regarding how the classical inverse document frequency formula may be modified to improve retrieval performance. Substantial experimental evidence is given to support the hypothesis.

Chapter 5 describes the Maximum Entropy Principle and, based on it, derives a strong theoretical result concerning the additivity of weights of evidence. This result can be used to understand two classical probabilistic retrieval models in terms of the Maximum Entropy Principle. It also serves as a theoretical foundation for the modeling that is described in the succeeding chapter.

In Chapter 6, the methodology used for the study of inverse document frequency in Chapter 4 is extended to deal with multiple sources of evidence. Retrieval data is analyzed and a stochastic model is developed. From the model, a ranking formula is derived with the properties mentioned above. Experimental results are given for tests of the ranking formula which indicate the viability of the overall approach to IR research and system development.

In the final chapter, we return to discuss the contributions of this research. The major contributions are analyzed in greater depth and detailed contributions are itemized. The chapter closes with a discussion of possible research directions that may be explored as continuations of this work. The possibilities for extension of the work emanate from a vision of the data driven approach as a fundamentally sound, pragmatic methodology applicable to all aspects of information retrieval research, and that the work reported here is only the initiation of a far more extensive and encompassing research agenda.

CHAPTER 2

RELATED WORK

> *The great mistake of the Greek Philosophers was that they spent so much time in theory, so little in observation. But thought should be the aide of observation, not its substitute.*
>
> Will Durant in discussion of Francis Bacon, in *The Story Of Philosophy: The Lives And Opinions Of The Great Philosophers*, [28, p. 142]

The work discussed in this document is a natural outgrowth of IR research conducted over the last forty years. Much of previous thought and experimentation on the IR problem directly motivated early stages of the work. The relationship to other lines of research only became apparent after the main thrust of this dissertation project had taken form. This section reviews the major research directions in information retrieval that are related to this dissertation, with the explicit objective of placing it in the context of past and current information retrieval research.

2.1 Binary Independence Models

It is generally agreed that the seminal paper on probabilistic information retrieval was "On Relevance, Probabilistic Indexing and Information Retrieval", written in 1960 by Maron and Kuhns [78]. In this paper, the idea of probabilistic indexing is introduced. The authors point out that it is not certain that a document assigned a given index term will be deemed relevant when that index term is given as the subject of a search. Rather, there is a probability that the document will be found relevant to a search using the term. They propose that having an indexer weight the association

of an index term with her estimate of this probability will result in an index that "can characterize more precisely the information content of a document" [p. 220]. They go on to develop search strategies based on a combination of heuristic techniques, statistical estimation assumptions and probabilistic reasoning. The resulting algorithm is able to respond to information needs expressed as Boolean combinations of index terms and order retrieved documents by probability of relevance.

2.1.1 Binary Independence given Relevance Information

In 1972, Sparck Jones, convincingly demonstrated that the weighting of query terms can significantly improve retrieval performance compared to unweighted *coordination match* ranking [103]. The weighting formula she proposed was an approximation of:

$$w_{sj} = -\log \frac{n}{N} = \log \frac{N}{n} \tag{2.1}$$

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. Ever since, some formulation of what has come to be called *inverse document frequency* (*idf*), has been used as part of the ranking formula of modern information retrieval engines.

In a letter to the Journal of Documentation later that year, 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 [87]; thus providing theoretical arguments for the use of w_{sj}. Together, in 1976, Robertson and Sparck Jones presented the Binary Independence Model [89], in which terms are weighted by:

$$w_{rsj} = \log \frac{p(occ \mid rel) \cdot (1 - p(occ \mid \overline{rel}))}{(1 - p(occ \mid rel)) \cdot p(occ \mid \overline{rel})} \tag{2.2}$$

where $p(occ \mid rel)$ is the probability of the term occurring in relevant documents[1], and $p(occ \mid \overline{rel})$ 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.

2.1.2 Binary Independence in the Absence of Relevance Information

Applying the probabilistic approach of Robertson and Sparck Jones, Croft and Harper worked, in 1979, with an equivalent formulation of w_{rsj} [23]:

$$w_{rsj} = \log \frac{p(occ \mid rel)}{1 - p(occ \mid rel)} - \log \frac{p(occ \mid \overline{rel})}{1 - p(occ \mid \overline{rel})} \tag{2.3}$$

Their goal was the development of a probabilistically justified weighting formula that could be used in a retrieval setting in the absence of, or prior to, relevance feedback. They make two assumptions:

- 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" [23, p. 287]; and
- the probability, $p(occ \mid \overline{rel})$, 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:

$$w_{ch} = k + \log \frac{N - n}{n} \tag{2.4}$$

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

[1]For the purposes of exposition, the original notation used by the authors discussed in this document has been replaced by the notation used later in this document.

essentially equivalent to eq. 2.1 for all but very high frequency terms. They note that w_{ch} of eq. 2.4 encompasses the Sparck Jones *idf* weight, w_{rsj} of eq. 2.1, as a special case, and report experimental results indicating that the inclusion of a constant term can significantly improve retrieval performance.

A more detailed mathematical treatment of both the Binary Independence and Combination Match models will be given in Chapter 5 where its relationship to the Maximum Entropy Principle will be shown.

2.1.3 Combination Match Anomalies

Robertson and Walker [90] 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 eq. 2.4, will assume negative values for high frequency terms. These anomalies cause them to modify the assumption of equal probability of occurrence in relevant documents, in favor of an assumption that this probability "increases from a non-zero starting point to reach unity" [p. 19] for a term that appears in all documents.

2.1.4 Relation of Binary Independence Models to this Dissertation

Central to the Combination Match model of Croft and Harper is the assumption of constant $p(occ \mid rel)$ 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". In Chapter 5 we will see that the Combination Match Model follows from a reasonable set of constraints and the *Principle of Maximum Entropy* [47]. This principle, enunciated by Jaynes, states that the probability distribution that has the maximum entropy of all those that conform to a given set of constraints is "the least biased estimate possible on the given infor-

mation; i.e. it is maximally noncommittal with regard to missing information" [60, pg. 620]. In this context, the entropy of a distribution is to be understood in the information theoretic sense defined by Shannon [97].

The Robertson/Walker adjustment of the Combination Match formula allows for an increase in $p(occ \mid rel)$ 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 $p(occ \mid rel)$ 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 $p(occ \mid rel)$ for the lowest frequency terms must be very very small. Unfortunately Robertson and Walker find themselves restricted to values above 0.5. The data indicate that only fairly high frequency terms can be expected to appear in as many as half of the relevant documents.

2.2 Probabilistic Indexing

For the most part, the idea of ranking by an estimation of the probability of relevance, introduced by Maron & Kuhns in 1960, lay dormant for the rest of the decade. Research in this period tended to focus on automating the indexing process. Inspired by the research and writings in the late 1950's by H. P. Luhn [74, 75], 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 [5].

2.2.1 Poisson Models

More pertinent to this dissertation is work during this period that investigated the relationship between distribution characteristics of word occurrences and the value of words for the purpose of indexing. Some years later, Bookstein and Swanson

conjecture that the distribution of word occurrence in a collection might be well modeled as a mixture of Poisson distributions. The assumption is that with respect to a specific word, documents are partitioned into k classes [7]. Given that a document is in a given class, the distribution of word occurrences are assumed to follow a Poisson distribution, with a mean specific to that document class. The probability of finding j occurrences of a word in a randomly selected document of the collection is then given by:

$$p(tf = j) = \sum_{i=1}^{k} p(Class = i) \cdot p(tf = j \mid Class = i) = \sum_{i=1}^{k} \pi_i \frac{e^{\lambda_i} \lambda_i^j}{j!} \tag{2.5}$$

where π_i is the probability that the document will belong to class i, and λ_i is the mean of the associated Poisson distribution.

In a two-part paper [55, 56], Harter concentrates on mixtures of two Poisson distributions as models for the distribution of specialty words. Using the method of moments to estimate values of the parameters, π_1, λ_1, and λ_2 (π_2 is restricted to $1 - \pi_1$), models were fit for specialty terms from a collection of Freud's works. Once the parameters were fit, χ^2 goodness-of-fit tests for each word were performed. He found that for 80% of the 183 terms tested, the null hypothesis that $\lambda_2 = 0$ (i.e. that terms were generated from a simple Poisson distribution) was rejected at the .05 level. Using an *ad-hoc* variant of the χ^2 test, he concludes that, with a confidence level of .95, a 2-Poisson distribution provides a close fit to the observed frequencies for between 65% and 95% of 36 randomly selected specialty words. Harter goes on to introduce the measure:

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}} \tag{2.6}$$

which he believes to be a plausible measure of "the effectiveness of a word as a potential index term" [p. 204]. Comparison of histograms of z for the classes of specialty and non-specialty words indicates that z is "reasonably successful in identifying spe-

cialty words" [p. 205]. In Part II of the paper, Harter analyzes the effectiveness of:

$$\beta = p(d \in \text{Class 1} \mid j) + z \tag{2.7}$$

where "the first, $p(d \in \text{Class 1} \mid j)$ is an estimate of the relative level of treatment of the concept w in the document d, while the second, z is an estimate of the overall effectiveness of w as a potential index term" [p. 284]. Exactly what is meant by "overall effectiveness as a potential index term" is not quite clear, and no attempt is made to justify the summing of these two numbers. By analyzing recall/precision graphs, based on a manually prepared index as the standard of correct indexing, he shows that ranking by β is superior to ranking by the raw term frequency, j, and that the difference is statistically significant.

Years later, Srinivasan and then Margulis continued this line of investigation. Srinivasan considered 2-Poisson and 3-Poisson models [104]. Data analysis using the much larger INSPEC collection showed that 43% of the terms were well fit by a 2-Poisson distribution. Allowing for 3-Poisson distributions did not extend the number of words that could be modeled, however. Margulis examined the hypothesis that the occurrence distributions of high frequency words follows an n-Poisson distribution, where n is permitted to vary from 2 to 8 [76]. He repeats the pattern of the Harter and Srinivasan experiments, but expands the family of mixture distributions he considers; examines full length documents from larger collections; uses maximum likelihood estimates in place of the method of moments; and does not restrict his study to known specialty terms.

2.2.2 Relation of Poisson Modeling to this Dissertation

The research on characteristics of term frequency distributions for the purposes of automatic indexing, particularly the investigation of Poisson distributions, raises points that may be compared and contrasted with this dissertation. A major point of

comparison is the emphasis placed in this work on the examination of available data for the generation and verification of hypotheses. Harter, Srinivasan and Margulis concern themselves first and foremost with the question of how term frequencies are distributed. They rely heavily on established statistical procedures for both fitting parameters and determining goodness of fit. The statistical assumptions with regard to the random generation of the sample from a presumed population are explicitly specified in Harter's paper.

The work reported in this dissertation emulates the attitude adopted by these researchers with respect to the exploration and statistical analysis of available data. The concentration on statistical characteristics of retrieval can be contrasted with other probabilistic IR work, in which *a priori* assumptions lead to models which are immediately tested via retrieval experiments, without direct examination of the data.

Pertinent also are Harter's studies of the relationship between raw term frequencies and document length. He uses visual inspection of scatterplots, corroborated by analysis of the product-moment coefficient of correlation, to justify his use of raw, as opposed to normalized, term frequencies. With respect to the careful scrutiny of the form of variables to be used for the purposes of modeling, the work presented in this document once again follows the example set in this earlier research.

On the other hand, the approach taken here may be contrasted with Harter's attempts to use his Poisson analysis as the basis of an automatic indexing procedure that will support probabilistic retrieval. In this phase of his work, the hypothesized relationship between the two classes of documents, and relevance to a search, is left unexamined. Evaluation of the formula for ranking potential index terms is based solely on comparative recall/precision curves, with no direct statistical analysis of β (eq. 2.7) interpreted probabilistically.

Harter uses decision theoretic arguments to motivate the use of β for ranking candidate indexing terms. Unfortunately, expected values for the costs of retrieving a

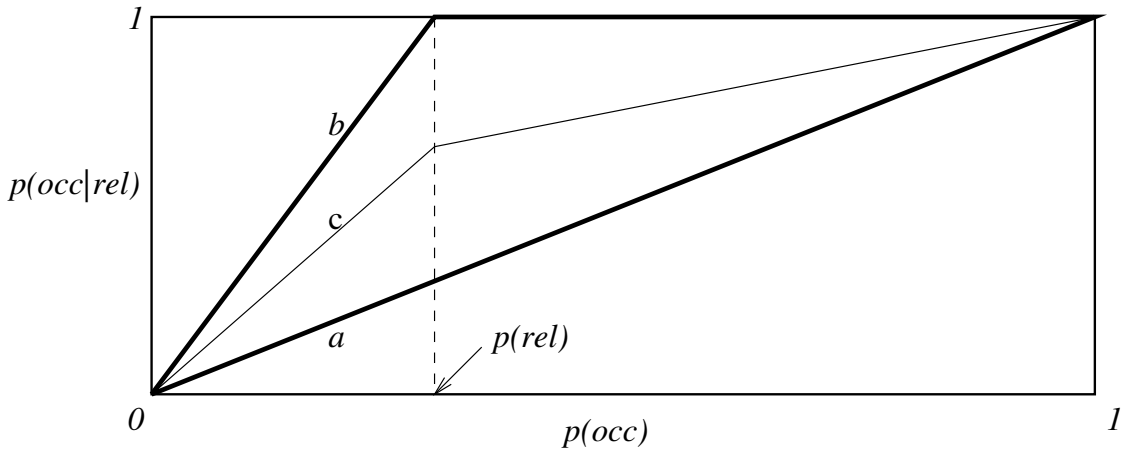

Figure 2.1. Term precision theory of $p(occ \mid rel)$ as a function of $p(occ)$ for a) random term; b) perfect term; c) linear combination of a and b.

non-relevant document, and failing to retrieve a relevant document, are not available. Nor are the probabilities, u_1 and u_2, that a searcher will find a class 1, respectively class 2, document to be relevant to a request. So Harter, replaces a function of these four parameters with the z measure of eq. 2.6. and then goes on to eliminate the decision procedure, in favor of ranking. This brings into question the purpose served by the decision theoretic machinery, given that in the end it is used neither for the derivation of the β measure, nor its evaluation in terms of the cost of erroneous decisions.

This dissertation extends the methodology employed by Harter in the first phase of his work to the direct study and explicit modeling of relevance as a function of observable, statistically analyzable features of a search situation.

2.3 Term Precision

In a series of papers in this same period, Salton and co-workers 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 [95]:

$$w_{tp} = \frac{p(occ \mid rel)}{1 - p(occ \mid rel)} \; / \; \frac{p(occ \mid \overline{rel})}{1 - p(occ \mid \overline{rel})} \tag{2.8}$$

Later, term precision was defined as the log of this quantity [96, 115], yielding the same weight as given by Robertson and Sparck Jones (eq. 2.2). The form they adopt for what amounts to $p(occ \mid rel)$ differs from that of both Croft/Harper and Robertson/Walker. The term precision model assumes a two-piece piecewise linear function with: $p(occ \mid rel) = 0$ at $p(occ) = 0$; $p(occ \mid rel) = 1$ at $p(occ) = 1$; and a change in slope at $p(occ) = p(rel)$ as shown in Figure 2.1. 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" [115, p. 159], sustained by 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.

2.3.1 Relation of Term Precision Model to this Dissertation

There is much similarity between the term relevance work and the research with regard to term weights presented in Chapter 4. Both are concerned with log relative odds as a weight; both result in an explanation of why *idf* weighting should be effective; and the relationship that is found in Chapter 4 between probability of occurrence in relevant documents vs. probability of occurrence in the whole collection is similar in general form, although significantly different in detail, to that assumed by Salton, *et al.* At the same time, there are major distinctions.

The major difference is one of approach. The term precision research starts from reasonable assumptions and goes on from there. The work here starts from the data. Techniques of exploratory data analysis are used to to look at the data to try to understand what it is that makes *idf* such a useful term weighting factor.

The model that results from exploratory analysis is different from that proposed in the term precision model. Term precision assumes that the probability of a query term

occurring in a relevant document can be modeled as a two-piece linear function of its probability of occurring in the collection as a whole. We will show in Chapter 4 that what they assume to be 2-piece linear is, in fact, exponential, and that the difference between linear and exponential growth is quite marked at the low-frequency end of the spectrum.

The term-weighting model of Chapter 4 is based on observed empirical evidence. The term precision work shows that *idf* is an approximation of the model that they assume to be right. But we conclude that their 2-piece linear assumption is not correct. If the goal is to understand how query terms really do operate, the difference between the two analyses is significant.

Finally, we will see in Chapter 4 that a model derived from examination of empirical data has predictive power. As a result of analyzing the data, a prediction as to how a modification of the traditional *idf* formulation might affect retrieval performance becomes evident. The prediction is borne out rather convincingly by the experiments that have been run. The approach taken in the term precision research could not have led to the kind of insights that can result from exploratory data analysis.

2.4 Regression

Regression strategies (explicitly or implicitly) assume a parameterized model and apply statistical techniques to fit the model to available data. Yu and Mizuno, for example, use linear regression to determine parameter settings for both a binary and non-binary model [116]. Fuhr and Buckley have used a least-square error criterion to determine coefficients for a polynomial weighting function of term-document pair descriptor variables [35, 34].

A group at the University of California, Berkeley has conducted extensive research into the use of logistic regression [37, 19]. Logistic regression is generally considered

a more 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.

Logistic regression [58] models the probability of a binary response variable. The logistic function of the probability is assumed to be a linear function of a pre-defined set of explanatory variables. This logistic function of the probability is formally equivalent to the log-odds, which plays a central role in weight of evidence defined in section 3.1, as well as the binary independence models as described in Section 2.1. At Berkeley, they have experimented with fitting the logistic as a function of explanatory variables that have been used in other investigations. Variables studied have included document frequency, term frequency in the document, and term frequency in the query [37]. Experimentation has also involved a variety of transformations of these variables. Statistical diagnostics and goodness-of-fit tests have been studied to determine which variables to include in the model and what transformation of these variables, if any, should be applied.

Most of the work at Berkeley has focused on the development of a term weighting strategy [19, 20, 38, 39]. As initially conceived, however, the modeling of term weights was only one component of a more encompassing *staged logistic regression* strategy. Staged logistic regression is a technique whereby multiple logistic regressions are performed, with each regression based on the results of regressions performed at previous stages [21]. In particular, the focus is on a two-stage process. In the first stage, logistic regression is used to develop a model for the probability of relevance as a function of a number of predictor variables. Once the parameters for this model are fit, corresponding term weights can be calculated as a function of pre-determined query/document features. These term weights are then used for ranking and a second

stage of logistic regression is effected in order to fit a model for the probability of relevance with the ranking scores, together with other variables, such as query length, as predictors.

2.4.1 Relation of Regression Research to this Dissertation

Regression models, like the others that have been discussed, result from *a priori* reasoning. This research does have a more empirical flavor. Data is used so that parameters can be estimated. That is to say, so that a specific 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? 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.

There are a number of distinctions between the work reported here and the logistic regression research conducted at Berkeley. First, this work has relied heavily on exploratory analysis of the data prior to attempting to fit a model. In particular, we study graphical representations of the data in order to become familiar with its behavior, search for underlying regularity and take advantage of insights that may be offered by unexpected patterns. The use of exploratory analysis has been an essential component of both the modeling of weight of evidence of term occurrence as a function of query/document features for individual query terms, and the modeling of how these weights are to be combined, over the set of terms in the query.

Another distinction between the two lines of research relates to the use of an additive model, both with regard to the linear combination of distinct facets of the evidence, such as the *idf* and *tf* values for a given term, and with regard to the linear combination of the weights associated with the different terms of the query. The work at Berkeley starts with the family of models and proceeds to determine the variable

forms to be included and fit the parameters. The approach taken here is founded, instead, on the Maximum Entropy Principle. In Chapter 5 it is shown that additive models follow from this principle and the distinction between assuming additivity and viewing the model as the probability distribution that maximizes entropy consistent with a set of constraints drawn from empirical observation is discussed.

Finally, and perhaps most important, the overall goal of the research presented here is different from that conducted at Berkeley. In their work, logistic regression has been used primarily as the basis of a learning algorithm. Based on training data, parameters for a general family of models are fit for a specific corpus with the goal of using the learned parameters for evaluation of previously unseen queries. In this work, regression is as much an exploratory tool, with emphasis on the discovery of underlying statistical patterns.

A more in depth understanding of the variables involved in the retrieval process should be useful for the development of improved retrieval strategies, including, potentially, learning algorithms based on fitting parameterized models. However, it is theory formation and not only improved retrieval that is the direct objective. The goal is not limited to the computation of coefficients that can be plugged into a general formula and used, perhaps very effectively, in a parameterized search engine. The goal is a model, the components of which can be understood in terms of intuitively reasonable statistical properties of retrieval.

2.5 Inference Network

In his doctoral dissertation in 1990 [109], Howard Turtle developed a probabilistic model for information retrieval formulated in terms of a Bayesian Network [71, 82, 13]. The inference network is a general framework which makes possible the consideration of multiple sources of evidence in the process of ranking documents in response to a user's information need. Evidence due to multiple document representations

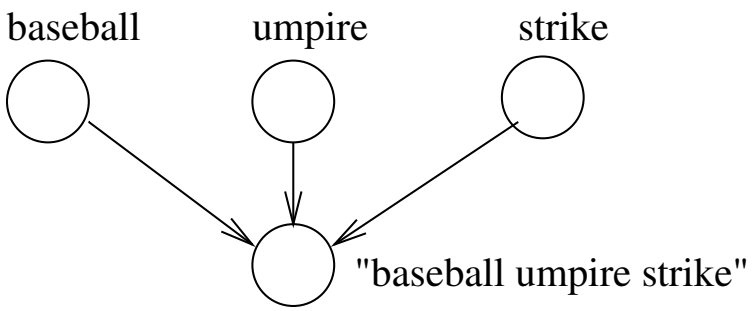

Figure 2.2. A query node dependent on three concept nodes

(e.g. titles, abstracts, document bodies, manually produced indices); multiple query formulations (e.g. Boolean, natural language) and even multiple belief systems can be combined in a principled way. An attractive aspect of the inference network approach is that it provides a direct, natural, computationally efficient, probabilistically motivated method for modeling query operators.

The nodes of the inference network correspond to propositions and are divided into two parts: the *document sub-network*, and the *query sub-network*. In the document sub-network, the propositions associated with the nodes pertain to the observation of: documents; representations of the documents; and representations of document content. Nodes of the query sub-network correspond to propositions regarding: the presence of query concepts; the satisfaction of queries; and the satisfaction of information needs.

This dissertation is directly concerned with only the *concept* and *query* nodes, and so this description will focus on this part of the network. In [108], Turtle and Croft provide a more general description of the inference network framework; and Callan, Croft and Harding discuss the implementation of the inference network in the INQUERY retrieval system in [11].

Within the inference network formulation, a concept node is associated with the presence of a "concept" in the document currently being analyzed. For unstructured queries, there is exactly one query node in the network. This query node is connected

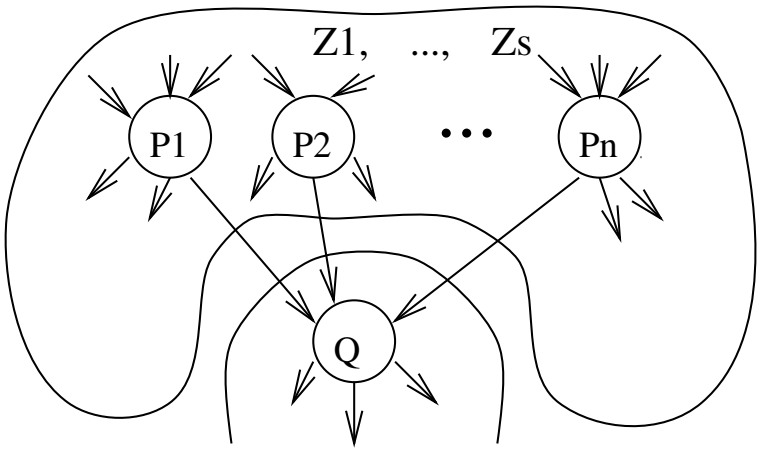

Figure 2.3. Conditional independence encoded in a Bayesian Network

to the concept nodes corresponding to the terms of the query. For example, Figure 2.2 shows the relevant part of the network for the query "`baseball umpire strike`". The leftmost node corresponds to a proposition asserting the presence in the document of the concept associated with the word `baseball`. Similarly, there are nodes for the `umpire` and `strike` concepts. The query node is associated with the proposition that the user's query is satisfied.

2.5.1 Bayes Nets

In general, a Bayesian Network encodes a joint probability distribution. The nodes of the network correspond to random variables. In the INQUERY inference network, each random variable may assume a value of either true or false. The topology of the network is interpreted as encoding a set of conditional independence relations among the variables. If the nodes corresponding to the variables, $P_1, \ldots, P_n$, are the immediate predecessors (*parents*) of a node, Q, and $Z_1, ..., Z_s$ are all other nodes that are not *descendants* of (i.e. are not reachable from) Q, as shown in Figure 2.3, then Q is considered to be conditionally independent of $Z_1, ..., Z_s$ given $P_1, \ldots, P_n$:

$$p(Q \mid P_1, \ldots, P_n, Z_1, \ldots, Z_s) = p(Q \mid P_1, \ldots, P_n)$$

Given probabilities for the root nodes (i.e. nodes with no parents), the network may be processed in a top-down fashion in order to produce the probabilities relevant to each of its nodes. As a consequence of the conditional independence assumptions implicit in the topology of a Bayesian Network, once the probabilities, $p_1, \ldots, p_n$, have been produced for the parents of a node, Q, the probability that Q assumes the value $y \in D_q$ is given by:

$$p(Q = y) = \sum_{x_1 \in D_1, \ldots, x_n \in D_n} p(Q = y \mid P_1 = x_1, \ldots, P_n = x_n) p(P_1 = x_1) \cdots p(P_n = x_n)$$

where $D_1, \ldots, D_n$, D_q are the sets of values that may be assumed by the variables, $P_1, \ldots, P_n$, and Q, respectively. For the INQUERY inference network, $D_1 = D_2 = \ldots D_n = D_q = \{\text{true,false}\}$.

2.5.2 Binary Valued Random Variables

In INQUERY each node corresponds to a proposition; that is, a variable that may take on one of two values: *true* or *false*. For example, in Figure 2.3, each P_i might correspond to the proposition that some document under consideration is about some concept, c_i, while Q corresponds to the proposition that the query is satisfied.

Since all the variables are binary valued in INQUERY, the dependence of a child on its parents can be given via the specification of:

$$p(Q \text{ is true} \mid P_1 = b_1, \ldots, P_n = b_n) \qquad \text{and}$$
$$p(Q \text{ is false} \mid P_1 = b_1, \ldots, P_n = b_n)$$

for each:

$$< b_1, \ldots, b_n > \in \{\text{true}, \text{false}\}^n$$

Equivalently, the values:

$$\begin{aligned} \bar{\alpha}_R &= p(Q \text{ is false} \mid & i \in R \Rightarrow P_i \text{ is true} \\ & & i \notin R \Rightarrow P_i \text{ is false}) \qquad \text{and} \\ \alpha_R &= p(Q \text{ is true} \mid & i \in R \Rightarrow P_i \text{ is true} \\ & & i \notin R \Rightarrow P_i \text{ is false}) \end{aligned}$$

must be specified for every possible subset, R, of $\{1, \ldots, n\}$. In terms of these conditional properties, the probability that a child node is true can be calculated once the probabilities of truth, $p_1, \ldots, p_n$, are known for each of its parent nodes:

$$\begin{aligned} p(Q \text{ is false}) &= \sum_{R \subseteq \{1,\ldots,n\}} \bar{\alpha}_R \prod_{i \in R} p_i \prod_{i \notin R} (1 - p_i) \qquad \text{and} \\ p(Q \text{ is true}) &= \sum_{R \subseteq \{1,\ldots,n\}} \alpha_R \prod_{i \in R} p_i \prod_{i \notin R} (1 - p_i) \end{aligned}$$

The set of coefficients involved is conveniently organized as a 2×2^n matrix:

	$P_{0\ldots000}$	$P_{0\ldots001}$	$P_{0\ldots010}$	$\ldots$	$P_{1\ldots111}$
Q false	$\bar{\alpha}_{0\ldots000}$	$\bar{\alpha}_{0\ldots001}$	$\bar{\alpha}_{0\ldots010}$	$\ldots$	$\bar{\alpha}_{1\ldots111}$
Q true	$\alpha_{0\ldots000}$	$\alpha_{0\ldots001}$	$\alpha_{0\ldots010}$	$\ldots$	$\alpha_{1\ldots111}$

where $\alpha_{b_1,b_2,\ldots b_n}$ is the probability that Q is true subject to the condition that the parents P_i such that $b_i = 1$ are true, and the parents P_i such that $b_i = 0$ are false; $\bar{\alpha}_{b_1,b_2,\ldots b_n}$ is the corresponding probability that Q is false and is equal to $1 - \alpha_{b_1,b_2,\ldots b_n}$. This matrix, known as a *link matrix*, may be visualized as linking the child node with the parent nodes as is shown in Figure 2.4

2.5.3 Link Matrices As Query Operators

INQUERY examines documents one by one. For each, the inference network is used to evaluate evidence that the document satisfies an information need expressed by the

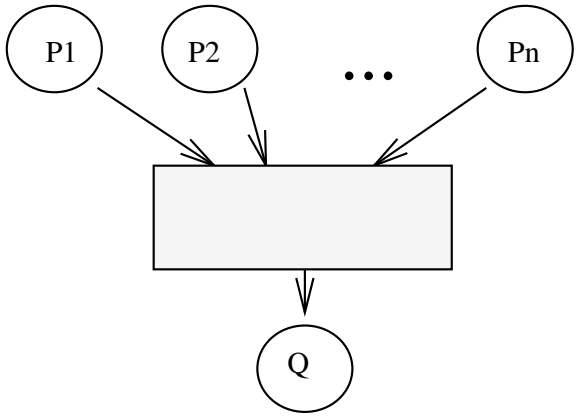

Figure 2.4. *Link matrix* links child to parents

user. A given link matrix form can be viewed as defining an operator for combining evidence. For example, suppose the propositions P_1, P_2, P_3, state that three queries q_1, q_2, q_3, respectively, have been (in some sense) *satisfied* by the document currently under scrutiny. A link matrix connecting the parents nodes, P_1, P_2, P_3, with the child node, Q, can be viewed as a way of forming a query, q, that is a composite of the individual sub-queries. The child node, Q, would correspond to the proposition that the combined query, q, has been satisfied. The specification of the coefficients of the link matrix defines the way the sub-queries are *combined* in that it gives all the information necessary for determining the probability that q has been satisfied given the probabilities, p_1, p_2, p_3, that the individual sub-queries have been satisfied.

In the example of Figure 2.2, this means that once the three probabilities, p_b, p_u, p_s that the document under scrutiny is about `baseball`, `umpire` and `strike` respectively, have been determined, the probability that the document is relevant to the query can be calculated. The computation requires that the requisite conditional probabilities have been specified. One such probability, for example, is $p(Q \mid \bar{B}U\bar{S})$, the probability that the query is satisfied, given that the document is about `umpire`, but not about either `baseball` or `strike`. There are 8 possible truth assignments for the set of three variables associated with the concept nodes of this example. The

probability that the query is satisfied is then given by:

$$
\begin{aligned}
p(Q) \;=\;& p(Q \mid \bar{B}\bar{U}\bar{S}) \cdot (1-p_b) \cdot (1-p_u) \cdot (1-p_s) \\
+\;& p(Q \mid \bar{B}\bar{U}S) \cdot (1-p_b) \cdot (1-p_u) \cdot p_s \\
+\;& p(Q \mid \bar{B}U\bar{S}) \cdot (1-p_b) \cdot p_u \cdot (1-p_s) \\
& \vdots \\
+\;& p(Q \mid BUS) \cdot p_b \cdot p_u \cdot p_s
\end{aligned}
\tag{2.9}
$$

One, admittedly arbitrary, way of defining a 3-ary query composition operator for forming the query, q, from the sub-queries q_1, q_2, q_3, might be to specify that q is to be considered satisfied with:

- 80% probability if q_1 is satisfied, independent of the whether or not q_2 and q_3 are satisfied;
- 50% probability if q_1 is not satisfied, but q_2 and q_3 are both satisfied
- 10% probability in any other situation.

This particular operator corresponds to the link matrix[2]:

P_{000}	P_{001}	P_{010}	P_{011}	P_{100}	P_{101}	P_{110}	P_{111}
.1	.1	.1	.5	.8	.8	.8	.8

Where the column under P_{011}, for example, gives the probability that Q is true given that P_1 is false, P_2 is true, and P_3 is true. Clearly, this link matrix has the desired effect. Typically, none of the parents will be known to be either true or false with certainty. Rather, evidence corresponding to each of P_1, P_2, P_3 will, in

[2]Since each column must sum to 1.0, only the row corresponding to Q = true is shown.

general, be estimated to be present with certain probabilities: p_1, p_2, p_3. Given these probabilities, the probability that Q is true can be calculated as:

$$\begin{aligned} p(Q \text{ is true}) = \quad & .1 \cdot \bar{p}_1\bar{p}_2\bar{p}_3 + .1 \cdot \bar{p}_1\bar{p}_2p_3 + .1 \cdot \bar{p}_1p_2\bar{p}_3 + .5 \cdot \bar{p}_1p_2p_3 \\ + \quad & .8 \cdot p_1\bar{p}_2\bar{p}_3 + .8 \cdot p_1\bar{p}_2p_3 + .8 \cdot p_1p_2\bar{p}_3 + .8 \cdot p_1p_2p_3 \end{aligned}$$

2.5.4 Relation of the Inference Network to this Dissertation

The initial motivation for the current research direction was earlier research pertaining to the use of inference networks in information retrieval. The objective of this research was to model Boolean query operators in terms of link matrix components of a Bayesian network. It resulted in the development of a class of link matrices called PIC matrices and a $O(n^2)$ algorithm for the evaluation of arbitrary matrices of this class for arbitrary inputs.

Experimentation demonstrated that this approach to modeling Boolean operators can be as effective as the previously reported *pnorm* formulation [33, 93, 94]. Whereas the *pnorm* formulation was developed within the vector space model [92], the PIC matrix approach is integrated in the inference network framework and has a probabilistic interpretation. The PIC matrix work is reported in [45], and in greater detail in [46].

While the modeling of Boolean query operators in the inference network was successful, problems were encountered in later attempts to extend this work. Similar difficulties were encountered when attempting to apply the PIC matrices to the retrieval of unstructured queries, and once again in the context of combining query expansion terms with original forms of queries automatically generated from descriptions of information needs.

Although the inference network is a probabilistic framework, there is no clear frequentist interpretation of the input, output or intermediate values manipulated by the system. This impedes the ability of the theory to guide research. Although, a

purely subjective Bayesian interpretation can be given to these values, this does not provide a concrete basis for determining the adequacy of the model. For example, the correct functioning of the system cannot be analyzed by comparing output values to proportions of relevant documents calculated over large samples. It is all the more infeasible to correlate the probabilities for the latent variables corresponding to intermediate nodes of the network with empirically observable frequencies.

2.6 Ranking by the Probability of the Query

Recently, Ponte and Croft have turned traditional probabilistic modeling on its head [85]. In their approach, documents are not ranked in order of the probability of relevance. Relevance is not modeled at all. Rather, each document is assumed to have been generated by a *language model*. The language model corresponds to a probability distribution over the collection vocabulary. For each document in the collection, the term occurrence pattern for the document is used to estimate the language model from which the document was produced.

A query submitted to this system is assumed to have been generated from a language model as well. When a query is submitted, documents are ranked according to the probability that the query was produced from the language model estimated to correspond to the document.

This model has achieved ad-hoc retrieval performance comparable to state-of-the-art systems. Ponte has also been successful in integrating relevance feedback into the model and has shown that it performs as well as existing approaches on the TREC routing task [84]. Local feedback has also been integrated easily into the model and experimental results again confirm the viability of the approach.

2.6.1 Relation of the Ponte/Croft Model to this Dissertation

An important component of the algorithm used in the Ponte/Croft system is the use of *smoothing* in the estimation procedure. The simplest, and most common, approach to estimating the probability of term occurrence for a document is to use the maximum likelihood estimator obtained by simply dividing the number of times a term occurs by the total number of occurrences over all terms appearing in the document. A more robust approach is to smooth this estimate by taking a weighted average of it with an estimate obtained from a larger sample. In the Ponte/Croft system a weighted average is taken of the estimate based on the document alone with an estimate based on the distribution of the term over the entire collection.

This smoothing method plays a significant role in producing the observed retrieval performance. Smoothing is also used heavily in this dissertation. Non-parametric smoothing techniques are used to produce regression curves for the purposes of exploratory data analysis. Also, in the section on future work, the possibility of research into the application of empirical Bayes analysis [77, 12] to the modeling of weight of evidence is discussed. The use of empirical Bayes techniques is similar to the approach taken in Ponte's work in that it utilizes a larger source of data to reduce the variance of estimates for parameters associated with a subset of the given sample.

CHAPTER 3

WEIGHT OF EVIDENCE AND EXPLORATORY DATA ANALYSIS

> *. . . philosophy is like horseradish. It is good if taken in small amounts in combination with other things. But it is not good in large amounts by itself. The risk with philosophy, as with horseradish, is the temptation to use ever stronger concentrations to maintain the sensation of that first taste. Soon you are serving up pure horseradish.*
>
> T. Seidenfeld paraphrasing Laura of Pasternak's *Dr. Zhivago* (chapter 13, section 16) in comment on *Weight of Evidence* by I. J. Good, [43].

The principal object of study in this thesis is the weight of evidence in favor of relevance provided by the values of features extracted from a retrieval situation. Section 3.1 gives the formal definition for weight of evidence and motivation for it. The method of study in this work relies heavily on statistical techniques that have been developed in recent years and are collectively referred to as exploratory data analysis. Section 3.2 provides a brief introduction of this approach to scientific understanding and hypothesis formation.

3.1 Weight of Evidence

We begin in Section 3.1.1 by giving the formal definition for the concept of weight of evidence. This is followed in Section 3.1.2 by the presentation of desiderata for the concept from which the definition can be derived. Sections 3.1.3 and 3.1.4 discuss interesting properties of weight of evidence as defined here and the connection of the concept to previous research in the area of information retrieval.

3.1.1 Formal definition of *Weight of Evidence*

I. J. Good formally defines the weight in favor of a hypothesis, h, provided by evidence, e, as [43, 41]:

$$woe(h:e) = \log \frac{O(h|e)}{O(h)} \tag{3.1}$$

where

$$O(h) = \frac{p(h)}{p(\bar{h})} = \frac{p(h)}{1-p(h)} \tag{3.2}$$

is the *prior* odds of the hypothesis, h being true, and

$$O(h|e) = \frac{p(h \mid e)}{p(\bar{h} \mid e)} = \frac{p(h \mid e)}{1-p(h \mid e)} \tag{3.3}$$

is the posterior odds of the hypothesis h being true conditioned on the evidence e having been observed. He believes this is a concept "almost as important as that of probability itself" [41, p. 249].

In [42, chap. 4], Good points out that, in various guises, the notion of weight of evidence had appeared in the work of others. As early as 1878, the quantity given in eq. 3.1 appears in the work of the philosopher Charles Sanders Pierce. Good credits Pierce with the original use of the term *weight of evidence*. More recently, Minsky and Selfridge also refer to this quantity and call it *weight of evidence* as well [80]. Turing had labeled the quantity, $\frac{O(h|e)}{O(h)}$, as the *factor in favor of the hypothesis h provided by the evidence e*, and Harold Jeffries made much use of the concept, referring to it as *support* [65].

Weight of evidence is related to Keynes's concept of the *amount of information*, which he defined as the log of $p(e \wedge h)/p(e)p(h)$ [42, chap. 11]. This is more commonly referred to today as *mutual information* and is discussed in Section 4.3 in connection with the relationship between weight of evidence and inverse document frequency. Keynes also used the term *weight of evidence*, but in a different sense from that used by Pierce, Minsky & Selfridge, and Good [42, chap. 15].

Although, we will primarily concentrate on the concept of weight of evidence as it is defined in eq. 3.1, two generalizations will also play an important role in this thesis.

First, weight of evidence can be conditional. That is, attention may be restricted to some sub-space of the full event space. The notation that will be used for conditional weight of evidence will be:

$$woe(h : e \mid c) \quad = \quad \log \frac{O(h|e,c)}{O(h|c)} \tag{3.4}$$

Second, weight of evidence, as given in eq. 3.1 and eq. 3.4 implicitly contrasts the hypothesis h against its negation $\bar{h}$. Often, however, we are interested in weight of evidence in favor of one hypothesis of interest, h, relative to another hypothesis of interest, h', that is not its negation. To meet this need we introduce the notation:

$$woe(h/h' : e \mid c) \quad = \quad woe(h : e \mid h \vee h', c) \tag{3.5}$$

By restricting the event space to that for which either h or h' holds, we have effectively defined the weight of evidence *in favor of* h *against* h' provided by e, (and, in the general case, conditioned on c).

3.1.2 Desiderata for a Concept of *Weight of Evidence*

Good elucidates three simple, natural desiderata for the formalization of the notion of weight of evidence [44, 43].

1. **Weight of evidence is some function of the likelihoods:**

$$woe(h : e) = f[p(e \mid h), p(e \mid \bar{h})] \tag{3.6}$$

   For example, let us suppose that a document is evaluated with respect to the query,

Clinton impeachment proceedings

and we discover the term *Hyde* in the document:

$$h = \text{document is relevant to the query}$$

$$e = \textit{Hyde} \text{ occurs in the document}$$

This first criterion states that the weight of evidence provided by finding *Hyde* in the document should be some function of the likelihood of finding *Hyde* in a document that is relevant to the query and the (for this example, presumably lower) likelihood of finding *Hyde* in a document that is not relevant to the query.

2. **The final (posterior) probability is a function of the initial (prior) probability and the weight of evidence:**

$$p(h \mid e) = g[p(h), woe(h : e)] \tag{3.7}$$

In the context of the same example, this states that the probability associated with the document being about the impeachment proceedings after having observed that the document contains the term, *Hyde*, should be a function of 1) the probability associated with the document being relevant to the query prior to obtaining knowledge as to the terms it contains, and 2) the weight of evidence associated with finding *Hyde*.

3. **Weight of evidence is additive:**

$$woe(h : e_1 \wedge e_2) = woe(h : e_1) + woe(h : e_2 \mid e_1) \tag{3.8}$$

This property states that the weight in favor of a hypothesis provided by two sources of evidence taken together 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. The weight of the

second piece of evidence is conditioned on the first in the sense that, $woe(h : e_2)$ is calculated on the subspace corresponding to the event, e_1. If, for example, the two pieces of evidence are:

$$e_1 = \textit{Hyde} \text{ occurs in the document}$$
$$e_2 = \textit{Henry} \text{ occurs in the document}$$

then the weight of evidence provided by finding both of the terms, *Hyde* and *Henry*, in the document is the sum of the evidence given by finding *Hyde* and the evidence given by finding *Henry*. The weight of evidence provided by occurrence of the term, *Henry*, is conditioned on having previously taken into consideration the evidence provided by encountering *Hyde* in the document.

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. 3.1:

$$woe(h : e) = \log \frac{O(h|e)}{O(h)} \tag{3.9}$$

The constant factor may be absorbed in the base of the logarithm. For the purposes of this dissertation all logarithms will be understood to be in base 10, which will make the scales shown on graphs easier to interpret.

3.1.3 Properties of Weight of Evidence

The formal definition of weight of evidence, together with conceptualization in terms of log-odds in place of standard probabilities, has a number of interesting, intuitively pleasing, and useful properties.

Weight of evidence adds to belief: From eq. 3.1, 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 (validity of a scientific theory, guilt of an alleged criminal, relevance

of a document to an information need) is equal to our initial belief plus the weight of whatever evidence we are presented with.

For example, our final belief that a given document will be about the impeachment proceedings will equal our (initial) belief that an arbitrary document will be about the impeachment plus the weight of evidence, such as the occurrence of the term *Hyde*, observed once the document is examined.

Weight of evidence can be positive, negative or zero: Positive *woe* causes our belief, in the form of log-odds, to increase; negative *woe* results in a decrease in our belief; and a weight of zero leaves the log-odds unaffected. Log-odds, $\log O(h|e)$, is a one-to-one, monotonically increasing, function of probability, $p(h \mid e)$. Hence, increasing, decreasing and stable log-odds correspond to increasing, decreasing and stable probabilities, respectively.

For example, presumably the weight,

$$woe(\text{about impeachment} : \textit{Henry} \text{ occurs} \mid \textit{Hyde} \text{ occurs})$$

would be positive, whereas the weight,

$$woe(\text{about impeachment} : \textit{Jekyll} \text{ occurs} \mid \textit{Hyde} \text{ occurs})$$

would be negative since the occurrence of *Jekyll* gives reason to believe *Hyde* does not refer to the congressman.

Log-odds can be positive, negative or zero: This is true of the log-odds itself, either prior to or resulting from, the accumulation of evidence. Log-odds of zero is equivalent to a probability of $\frac{1}{2}$. It is associated with a hypothesis that is neither favored nor disfavored in comparison to its contradiction. Positive log-odds ($\frac{1}{2} <$ probability ≤ 1) is associated with a hypothesis that is expected to hold, and negative log-odds ($0 \leq$ probability $< \frac{1}{2}$) with a hypothesis expected to be found to be false.

On the log-odds scale, the entire range from $-\infty$ to $+\infty$ is used: On the regular, 0 to 1, probability scale it becomes very difficult to conceptualize the difference between two values that are very small or, between two values that are very close to 1. But in many cases, this is where the action is. It then becomes useful to operate on a scale that, for example, causes us to conceive of the difference between 0.99 and 0.999 as equivalent to the difference between 0.999 and 0.9999. Of course, this comes at a price: the loss of the direct correspondence which exists with the regular probability scale, between belief values and ratios of occurrence.

Weight of evidence measures likelihood of what has been observed: The following formulation of weight of evidence, algebraically equivalent to that given in eq. 3.1, is compatible with our intuitions, as well as computationally useful.

$$woe(h:e) = \log \frac{p(e \mid h)}{p(e \mid \bar{h})} \tag{3.10}$$

From this expression, we see that the weight of evidence can be viewed as how much more likely we would be to see the evidence given that the hypothesis were true, relative to the likelihood of observing the same evidence were it to be false; the ratio in this case being measured on a log scale. The relation given in eq. 3.10, allowing for an arbitrary alternative hypothesis, h', and conditioned on some event c, would, in full generality, be:

$$woe(h/h' : e \mid c) = \log \frac{p(e \mid h, c)}{p(e \mid h', c)} \tag{3.11}$$

Weight of evidence provides a useful notion of independence: In probability theory, independence is defined so that two events, a and b, are considered *independent* if the probability of one is unaffected by knowledge of the other:

$$p(a \mid b) = p(a) \tag{3.12}$$

This definition is useful in large part because it corresponds to our intuitive notion of causal independence. We believe that the probability of a die showing 6 is independent of another die coming up 6. We believe this because we do not believe that one die has any causal influence on the other; nor do we believe in an external causal influence affecting both of the dice.

This notion of independence is less applicable in an environment in which we must assess the probability of a hypothesis in light of multiple sources of evidence. In this case, the weight of one piece of evidence may be independent of another, even though the two pieces of evidence are not probabilistically independent in the sense of eq. 3.12. For example, it may be the case that the weight of evidence in favor of death from heart disease provided by learning that a subject does not engage in regular exercise may be independent of whether or not the subject is overweight,

$$woe(\text{death} : \text{no exercise} \mid \text{overweight}) = woe(\text{death} : \text{no exercise})$$

This could be the case even though the probability that the subject does not exercise increases upon learning that the subject is overweight.

$$p(\text{no exercise} \mid \text{overweight}) > p(\text{no exercise})$$

Independence of weight of evidence is a different form of independence. This type of independence will hold when: the factor by which our belief that the subject will die of heart disease increases, upon learning that the subject does not exercise, is the same if we know the subject is overweight as it would be if we had no knowledge of the subject's weight.

Independence of evidence in the sense of:

$$woe(h : e_2 \mid e_1) = woe(h : e_2) \tag{3.13}$$

will play an important role in the probabilistic retrieval model developed in this thesis. As we will see, independence of weight of evidence, follows from the Maximum Entropy Principle. Based on this, evidence in favor of relevance given by the *tf-idf* value for a query term will be considered independent of the *tf-idf* values associated with other terms.

3.1.4 Weight of Evidence and Information Retrieval

There is nothing new about using either log-odds or weight of evidence in information retrieval. The Robertson/Sparck Jones term weight discussed in Section 2.1 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. 2.2, 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.

Finally, as a lead-in to what follows, we mention that Tukey, [107], recommends the use of what he calls *folded logs* or *flogs* as a natural and useful transformation of counted fractions for the purposes of exploratory data analysis. The folded log is defined to be:

$$\text{flog} = \frac{1}{2}\log_e f - \frac{1}{2}\log_e(1-f) \tag{3.14}$$

where f is the fraction of interest. With the fractions viewed as probabilities, *flog* is proportional to log-odds.

3.2 Exploratory Data Analysis

Hartwig and Dearing define *Exploratory Data Analysis* (EDA) as "a state of mind, a way of thinking about data analysis – and also a way of doing it" [57, p. 9]. They advance adherence to two principles. First, that one should be skeptical of data sum-

maries 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 article on EDA in the International Encyclopedia of Statistics says that it is the "manipulation, summarization, and display of data to make them more comprehensible to human minds, thus uncovering underlying structure in the data and detecting important departures from that structure" [3]. It goes on to point out that "these goals have always been central to statistics and indeed all scientific inquiry", but that there has been a renaissance in the latter part of this century. This is due in no small part to the accessibility of powerful electronic computers for: the accumulation of voluminous quantities of data, high-speed calculation and efficient graphical display.

EDA embodies a set of useful methods and strategies, fomented primarily by John W. Tukey [107]. Four distinguishing aspects of this practice, each of which plays an important role in the probability modeling discussed in this dissertation, are:

graphical displays: The emphasis in exploratory data analysis is on making the most of graphical displays of the data, a historical review of which is given in [6]. 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.

smoothing: Smoothing and non-parametric regression techniques are used with the objective of identifying the component of the data considered to be the *signal* from that which, for the purposes of the analysis at hand, are to be treated as *noise*.

study of residuals: A residual is the difference between the observed value of a response variable and the value predicted by a given model. By studying graphs of residuals against potential predictor variables, possibilities for extending the model can be explored.

transformation of variables: Supported by the production of graphical displays, features of the data can be transformed in a variety of ways in order to make more evident underlying regularities in the data.

What is most important is that EDA invites the researcher in information retrieval, fortunate at this point to have available a significant body of useful data at her disposal, to approach the task with the altered "state of mind" of which Hartwig and Dearing speak. It is the goal of this dissertation to take advantage of this methodology with the hope of letting the data itself guide development of an alternate approach to the development of IR systems and IR theory.

CHAPTER 4

ANALYSIS OF THE RELATIONSHIP BETWEEN DOCUMENT FREQUENCY AND THE WEIGHT OF EVIDENCE OF TERM OCCURRENCE

> *...while man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to.*
>
> Arthur Conan Doyle in *Sherlock Holmes: The Sign of Four*, Chapter 10.

This chapter reports on a study of the relationship between the weight of evidence in favor of relevance provided by the occurrence of a query term in a document and the frequency with which the term is found to occur in the collection as a whole. From the empirical study of retrieval data, statistical regularities are observed. Based on the patterns uncovered, a theory of term weighting is developed. The theory proposes *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 4.4, experiments are presented that corroborate this prediction.

4.1 Data Preparation

The study presented here involved data from queries 051-100 from the first Text REtrieval Conference (TREC) and the Associated Press (AP) documents from TREC volume 1 [51]. 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, a set of quality query terms was desired that would keep to a minimum the *noise* in the data to be analyzed. For this reason the concepts field was used.

Initially, the plan was for all query terms to be plotted. Two problems immediately presented themselves. First, rare terms are likely to have zero counts and this is problematic. 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 (relevant documents in which term occurs, relevant documents in which term does not occur, non-relevant documents in which term occurs, non-relevant documents in which term does not occur). This is a common approach, taken for instance in [89], where, for the purpose of estimating w_{rsj}, 0.5 is added to each count.

For the purposes of data analysis, however, there is a problem with this approximation. The choice of constant is to a large degree arbitrary. For many of the plots of interest, the shape of the plot at the low frequency end will vary considerably with the value chosen for the constant. 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 4.4 where $p(occ \mid rel)$ is plotted against $\log O(occ)$ for all terms for

which it has a finite value. That this variable increases with increasing *df* is somewhat evident, but subtler details of the relationship are obscured.

In order to confront both of these problems, data points were *binned*. Query terms were sorted in order of document frequency. Then for some bin size, k, sequences of k query terms[1] were grouped together in bins. Each bin was then converted into a single *pseudo-term* by averaging all counts (number of relevant documents in which term occurs, number of relevant documents in which term does not occur, number of non-relevant documents in which term occurs, number of non-relevant documents in which term does not occur). 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 4.4 is shown at the right of the same figure. Although we will focus on the binned plots, 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.

Although all analyses reported in this chapter were done using the binning technique described in the previous paragraphs, some experimentation was also conducted using an adaption of kernel regression methods [27, 59, 99]. A direct application of kernel regression does not work because it would involve calculating weighted averages over sets of estimates that include infinite values. Hence, averaging is done instead over the raw counts, as is done with the binning technique. As with binning, terms are ordered by document frequency. Then, each term is replaced by a *smoothed* version of the term.

All counts for the term are replaced by the weighted average of counts for all terms. Weights for the averaging are determined by some *kernel function*. Often,

[1] More precisely, k or $k + 1$ query terms, so that no bin was much smaller than the rest.

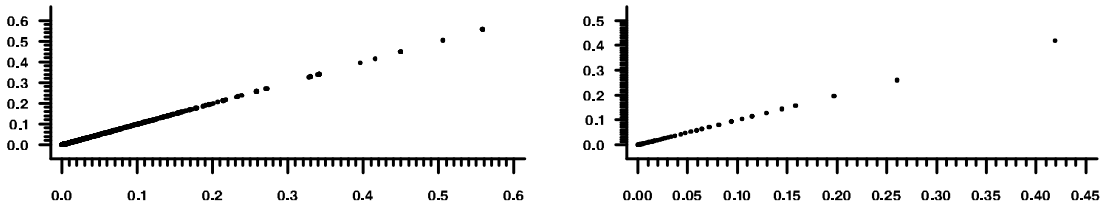

Figure 4.1. $p(occ \mid \overline{rel})$ as function of $p(occ)$

a Gaussian kernel function is used. The Gaussian is centered over the point being estimated, x_0. The weight assigned to each of the points, x', included in the average is the value of the Gaussian of the difference between the *idf* value at x' and the *idf* value at x_0. Once the smoothed version has been produced for all terms, the calculations of probabilities and weights can be done as above. How smooth a curve is produced can be adjusted by modifying the variance of the Gaussian function.

This approach to smoothing presents some advantages over the binning technique. In particular, discontinuities in the resulting plot due to the discrete nature of binning are smoothed over. However, the author did not become familiar with non-parametric approaches until after the phase of research reported in this chapter was complete.

4.2 Plotting the Data

Taking a lead from the Croft and Harper formulation of eq. 2.3, the data analysis begins by focusing on the components, $p(occ \mid rel)$ and $p(occ \mid \overline{rel})$, and how these components correlate with document frequency. Because the goal is to compare various document sets of differing sizes, we prefer not to plot the data in terms of absolute document frequencies. Instead, we plot against $p(occ) = \frac{df}{N}$, the probability that the term will occur in a document chosen randomly from the collection.

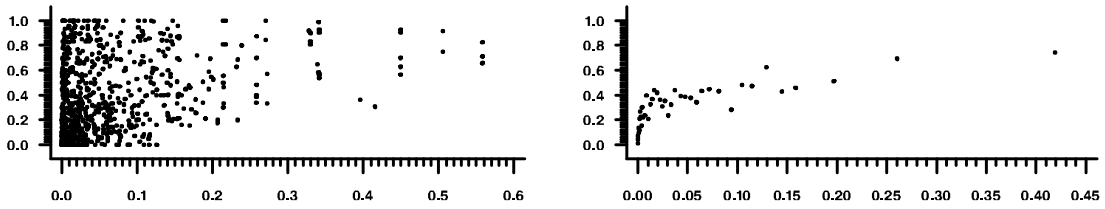

Figure 4.2. $p(occ \mid rel)$ as function of $p(occ)$

4.2.1 Occurrence in Non-relevant Documents

With respect to $p(occ \mid \overline{rel})$, we see in Figure 4.1 that it is well approximated by $p(occ)$. We will return to analyze the variable, $p(occ \mid \overline{rel})$, in more detail.

4.2.2 Occurrence in Relevant Documents

More interesting is Figure 4.2, which shows a plot of $p(occ \mid rel)$ as a function of $p(occ)$. We see from this scatter plot that as the document frequency (equivalently, probability of term occurrence) gets small (left end of graph), the probability of the term occurring in a relevant document tends to get small (lower on the graph) as well.

This plot of $p(occ \mid rel)$ vs. $p(occ)$ gives us reason to question the advisability of the assumption of equal probability of term occurrence in the relevant documents, used by Croft and Harper in [23] as the basis of eq. 2.4. This also puts into question the assumptions made by Robertson and Walker in [90]. Both the assumption that $p(occ \mid rel)$ increases from a non-zero starting point and that the increase is linear with increasing $p(occ)$ contradict the evidence provided by Figure 4.2. We return to discuss these points further in Section 4.5.

A glance at this graph suggests that a re-expression of variables may be indicated. The histogram shown at the left in Figure 4.3 confirms that, as both intuition and Figure 4.2 suggest, the distribution of document frequencies is highly skewed. With

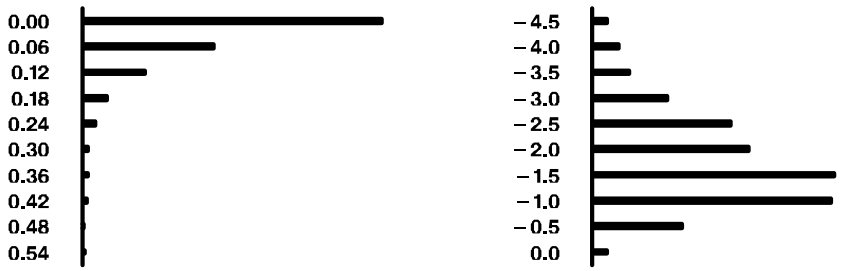

Figure 4.3. Histograms for $p(occ)$ and $\log O(occ)$

this type of skew, a logarithmic[2] transformation is often found to be beneficial [107]. Here, we go one step further and re-express the variable as:

$$\log O(occ) = \log \frac{p(occ)}{1 - p(occ)} \tag{4.1}$$

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, as discussed in Chapter 3. The histogram at the right in Figure 4.3 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, $p(occ \mid rel)$, is re-plotted as a function of $\log O(occ)$ in Figure 4.4. The plot against the log-odds shows that the decrease in $p(occ \mid rel)$ 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 4.2).

[2]As mentioned earlier, in order to aid intuitive comprehension, all logarithms in this document are logarithms to the base 10.

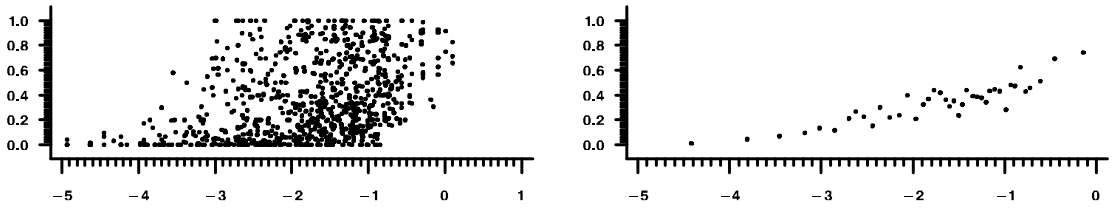

Figure 4.4. $p(occ \mid rel)$ as function of $\log O(occ)$

4.2.3 p(occ | rel) Relative to p(occ)

Despite the transformation of the independent variable, looking at $p(occ \mid rel)$ 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.

Take for example, two terms: t_1 with probability of occurrence, $p(t_1) = .2$, and t_2 with probability of occurrence, $p(t_2) = .0001$. We would expect that $p(t_1 \mid rel)$ is at least .2. In contrast, we could hardly expect a term which only appears in one of every ten thousand documents in the collection to appear in as many as two out of ten of the relevant documents. In fact, if the probability of relevance for the query is, say, one in a thousand, simple algebra shows that it will not be possible for $p(t_2 \mid rel)$ to be greater than 0.1:

$$p(occ \mid rel) = \frac{p(occ \wedge rel)}{p(rel)} \leq \frac{p(occ)}{p(rel)} = \frac{.0001}{.001} = .1$$

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 4.5 shows a plot of the ratio

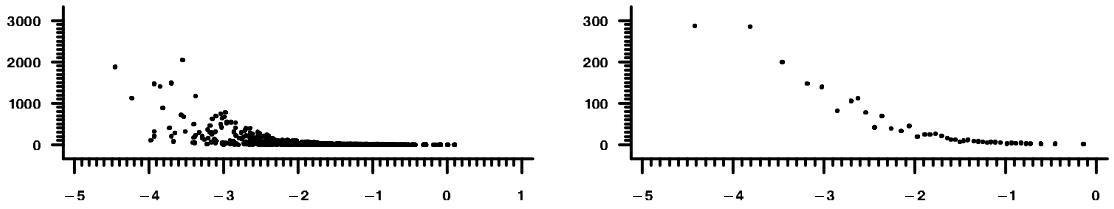

Figure 4.5. $\frac{p(occ \mid rel)}{p(occ)}$ as function of $\log O(occ)$

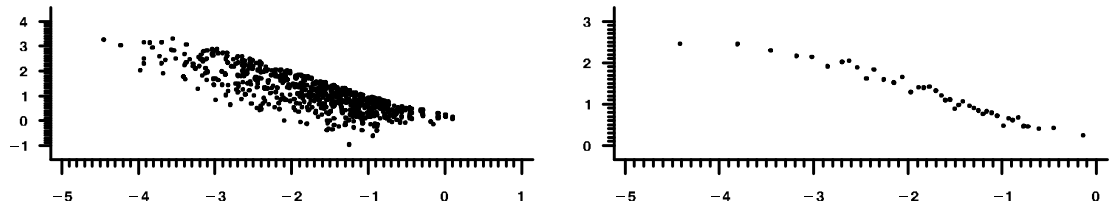

Figure 4.6. $\log \frac{p(occ \mid rel)}{p(occ)}$ as function of $\log O(occ)$

$$\frac{p(occ \mid rel)}{p(occ)}$$

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 $\frac{p(occ \mid rel)}{p(occ)}$ to be investigated.

4.2.4 Log of the Ratio of p(occ | rel) to p(occ)

In Figure 4.6 the log of the ratio $\frac{p(occ \mid rel)}{p(occ)}$ is plotted against the logarithm of the odds of occurrence in the collection. In the plot, we observe:

- a roughly linear overall increase in $\log \frac{p(occ \mid rel)}{p(occ)}$ 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:

$$\frac{p(occ \mid rel)}{p(occ)} = \frac{p(occ \wedge rel)}{p(rel)p(occ)} = \frac{p(rel \mid occ)}{p(rel)} \leq \frac{1}{p(rel)} \tag{4.2}$$

If we assume that, for a given query, the probability of relevance across the entire collection is approximately one in a thousand[3], then $\log \frac{p(occ \mid rel)}{p(occ)}$ 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 $\log \frac{p(occ \mid rel)}{p(occ)}$ 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, $\frac{p(occ \mid rel)}{p(occ)}$, is greater than 1, and its logarithm greater than 0. Since we conclude that $\log \frac{p(occ \mid rel)}{p(occ)}$ 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 4.6 (and

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

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.

Finally, we note that the quantity:

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

has connections to information theory. Often referred to as *mutual information*, it has frequently been used as a measure of variable dependence in information retrieval. It has been used in attempts to include co-occurrence data in models with weaker independence assumptions [111]; for the purposes of corpus-specific stemming [24]; and for term selection in query expansion based on relevance feedback [50]. It is also often used as a measure of variable dependence in computational linguistics [15].

In a very important sense, it can be taken as a measure of the information about one event provided by the occurrence of another [31]. In our context, it can be taken as a measure of 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.

4.3 Mutual Information and *idf*

In this section, we show how the general relationship observed between

$$\mathrm{MI}(occ, rel) = \log \frac{p(occ \mid rel)}{p(occ)}$$

and *df* can be used to explain why inverse document frequency should be expected to produce good retrieval performance when used for term weighting.

4.3.1 Δwoe

Our interest 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 (3.1), together with that for mutual information, Δwoe can be expressed as:

$$\begin{aligned}
\Delta woe &= woe(rel : occ) - woe(rel : \overline{occ}) \\
&= \log \frac{p(occ \mid rel)}{p(\overline{occ} \mid rel)} - \log \frac{p(occ \mid \overline{rel})}{p(\overline{occ} \mid \overline{rel})} \\
&= \log p(occ \mid rel) - \log p(\overline{occ} \mid rel) - \log p(occ \mid \overline{rel}) + \log p(\overline{occ} \mid \overline{rel}) \\
&= \mathrm{MI}(occ, rel) + \log p(occ) \\
&\qquad - \log p(\overline{occ} \mid rel) - \log p(occ \mid \overline{rel}) + \log p(\overline{occ} \mid \overline{rel})
\end{aligned} \tag{4.3}$$

4.3.2 Δwoe ≈ MI(occ,rel)

We will now argue that:

1) $\log p(occ \mid \overline{rel}) \approx \log p(occ)$

2) $-\log p(\overline{occ} \mid rel)$ is not too big;

3) $\log p(\overline{occ} \mid \overline{rel}) \approx 0$

This done, we will be able to reduce eq. 4.3 to the following approximation for Δwoe:

$$\Delta woe \approx \mathrm{MI}(occ, rel)$$

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 every effort has been made 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 4.6, and similar plots for other collections that have been 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) $\log \mathbf{p}(\mathbf{occ} \mid \overline{\mathbf{rel}}) \approx \log \mathbf{p}(\mathbf{occ})$: Throughout the range of document frequencies with which we are concerned, $\log p(occ \mid \overline{rel})$ cannot be very different from $\log p(occ)$.

$$p(occ \mid \overline{rel}) = \frac{p(occ \wedge \overline{rel})}{p(\overline{rel})} = \frac{p(occ) - p(occ \wedge rel)}{1 - p(rel)} \tag{4.4}$$

and, since $p(rel) \geq p(occ \wedge rel) \geq 0$,

$$\frac{p(occ) - p(rel)}{1 - p(rel)} \leq p(occ \mid \overline{rel}) \leq \frac{p(occ)}{1 - p(rel)} \tag{4.5}$$

We assume $p(rel)$ is small. If the probability of occurrence is fairly large relative to the probability of relevance, the range given in eq. 4.5 will be a small interval about $p(occ)$, the size of which will be small relative to $p(occ)$. Hence, the order of magnitude of $p(occ \mid \overline{rel})$ will be the same as that of $p(occ)$ and $\log p(occ \mid \overline{rel})$ will be close to $\log p(occ)$.

For example, if $p(occ) = .01$ with $p(rel) = .001$, then,

$$\frac{.01 - .001}{1 - .001} = .009009 \leq p(occ \mid \overline{rel}) \leq \frac{.01}{1 - .001} = .010010 \tag{4.6}$$

giving $-2.0453 \leq \log p(occ \mid \overline{rel}) \leq -1.9996$, whereas $\log p(occ) = -2.00$.

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

$$\begin{aligned} p(occ \mid \overline{rel}) = \frac{p(occ \wedge \overline{rel})}{p(\overline{rel})} &\approx p(occ \wedge \overline{rel}) \\ &= p(occ) - p(occ \wedge rel) \\ &= p(occ) - p(occ \mid rel) \cdot p(rel) \\ &= p(occ) \cdot \left(1 - \frac{p(occ \mid rel)}{p(occ)} \cdot p(rel)\right) \end{aligned} \tag{4.7}$$

For a probability of occurrence greater than 1 in 10,000, the data indicate (Figure 4.5) that $\frac{p(occ \mid rel)}{p(occ)}$ can be expected to be no more than 300. Given typical values for $p(rel)$, $(1 - \frac{p(occ \mid rel)}{p(occ)} \cdot p(rel))$ can still be expected to be less than one half of the value of $p(occ)$, and $\log_{10} p(occ \mid \overline{rel})$ can be expected to be not too different from $\log_{10} p(occ)$.

Since there is reason to believe that the growth of $\frac{p(occ \mid rel)}{p(occ)}$ will continue to increase as document frequency decreases, concluding that $p(occ \mid \overline{rel}) \approx p(occ)$ 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 (eq. 2.4) in [23] also depends on $p(occ \mid \overline{rel})$ 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 $\frac{p(occ \mid rel)}{p(occ)}$ relative to $p(rel)$. In theory, at least, $p(occ \mid \overline{rel})$ could be an arbitrarily small fraction of $p(occ)$. Equivalently, $\frac{p(occ \mid rel)}{p(occ)} \cdot p(rel) = \frac{p(occ \wedge rel)}{p(occ)}$ could be arbitrarily close to 1, and so, $p(occ \mid \overline{rel})$ in eq. 4.7 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 $\frac{p(occ \mid rel)}{p(occ)}$ supports a conclusion,

$$p(occ \mid \overline{rel}) \approx p(occ)$$

which cannot be rigorously maintained in its absence.

2) $-\log p(\overline{occ} \mid rel)$ is not too big : Though it may not be negligible, the quantity, $-\log p(\overline{occ} \mid rel)$ cannot be too big. The conditional probability, $p(\overline{occ} \mid rel)$, is simply $1 - p(occ \mid rel)$, and we see in Figure 4.2 that $p(occ \mid rel)$ is below .9, even for the most frequent pseudo-term. It follows that $p(\overline{occ} \mid rel) > .1$, and hence $-\log p(\overline{occ} \mid rel) < 1$.

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 $-\log p(\overline{occ} \mid rel)$ is achieved by only a small percentage of query terms – those which appear in more than 25% of all documents. Also, $\log p(\overline{occ} \mid rel)$ 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 $\log p(\overline{occ} \mid rel)$ should not overwhelm the overall effect of the more important component, $\mathrm{MI}(occ, rel)$, of Δwoe given in eq. 4.3.

3) $\log p(\overline{occ} \mid \overline{rel}) \approx 0$: 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, $p(\overline{occ} \mid \overline{rel}) > p(\overline{occ})$. In this study, $p(\overline{occ})$ is found to be

greater than .7 for all pseudo-terms. Equivalently, $0 > \log p(\overline{occ} \mid \overline{rel}) > -0.15$. This component too, has a minimal effect on Δwoe.

4.3.3 *idf* approximates Δwoe

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

$$\Delta woe \approx \mathrm{MI}(occ, rel) \approx k_2 - k_1 \log O(occ) \tag{4.8}$$

But, k_2 can be ignored. By casual inspection of Figure 4.6, we see that any reasonable linear approximation of the plot of $\mathrm{MI}(occ, rel) = \log \frac{p(occ \mid rel)}{p(occ)}$ as a function of $\log O(occ)$ will have an intercept value relatively close to 0. We are now left with:

$$\Delta woe \approx -k_1 \log O(occ) = k_1(-\log O(occ)) \tag{4.9}$$

Once the constant k_2 has been eliminated, the remaining constant, k_1, becomes irrelevant for the purposes of ranking. It will affect only the scale of the scores obtained, having no affect on the ranking order itself. And so we conclude that the *idf* formulation,

$$idf = -\log O(occ) = \log \frac{N - n_i}{n_i} \tag{4.10}$$

should produce good retrieval performance.

4.4 Improving on IDF

We have shown that by accepting some empirically motivated assumptions concerning query terms, the quantity Δwoe can be approximated by $\mathrm{MI}(occ, rel)$. 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 4.1, 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 can 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 4.6 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 [11] on each of the ad-hoc tasks for TREC 1 through TREC 6 [53]. Queries were formed by taking all words from both the title and description. After the union of the words in the title and description fields was produced, the following processing steps were applied:

stopword removal: *stopwords* taken from a fixed list of non-content bearing English words (articles, prepositions, etc.) were removed.

elimination of duplicates: any term appearing more than once in the title appears just once in the query.

stemming: all words were converted to a canonical form based on the *k-stem* algorithm [72].

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

[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.

$-\log O(occ) = 1.0$ to $-\log O(occ) = 3.0$; and maintained at a constant value for all terms for which $-\log O(occ)$ exceeded 3.0.

	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 4.1. 3-piece Piecewise-linear vs. Linear Versions of *idf*

The results of these tests are summarized in Table 4.1. The test version outperforms the baseline system in terms of average precision, on all six query sets, by 20% or more 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. The test results showed statistically significant improvement at the 5% level on all test sets according to both statistical measures. Improvement at the .1% level was observed in three of the six runs according to the sign test and five of the six according to the wilcoxon. 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.

4.5 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 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 $p(occ \mid rel)$ 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 these data supplies us with sufficient reason for assigning unequal probabilities for $p(occ \mid rel)$ 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. 2.3, 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 $p(occ \mid rel)$ 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 $p(occ \mid rel)$ 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 $p(occ \mid rel)$ for the lowest frequency terms must be very very small. Unfortunately they find themselves restricted to values above 0.5. Figure 4.2 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 $p(occ \mid rel)$ predicted by the model comes closest to approximating the plot shown in Figure 4.2. 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, $p(occ \mid rel)$, 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? 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 can 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.

4.6 Summary

In this chapter we have analyzed the discriminatory power of a term as a function of its frequency of occurrence in the collection. We have seen how an approach based on exploratory data analysis can lead to the use of $-\log \frac{df}{N}$ as a feature for term weighting. Through the use of EDA we recapitulate advances originally made in information retrieval due to intuition; in this case, the intuition of Sparck Jones in the early seventies [103]. The EDA approach, however, has led to a more precise formulation of the way document frequency can be utilized for term weighting.

At this stage we have only studied the use of document frequency in isolation. Two point should be noted. First, it has been implicitly accepted that term weights based on document frequency should be added across query terms appearing in a document in order to use weights of evidence for retrieval ranking. Second, modern IR systems make more sophisticated use of available sources of evidence.

These two points are addressed in the following two chapters. First, in Chapter 5, we present a theoretical foundation for probabilistic modeling based on the Maximum Entropy Principle. We derive a formulation for combining different sources of evidence. Based on this foundation, we proceed, in Chapter 6, to expand the techniques introduced here to sources of evidence traditionally used in information retrieval that have not yet been considered.

CHAPTER 5

THE MAXIMUM ENTROPY PRINCIPLE AS A FOUNDATION FOR THE DERIVATION OF PROBABILISTIC RETRIEVAL MODELS

> *Tout le monde y croit (la lois des erreurs) par ce que les mathématciens s'imaginent que c'est un fait d'observation, et les observateurs que c'est un théorème de mathèmatiques.*
>
> Henri Poincaré in the preface to *Thermodynamique*, quoted by Mark Kac in *Enigmas of Chance: An Autobiography* [67, p. 48]

In the experimentation reported in the previous chapter, it was implicitly assumed that the weights associated with the query terms should be added in order to produce the ranking status value for document retrieval. Additivity follows from independence assumptions, which have been made in one form or another in much of the work on probabilistic retrieval models. In particular, the Binary Independence and Combination Match models mentioned in the section on related work have been based on these assumptions. In [47], Greiff and Ponte have shown that the independence assumptions of both of these models, as well as the assumption of equal probability of occurrence in relevant documents in the Combination Match model, can be derived from what has come to be known as the Maximum Entropy Principle.

In this chapter, we begin with a summary of the principal arguments given in [47]. We then go on to extend this work by proving a theorem that gives general conditions from which the additivity of weights of evidence follows. This theorem, which generalizes arguments used in [47] with respect to both the Binary Independence and

Combination Match models will be used in the chapter following this one to support the addition of weights of evidence in the log-odds model developed there.

In order to rank documents in response to a query, a probabilistic system will calculate a probability of relevance for each document. This calculation will be based on some joint probability distribution over the relevance variable and variables corresponding to the evidence used by the system. The system, however, will not have full knowledge of such a distribution. In the Binary Independence and Combination Match models, a probability distribution is chosen by making strong assumptions concerning the distribution, which, together with parameters estimated from the data, allows the desired probability of relevance to be calculated. We will show how these formal models can be derived from the Maximum Entropy Principle, which counsels us to select the probability distribution with maximum entropy of all those that conform to an accepted set of constraints.

We adopt a probabilistic attitude with respect to information retrieval in this chapter, where *probability* of relevance shall be understood as the system's judgment that a document will be relevant based on all information it has available to it. These probabilities will have to be determined in the absence of total knowledge concerning all aspects of the distribution. In order to rank a document according to the probability that it will be judged relevant to the query, an IR system must adopt a probability distribution of relevance conditioned on the evidence it considers. Available knowledge will constrain this distribution, but will not leave it fully determined. The Maximum Entropy Principle provides a reasonable methodology for fully determining the otherwise underconstrained distribution.

In the next section, we introduce the Maximum Entropy Principle in the context of a Bayesian view of probabilities. We go on to give a brief review of the Binary Independence Model (BIM) developed by Robertson and Sparck Jones. We also review

the Croft and Harper adaptation of the basic BIM idea to applications for which no relevance judgments are presumed to be available.

In Section 5.3, we show how the essence of the Binary Independence Model can be derived from the Maximum Entropy Principle. With the development of the model established, we discuss the assumptions of the Binary Independence Model, in the light of the maximum entropy approach. In particular, we show that linked dependence, which is assumed in the Binary Independence Model is, in a sense, a consequence of the maximum entropy (MAXENT) model in that it is a characteristic of the resulting probability distribution. In section 5.4 we go on to show how the work of Croft and Harper can also be reproduced from the maximum entropy standpoint. Again the approach taken by the original authors is compared to that adopted with MAXENT.

In Section 5.5, we discuss the two models that have been developed with the MAXENT approach. More specifically, we discuss the differences between making assumptions concerning the probabilities of events and constraining the probability distribution. We will review how constraints have been chosen to reflect prior information in the models and, generally, what kinds of prior information can be incorporated in a MAXENT distribution.

5.1 Bayesian Reasoning and Maximum Entropy

This section begins with a review of Bayesian reasoning. Based on a Bayesian outlook, the Maximum Entropy Principle (MEP) is defined. In Section 5.1.3, we give an example of how thinking in terms of the MEP can be used to solve a problem analogous to that facing the researcher in information retrieval. The section concludes with a review of previous work in the area of information retrieval pertaining to the application of the Maximum Entropy Principle, and the motivation behind adopting it in this thesis.

5.1.1 Bayesian Reasoning

A clear distinction is made in Statistics with regard to those who consider themselves frequentists and others who tend to be known as Bayesians. Frequentists view a probability as a real characteristic of a physically reproducible experimental setup. A clear example of this would be the repeated throwing of a coin or pair of dice. Another would be the random sampling of a physically existing population such as that which is done for the purposes of medical testing or political polling.

Bayesians can be distinguished in two important ways. First is a far greater tendency to call on Bayes law:

$$p(H|E,K) = \frac{p(E|H,K) \cdot p(H|K)}{p(E|K)} \tag{5.1}$$

when reasoning probabilistically. The second is that Bayesians have a wider view of what a probability is. For a Bayesian, a probability is interpreted as the plausibility of a proposition. While these propositions can refer to repeatable events, such as coin tosses, they may also refer to propositions that are not easily or naturally given a frequentist interpretation. Propositions referring, for example, to whether Albert Gore will be elected president of the United States in the year 2000, or whether Lizzy Borden was actually guilty of what she was accused are anathema to the frequentist, but considered grist for the probabilistic mill by the Bayesian.

These two facets of the Bayesian are not unrelated. Often, H is a statistical hypothesis and E is data that has been collected. K is included to emphasize that, for the Bayesian, all probabilities are conditioned on the background knowledge possessed by the person (or machine) making the probability assessment, as well as other information such as the data from an experiment. In such cases, $p(E|H,K)$ is the likelihood of seeing the evidence we have observed given that the hypothesis is true, and $p(H|K)$ is the *prior* probability that H is true before any data have been observed. $p(E|K)$ is the probability assigned to seeing the evidence without any knowledge

of which of the possible hypotheses may be true. In general, it may be calculated by summing the product of the likelihoods and prior probabilities over all possible hypotheses:

$$\begin{aligned} p(E = e \mid K) &= \sum_{i=1}^{n} p(E = e \wedge H = h_i \mid K) \\ &= \sum_{i=1}^{n} p(E = e \mid H = h_i, K) \cdot p(H = h_i \mid K) \end{aligned} \tag{5.2}$$

where e is the observed evidence, each h_i is one of the possible hypotheses, and the summation is over all possible hypotheses. Equivalently, $p(E|K)$ can be viewed as a normalization constant chosen to make the sum of probabilities over all possible hypotheses conditioned on the evidence, e, sum to 1.

For a frequentist, this type of reasoning is not considered valid unless the probability of a hypothesis can be given a frequency interpretation. Often it is not possible, or at least it is very unnatural, to conceive of the hypothesis as a random event. For the Bayesian who views the probability of a hypothesis as a measure of its plausibility, this does not present a problem. We see then that the two aspects of the Bayesian outlook, the utilization of Bayes law and the interpretation of the meaning of a probability, are intimately intertwined. The reader is referred to [32, 49] for more in depth discussions of these issues.

5.1.2 The Maximum Entropy Principle

At the end of the 19th century, primarily as a result of the work of Maxwell, Boltzmann and Gibbs [63], the area of Statistical Mechanics was born. As a consequence, the entropy of a physical system became associated with a probability distribution of the phase space of possible atomic configurations.

In 1948, Claude Shannon published *The Mathematical Theory of Communication* and established the foundations of Information Theory. From three intuitively appealing desiderata, Shannon developed a formal expression for a measure of "how

much 'choice' is involved in the selection of an event or of how uncertain we are of the outcome" [97]. He showed that for a probability distribution, $\mathbf{p} = (p_1, \ldots, p_k)$, over k possible elementary events, the quantity:

$$H(\mathbf{p}) = -\sum_{i=1}^{k} p_k \log p_k \tag{5.3}$$

is, within a constant factor, the unique quantity in accord with his assumptions[1]. Since the form of the expression is recognized as the expression given for the physical property of entropy in formulations of Statistical Mechanics he calls the quantity *entropy* and adopts the symbol H, recalling Boltzmann's H-theorem.

In 1957, Edwin Jaynes "converted Shannon's measure to a powerful instrument for the generation of statistical hypotheses and ... applied it as a tool in statistical inference" [106]. In a pair of seminal articles, [60, 61], Jaynes demonstrates that by viewing it as a problem of statistical inference, Statistical Mechanics can be derived without depending on "additional assumptions not contained in the laws of mechanics" [60]. His method of inference is based on what has come to be known as the *Maximum Entropy Principle*. In his own words, this principle states that the maximum entropy estimate is: "the least biased estimate possible on the given information; i.e. it is maximally noncommittal with regard to missing information" [60, pg. 620]. This maximum entropy estimate is obtained by determining that probability distribution associated with a random variable, A, over a discrete space $(a_1, \ldots, a_n)$ which has the greatest entropy subject to constraints on the expectations of a given set of functions of the variable. That is, the distribution that maximizes eq. 5.3 subject to a set of constraints:

[1] This constant factor can be identified with the base chosen for the logarithm in the expression of entropy. Consistent with the rest of this document all uses of the *log* symbol in this chapter will refer to base 10.

$$
\begin{aligned}
E(g_1(A)) = \sum_{i=1}^{n} p_1 \cdot g_1(a_i) &= G_1 \\
&\vdots \\
E(g_m(A)) = \sum_{i=1}^{n} p_m \cdot g_m(a_i) &= G_m
\end{aligned}
$$

These constraints embody the knowledge that we wish to incorporate in *our* distribution of the probability over the possible elementary events.

5.1.3 The Brandeis Dice Problem

The example given here is an adaptation of the "Brandeis Dice Problem" originally presented as an illustration of the maximum entropy approach in [62].

Suppose that we are given a large number of dice and the task of ranking them. Once the dice are ordered, each will be thrown one time, and our goal is to get as large a number of 4's as we can. Suppose, furthermore, that experiments have been run on the dice. Each die has been thrown a large number of times, but the only knowledge we have of these experiments is the average value produced by each die. Following the Probability Ranking Principle [88], we decide to rank the dice by the probability of their producing 4's. How are we to arrive at this probability?

Of some things we feel sure. A die whose average is very close to either 1 or 6 should rank very low. We know that a die that produced an average close to 1 must have produced almost all 1's and hence could have produced only a few 4's at best. The frequency of 4's was low in the experimental trials, and common sense dictates assigning a very low probability to its producing 4 the next time it is thrown. Similarly for an average close to 6. Somehow common sense also tells us that dice that produce sample means above 3.5 should be ranked higher than those that produced sample means below 3.5. A die that produced an average greater than 3.5 has exhibited a tendency toward the higher numbers, whereas a die that produced an average below 3.5 has exhibited a tendency toward the lower numbers. It is reasonable, then, to

assign a higher probability of producing a 4 to a die that has displayed an affinity for higher numbers. But, how are we to compare, for example, a die with an average of 3.7 against a die with an average of 4.2?

There are a number of formulations that might be employed here which are reasonable and are analogous, in fact, to approaches taken in a variety of real applications. For example, we might choose the probability distribution with maximum variance. This would result in all of the probability mass placed on the 1 and the 6 in such a way that mean, μ, was respected. That is,

$$\mathbf{p} = < p_1 = \tfrac{6-\mu}{5},\ p_2 = 0,\ p_3 = 0,\ p_4 = 0,\ p_5 = 0,\ p_6 = \tfrac{\mu-1}{5} >$$

Another approach that has been used in statistics is to choose the distribution that minimizes the sum of the squares of the probabilities. Unfortunately, this can lead to negative values for some of the probabilities in some cases.

The maximum entropy solution to this problem is to assign to each die the probability distribution over the six possible numbers that has maximum entropy. From this distribution, we can determine the probability of each die coming up 4, and then rank the dice based on these probabilities. The MAXENT solution has the following attractive properties:

- The probability associated with each die accords with the data in that, under this distribution, the expectation of the number to appear on a given toss is equal to the experimental average.

- Of all distributions that conform with the data in this way, it is that which has the maximum uncertainty associated with it, in the sense of uncertainty that follows from the Shannon desiderata. The probability is "spread out" as much as possible in accord with the constraints that have been imposed. In this way, it may be said to include all the knowledge available and nothing more. In the words of Jaynes, it is the least biased distribution possible.

- The method is logically consistent. We are guaranteed that anomalies (negative probabilities, for example) will not occur if we follow the MAXENT procedure.

- The results accord with common sense. For example, the probability distribution for a die with an average close to 1 will be highly peaked around 1.

The probability distribution associated with a die whose average is 4.0 is given by Golan, Judge and Miller [40, Table 2.3.1, pg. 14] as

$$\mathbf{p} =< 0.103, 0.123, 0.146, 0.174, 0.207, 0.247 >$$

The expected value for this distribution is 4.0, and of those with expectation of 4.0, the probability is the most evenly distributed. An average of 5.0 corresponds to a distribution of

$$\mathbf{p} =< 0.021, 0.038, 0.072, 0.136, 0.255, 0.478 >$$

which is even more skewed toward higher numbers, as we would expect. An average of 3.5 corresponds to

$$\mathbf{p} =< 0.167, 0.167, 0.167, 0.167, 0.167, 0.167 >$$

which is as spread out as a distribution over six possibilities can be, and is the same probability distribution we associate with a die about which we have no information. It is interesting to note here that the MAXENT approach allows such a die (one for which the experimental average was missing for some reason) to be included in the ranking along with the rest. Values for the probability of throwing a 4, associated with dice with averages of 2.0, 3.0, 3.5, 4.0, and 5.0 are, respectively, 0.072, 0.146, 0.167, 0.174 and 0.136.

We have modified The Brandeis Dice Problem so that the example is suggestive of the problem faced in the design of information retrieval systems. It is now time to address directly the issue of how the MAXENT approach pertains to IR modeling.

5.1.4 The MAXENT Approach and Probabilistic IR Modeling

Since the publication of Jaynes' articles, the Maximum Entropy Principle has been applied to practical problems in diverse areas [29], including image reconstruction [48], spectral analysis [8], reliability engineering [105] and economics [40].

In two papers in the early '80s, Cooper and Huizinga [22] and Cooper [16], make a strong case for applying the maximum entropy approach to the problems of information retrieval. Cooper points out that, "A common criticism of most probabilistic approaches to information retrieval system design is that they involve the use of unrealistic simplifying assumptions concerning statistical independence" [16]. Cooper and Huizinga state that one might "forgive serious oversimplifications in particular cases if the assumptions were in some sense correct on the average, or if they constituted a best guess in some *cogent statistical sense*, but no convincing arguments have been advanced showing that the assumptions are supportable even in this weak sense[2]" [22, pg. 101].

In these papers, firm first steps are taken in the direction of applying maximum entropy to information retrieval. The maximum entropy approach is used to incorporate the idea of term precision weighting [95] in a probabilistic context. They show how probability-of-relevance computations based on MAXENT result in an expressive request language combining the capabilities of both Boolean and "weighted-request" retrieval systems.

In [68, 69], Kantor and Lee extend the analysis of the Maximum Entropy Principle in the context of information retrieval. In [73] they explore the use of maximum entropy to resolve user estimates of conditional relevance probabilities that may be inconsistent with available term occurrence data. Very recently, [70], they have conducted experiments to test the performance of the MEP as a method of document

[2]italics added

retrieval. While they outperform two simpler methods on small collections, they report discouraging results on large document sets and conclude that the MEP, in general, does not appear to present advantages over more "naive" methods.

In contrast to the work of Kantor and Lee in [70], our interest is not in the development of an alternative retrieval algorithm based on the MEP. Our intent is rather to consider the conceptual basis for traditional approaches to probabilistic retrieval. The goal in what follows will be to analyze classical probabilistic IR models in light of the Maximum Entropy Principle. The primary objectives are to: 1) show that traditional approaches to probabilistic retrieval modeling can be reproduced using the MAXENT methodology; and 2) compare and contrast the classical and MAXENT approaches. The reasons for undertaking this study is our belief that:

- The MAXENT approach is, in a sense, more basic than previous approaches. We believe that maximum entropy allows for the development of probabilistic models from conceptually simpler, more fundamental principles. We recognize that opinions will differ as to what is to be considered conceptually simpler and more fundamental. We shall try to avoid taking a dogmatic stand in what follows and stay to our goal of presenting an alternative view and the reasons we believe this view to be worthy of consideration.

- The MAXENT approach adopts a different philosophical attitude with respect to the role of probability theory, and the meaning of "probability". This difference we believe to be pertinent when the probability calculus is applied to the problem of information retrieval. We find this distinction to be more than an abstract issue of philosophical interpretation, but one with practical repercussions that can affect how the IR problem is viewed; the types of solutions researchers are predisposed to consider; the methodologies and tools brought to bear; the formulation of proposed solutions; and ultimately the design of retrieval systems.

- Maximum entropy offers a formal, mathematically consistent technique for the combination of evidence. The justification of this technique, felt to be compelling by some, less so by others, can be said, at the least, to be reasonable. In cases of sufficient simplicity, for which common sense suggests a solution, MAXENT is found to accord with common sense.

- Maximum entropy can be viewed as a methodology of research. The researcher, intent on modeling some aspect of nature stochastically, chooses an elementary event space as best she can based on her knowledge of the phenomenon under study. She further constrains the probability distribution over this space using whatever information she has available. She then may mathematically derive the form of the maximum entropy distribution. If this distribution is satisfactory, all is well, and she is done. The results, however, may not be acceptable. The derived distribution may not predict something known to be true; in Statistical Mechanics, for example. Or, an application utilizing the distribution, for image reconstruction perhaps, may produce results inferior to what we have reason to suspect is possible. If so, this is where, according to Jaynes, the maximum entropy approach can be most valuable. Jaynes recounts how classical statistical mechanical theory was unable to predict some thermodynamic properties such as heat capacities. This state of affairs forced the search for additional constraints. The nature of this constraint lay in the discreteness of possible energy states. Jaynes asserts as "historical fact that the first claims indicating the need for the quantum theory ... were uncovered by a seeming unsuccessful application of the principle of maximum entropy" [64, p. 1125].

  If the distribution is not living up to expectations, then something known about the problem has not been taken into account and MAXENT points a finger in the direction that needs to be explored. There may be a way of using this knowledge to further constrain the distribution. If this extra piece of knowledge can be

identified, a way of incorporating the knowledge in the form of one or more new constraints can be designed and the process may continue. If no more constraints can be found and the results are still not adequate, the researcher must begin to question the specification of the elementary event space over which the probability distribution is defined. After serious contemplation, the space may, in retrospect, be thought not to be the best. The researcher may want to modify the space so as to better conform to her prior knowledge with respect to the nature of her problem.

In this thesis, the following view of a probabilistic retrieval system is adopted. The ranking score of a document is *the system's* probability that the document in question will be found to be relevant to a given query. In arriving at this probability, the system brings to bear all general knowledge it has concerning the relevance of documents to queries. This is combined with knowledge of the characteristics of the particular document collection being searched and the specific query/document pair currently under scrutiny. In the case of the Binary Independence Model, knowledge gleaned from the user in the process of relevance feedback is used as well.

For a given query, the system will arrive at a joint probability distribution over the elementary event space $\Omega = \mathbf{X} \times R$, where $\mathbf{X}$ is a vector of document attributes and R=$\{0, 1\}$ corresponds to judgments of relevance. Knowledge built into the system in combination with knowledge of the statistical characteristics of the document collection are used to constrain the probability distributions that will be considered. Of the set of probability distributions satisfying these constraints, the unique distribution that maximizes the entropy will be chosen. The distribution can then be used to assign the system's probability of relevance.

5.2 Binary Independence and Combination Match Models

This section presents the mathematical foundations of both the Binary Independence and Combination Match models. In the following two sections, we will see how the same models can be derived from the Maximum Entropy Principle.

5.2.1 Binary Independence Model

The *Binary Independence Model* (BIM), developed by Robertson and Sparck Jones [89, 111], adopts a probabilistic approach to the development of a ranking formula. It is designed to be applicable in an environment in which the relevance of some of the documents will have been judged prior to the application of the BIM ranking formula.

In the Binary Independence Model, the focus is on the odds of relevance, conditioned on the occurrence pattern of the query terms that is observed in a given document:

$$O(rel|x_1,\ldots,x_s) \;=\; \frac{p(rel|x_1,\ldots,x_s)}{p(\overline{rel}|x_1,\ldots,x_s)}$$

where $(x_1,\ldots,x_s) \in \{0,1\}^s$ are the values of $(X_1,\ldots,X_s)$ corresponding to the occurrences of the s query terms in a given document. For the purposes of clarity of exposition, rel and $\overline{rel}$ shall be used interchangeably with 0 and 1, respectively, for the values of the relevance variable, R. The application of Bayes law in both the numerator and the denominator gives:

$$\begin{aligned} O(rel|x_1,\ldots,x_s) &= \frac{p(x_1,\ldots,x_s|rel)\;\cdot\;p(rel)/p(x_1,\ldots,x_s)}{p(x_1,\ldots,x_s|\overline{rel})\;\cdot\;p(\overline{rel})/p(x_1,\ldots,x_s)} \\ &= \frac{p(x_1,\ldots,x_s|rel)\;\cdot\;p(rel)}{p(x_1,\ldots,x_s|\overline{rel})\;\cdot\;p(\overline{rel})} \\ &= \frac{p(x_1,\ldots,x_s|rel)}{p(x_1,\ldots,x_s|\overline{rel})}\;\cdot\;O(rel) \end{aligned} \tag{5.4}$$

The key assumption in the Binary Independence Model is that query term occurrences are independent in both the relevant and non-relevant sets. Formally:

$$\forall(x_1,\ldots,x_s)\in\{0,1\}^s: \qquad p(x_1,\ldots,x_s|rel) = \prod_{i=1}^{s} p(x_i|rel) \tag{5.5}$$

$$p(x_1,\ldots,x_s|\overline{rel}) = \prod_{i=1}^{s} p(x_i|\overline{rel}) \tag{5.6}$$

From which, it immediately follows that:

$$\forall(x_1,\ldots,x_s)\in\{0,1\}^s: \frac{p(x_1,\ldots,x_s|rel)}{p(x_1,\ldots,x_s|\overline{rel})} = \prod_{i=1}^{s}\frac{p(x_i|rel)}{p(x_i|\overline{rel})} \tag{5.7}$$

William Cooper later emphasized that equation eq. 5.7 is all that really needs to be assumed [17]. This "linked dependence assumption" is weaker than the pair of conditional independence assumptions, eq. 5.5 and eq. 5.6, and is a fairer statement of the properties that need be assumed to hold, in order for the application of the Binary Independence Model to be valid.

Under the linked dependence assumption the expression given in eq. 5.7, may be substituted for the fraction in eq. 5.4, giving:

$$\begin{aligned} O(rel|x_1,\ldots,x_s) &= \prod_{i=1}^{s}\frac{p(x_i|rel)}{p(x_i|\overline{rel})} \cdot O(rel) \\ &= \prod_{x_i=1}\frac{p(X_i=1|rel)}{p(X_i=1|\overline{rel})} \cdot \prod_{x_i=0}\frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})} \cdot O(rel) \\ &= \prod_{x_i=1}\frac{p(X_i=1|rel)}{p(X_i=1|\overline{rel})}\frac{p(X_i=0|\overline{rel})}{p(X_i=0|rel)} \cdot \prod_{i=1}^{s}\frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})} \cdot O(rel) \end{aligned}$$

Taking the log of both sides yields:

$$\log O(rel|x_1,\ldots,x_s) = \sum_{x_i=1}\log\frac{p(X_i=1|rel)}{p(X_i=1|\overline{rel})}\frac{p(X_i=0|\overline{rel})}{p(X_i=0|rel)} \tag{5.8}$$

$$+ \sum_{i=1}^{s} \log \frac{p(X_i = 0|rel)}{p(X_i = 0|\overline{rel})} \; + \; \log O(rel)$$

The Binary Independence Model supposes that relevance feedback information is available and that the probabilities in eq. 5.8 can be estimated from the set of documents judged relevant and non-relevant:

$$\begin{aligned} p(X_i = 1|rel) &= \xi_i \\ p(X_i = 1|\overline{rel}) &= \bar{\xi}_i \end{aligned}$$

giving,

$$\log O(rel|x_1, \ldots, x_s) = \sum_{x_i=1} \log \frac{\xi_i(1 - \bar{\xi}_i)}{\bar{\xi}_i(1 - \xi_i)} \; + \; \sum_{i=1}^{s} \log \frac{1 - \xi_i}{1 - \bar{\xi}_i} \; + \; \log O(rel) \qquad (5.9)$$

The result is an additive formula for the calculation of the log-odds of relevance, conditioned on the occurrence pattern of the query terms. The increase in the log-odds in favor of a hypothesis, from $\log O(rel)$ to $\log O(rel|x_1, \ldots, x_s)$ in this case, can be understood as "weight-of-evidence" as defined in Chapter 3. The formula allows the weight of evidence in favor of relevance provided by the occurrence pattern of the query terms, relative to that provided by a document in which no query terms are present, to be calculated by adding:

$$\log \frac{\xi_i(1 - \bar{\xi}_i)}{\bar{\xi}_i(1 - \xi_i)} \qquad (5.10)$$

for each query term that appears in the document. From a practical standpoint, it is important that the calculation involves only terms that appear in the document.

5.2.2 The Combination Match Model without Relevance Information

In 1979, Croft and Harper adapt the work of Robertson and Sparck Jones to develop a probabilistic retrieval model that does not depend on the availability of

relevance information. In the place of relevance feedback data they use collection statistics to estimate the probability of a query term appearing in a non-relevant document. Croft and Harper rewrite the sum of the BIM term weights, eq. 5.10, as:

$$\sum_{x_i=1}^{s} \log \frac{p(X_i = 1|rel)(1 - p(X_i = 1|\overline{rel}))}{p(X_i = 1|\overline{rel})(1 - p(X_i = 1|rel))}$$

$$= \sum_{x_i=1}^{s} \log \frac{p(X_i = 1|rel)}{1 - p(X_i = 1|rel)} + \sum_{x_i=1}^{s} \log \frac{1 - p(X_i = 1|\overline{rel})}{p(X_i = 1|\overline{rel})} \qquad (5.11)$$

They estimate the value of $p(X_i = 1|\overline{rel})$ as $\frac{n_i}{N}$, where n_i is the number of documents in which term i appears and N is the total number of documents in the collection. They also assume that the probability of appearing in a relevant document is the same for all terms in the query, an assumption we will examine further later on. The first term of eq. 5.11 is then simply a constant, C, times the number of query terms that appear in the document. Viewing this constant as a weighting factor, they conclude that the best ranking function is a weighted combination:

$$C \cdot \sum_{x_i=1}^{s} 1 + \sum_{x_i=1}^{s} \log \frac{N - n_i}{n_i} \qquad (5.12)$$

of a simple coordination match and a match using *idf* weights, which they call the *Combination Match Model* (CMM). They determine the value for C empirically, based on the quality of the resulting retrieval performance. This formula suggests a probabilistic justification of the use of inverse document frequency for the weighting of terms, which was originally proposed by Karen Sparck Jones [103].

5.3 Basic BIM-MAXENT Model

In this section, we derive a retrieval model based on the Maximum Entropy Principle. The model, which we shall refer to as BIM-MAXENT, will be constrained in such a way as to be consistent with the assumptions made in the Binary Independence

Model of Robertson and Sparck Jones. Our goal is to reproduce the ranking formula. Subsequently, we will analyze the constraints placed on the probability distribution in our maximum entropy model and compare them with the assumptions on which the Binary Independence Model is based.

Our goal is to maximize the entropy of the probability distribution:

$$H(p) = -\sum_{w \in \Omega} p(\omega) \log p(\omega) \tag{5.13}$$

where each ω is an elementary element of the event space $\Omega = \mathbf{X} \times R$. Each elementary event corresponds to the observation of a document with respect to a given query. Associated with each observation are the random variables, $X_1, \ldots, X_s$, & R, where s is the number of terms in the query. Each of these variables is binary, with $X_i = 1$ corresponding to the occurrence of term i in the document, and $R = 1$ corresponding to the document being relevant to the query. Hence, the sum in eq. 5.13 is taken over all possible (binary) assignments $(x_1, \ldots, x_s, r)$ to $(X_1, \ldots, X_s, R)$.

5.3.1 Constraints

In the maximum entropy model, the probability distribution over these elementary events will be constrained in three different ways:

- For each query term, the probability of its occurring in a document known not to be relevant to the query will be constrained. These probabilities may be constrained independently.

- For each query term, the probability of its occurring in a document known to be relevant to the query will be constrained. As with the probabilities conditioned on non-relevance, the probability associated with each query term may be constrained independently of the rest.

- The prior probability of relevance (i.e., the probability that an arbitrary document is relevant before any of the term occurrence variables is observed) will be constrained.

Formally these three constraints can be expressed as:

$$p(X_i = 1|R = 0) = \bar{\xi}_i \qquad i = 1, \ldots, s \tag{5.14}$$

$$p(X_i = 1|R = 1) = \xi_i \qquad i = 1, \ldots, s \tag{5.15}$$

$$p(R = 1) = \rho \tag{5.16}$$

The constraints given in eq. 5.14 and eq. 5.15 are analogous to probabilities that, in the Binary Independence Model, are estimated as a result of relevance feedback. There, the values ξ_i and $\bar{\xi}_i$ are estimated from documents judged to be relevant and non-relevant respectively.

No attempt is made to estimate the value ρ in the Binary Independence Model. The prior odds of relevance does enter into the odds of relevance conditioned on the term occurrence pattern given in eq. 5.8. However, it is not needed for the purposes of ranking. We include constraint eq. 5.16 in order to fully mimic the log-odds of relevance formula developed in the BIM model. This constraint has something of a subordinate status in our model, also. If no reasonable value for it can be assigned, it may be treated as a parameter in the resulting probability distribution. We will see that, for the purposes of ranking, the parameter may be left undetermined.

In order to "implement" the constraints discussed above, we focus on certain *features* of the elementary events. These features are random variables; functions associating a real number with every element of ω. The features we will need are:

$$\bar{g}_i(\omega) = \left\{ \begin{array}{ll} 1 & \text{if } X_i(\omega) = 1 \wedge R(\omega) = 0 \\ 0 & \text{otherwise} \end{array} \right\} \qquad i = 1, \ldots, s$$

$$g_i(\omega) = \left\{ \begin{array}{ll} 1 & \text{if } X_i(\omega) = 1 \wedge R(\omega) = 1 \\ 0 & \text{otherwise} \end{array} \right\} \qquad i = 1, \ldots, s$$

$$g_R(\omega) = R(\omega)$$

The desired constraints on the probability distribution can be effected by constraining the expectations of these features such that:

$$E[\bar{g}_i(\omega)] = \bar{G}_i \equiv \bar{\xi}_i \cdot (1 - \rho) \qquad i = 1, \ldots, s \tag{5.17}$$

$$E[g_i(\omega)] = G_i \equiv \xi_i \cdot \rho \qquad i = 1, \ldots, s \tag{5.18}$$

$$E[g_R(\omega)] = G_R \equiv \rho \tag{5.19}$$

In eq. 5.19 we constrain the probability $p(R = 1)$ to ρ directly, since the expected value of a binary variable is simply the probability that the variable equals 1, and in eq. 5.17, we are effectively constraining $p(X_i = 1|R = 0)$ to $\bar{\xi}_i$.

5.3.2 Probability of an Arbitrary Event

To maximize the entropy subject to these constraints, we apply the Lagrange method of undetermined multipliers [14]. Introducing the multipliers, λ_0'; $\bar{\lambda}_1, \ldots, \bar{\lambda}_s$; $\lambda_1, \ldots, \lambda_s$; and λ_R, the problem of maximizing H in conformance with the constraints, 5.17 – 5.19, is transformed into the maximization of the unconstrained function:

$$\begin{aligned} H'(p) = & - \sum_{\omega \in \Omega} p(\omega) \log p(\omega) + \lambda_0'(1 - \sum_{\omega \in \Omega} p(\omega)) \\ & + \bar{\lambda}_1 \cdot [\sum_{\omega \in \Omega} p(\omega)\bar{g}_1(\omega) - \bar{G}_1] + \ldots + \bar{\lambda}_s \cdot [\sum_{\omega \in \Omega} p(\omega)\bar{g}_s(\omega) - \bar{G}_s] \\ & + \lambda_1 \cdot [\sum_{\omega \in \Omega} p(\omega) g_1(\omega) - G_1] + \ldots + \lambda_s \cdot [\sum_{\omega \in \Omega} p(\omega) g_s(\omega) - G_s] \\ & + \lambda_R \cdot [\sum_{\omega \in \Omega} p(\omega) g_R(\omega) - G_R] \end{aligned} \tag{5.20}$$

where the term, $\lambda_0'(1 - \Sigma p(\omega))$, corresponds to the constraint, applicable to any probability distribution, that the $p(\omega)$ must sum to 1. Taking the partial derivative

with respect to $p(\omega)$, for a specific event, ω, and setting the derivatives (one for each ω) equal to zero, we get:

$$p(\omega) = e^{\left[\lambda_0 + \left(\sum_{i=1}^{s} \bar{\lambda}_i \bar{r} x_i\right) + \left(\sum_{i=1}^{s} \lambda_i r x_i\right) + \lambda_R r\right]}$$

where λ_0 is used for $\lambda_0' - 1$, r is 1 if ω corresponds to a relevant document and 0 otherwise; $\bar{r} = (1 - r)$ is 1 if ω corresponds to a non-relevant document; and for $i = 1, \ldots, s$: x_i is 1 when term i occurs in the document and 0 otherwise. It is not difficult to prove (see, for example, Chapter 4 of [105]) that this solution will always be, not only a maximum, but a global maximum for the entropy.

5.3.3 BIM-MAXENT Ranking Formula

As we saw in the introduction, the ranking formula developed for traditional probabilistic systems is based on the calculation of the odds of relevance given the occurrence pattern of the query terms. Based on the model developed in the previous section, the conditional odds of relevance for the maximum entropy distribution can be calculated as:

$$\begin{aligned} \log O(rel|x_1, \ldots, x_s) &= \left(\sum_{i=1}^{s} (\lambda_i - \bar{\lambda}_i) x_i\right) + \lambda_R \\ &= \left(\sum_{x_i=1} (\lambda_i - \bar{\lambda}_i)\right) + \lambda_R \end{aligned} \tag{5.21}$$

This gives an expression for the log-odds of relevance in terms of the parameters, $\lambda_1, \ldots, \lambda_s$; $\bar{\lambda}_1, \ldots, \bar{\lambda}_s$; and λ_R. We will need to determine the values of these parameters in terms of the constraining factors, $\xi_1, \ldots, \xi_s$; $\bar{\xi}_1, \ldots, \bar{\xi}_s$; and ρ, in order to transform this ranking formula to one in terms of parameters that can be set from the data that will be available at the time of retrieval.

In [47] it is shown that for the maximum entropy solution:

$$O(X_i = 1|rel) = e^{\lambda_i} \qquad i = 1, \ldots, s \tag{5.22}$$

$$O(X_i = 1|\overline{rel}) = e^{\bar{\lambda}_i} \qquad i = 1, \ldots, s \tag{5.23}$$

$$O(rel) = e^{\lambda_R} \prod_{i=1}^{s} \frac{(e^{\lambda_i} + 1)}{(e^{\bar{\lambda}_i} + 1)} \tag{5.24}$$

and the values for the Lagrange multipliers must be:

$$\lambda_i = \log \frac{\xi_i}{1 - \xi_i} \tag{5.25}$$

$$\bar{\lambda}_i = \log \frac{\bar{\xi}_i}{1 - \bar{\xi}_i} \tag{5.26}$$

$$\lambda_R = \log \frac{\rho}{1 - \rho} + \sum_{i=1}^{s} \log \frac{1 - \xi_i}{1 - \bar{\xi}_i} \tag{5.27}$$

Since e^{λ_0} is a factor in the probability of each elementary event and λ_0 plays no other role, λ_0 is nothing more than the log of the normalization constant which forces the probabilities over elementary events to sum to 1.

Also, it is worth observing that λ_R is simply the log-odds of relevance of a document for which none of the query terms occurs:

$$\log O(rel|0, \ldots, 0) = \log \frac{p(0, \ldots, 0, rel)}{p(0, \ldots, 0, \overline{rel})} \tag{5.28}$$

$$= \log \frac{e^{[\lambda_0 + \lambda_R]}}{e^{\lambda_0}} \tag{5.29}$$

$$= \lambda_R \tag{5.30}$$

Substituting the values of the parameters derived in eq. 5.25, 5.26, 5.27 for the conditional log-odds of relevance given in eq. 5.21, we have:

$$\log O(rel|x_1, \ldots, x_s) = \sum_{x_i = 1} \log \frac{\xi_i(1 - \bar{\xi}_i)}{\bar{\xi}_i(1 - \xi_i)} + \sum_{i=1}^{s} \log \frac{1 - \xi_i}{1 - \bar{\xi}_i} + \log \frac{\rho}{1 - \rho} \tag{5.31}$$

and this is the ranking formula, eq. 5.9, of the Binary Independence Model.

5.3.4 Discussion of the BIM-MAXENT Model

Of the distributions that conform to the constraints, that with maximum entropy is the distribution of the Binary Independence Model. Two points are worthy of further discussion. First, we have not assumed independence in any form. The linked dependence condition, while not assumed, can however be shown to be a property of the derived maximum entropy distribution. Also, we have included a constraint on the prior probability of relevance. A value for this is not needed if the formula is only to be used for ranking. Nonetheless, we might like to consider estimating this probability in order to produce a ranking status value that can be interpreted as a probability. We begin with a discussion of linked dependence.

5.3.5 Linked Dependence as a Consequence of Maximum Entropy

We have not explicitly encoded the linked dependence assumption in the development of the BIM-MAXENT model. It has not been necessary. Rather than assume that the query term occurrences are conditionally independent random variables, we have chosen a probability distribution that maximizes entropy subject to a set of constraints. There has been no need to explicitly assume independence.

We shall defer further discussion of the distinction between constraints and assumptions for the moment. For now, we will show that although independence has not been assumed, the form of the independence conditions is a consequence of the Maximum Entropy Principle. More precisely stated, a property of the probability distribution that maximizes uncertainty is equivalent to a property of the *physically real* probability distribution that is assumed to hold in traditional models.

In [47] it is shown that, for an arbitrary configuration $(x_1, \ldots, x_s) \in \{0,1\}^s$:

$$\frac{p(x_1, \ldots, x_s|rel)}{p(x_1, \ldots, x_s|\overline{rel})} = e^{(\sum_{i=1}^{s}(\lambda_i - \bar{\lambda}_i)x_i)} \cdot \prod_{i=1}^{s} \frac{(e^{\bar{\lambda}_i} + 1)}{(e^{\lambda_i} + 1)}$$

whereas for each $i = 1, \ldots, s$:

$$p(X_i = x_i|\overline{rel}) = \frac{e^{\bar{\lambda}_i x_i}}{1 + e^{\bar{\lambda}_i}}$$

from which it follows that:

$$\frac{p(x_1, \ldots, x_s|rel)}{p(x_1, \ldots, x_s|\overline{rel})} = \prod_{i=1}^{s} \frac{p(X_i = x_i|rel)}{p(X_i = x_i|\overline{rel})}$$

which is the form of the linked dependence assumption discussed in Section 5.2.1.

Linked dependence, then, is not assumed. It is a property of the constrained maximum entropy distribution. There is, we believe, a significant difference between making (possibly unwarranted) assumptions and constraining the distribution. The difference is discussed in greater detail in Section 5.5.

5.3.6 Prior Probability Of Relevance

The constraints imposed on the BIM-MAXENT model include a constraint on the prior probability of relevance, $p(rel) = \rho$. It is important to note, however, that it is not necessary for the system designer to actually set ρ to a particular value. If the goal is simply to rank documents according to the probability of relevance, without making any claims as to the interpretability of the resulting ranking status value, the value assigned to ρ becomes irrelevant. It can be ignored here, as it is in the Binary Independence Model, inasmuch as the value used will not affect the order in which documents are ranked.

Even if we wanted to produce the system's probability of relevance, as opposed to a (less naturally interpretable) ranking score, we might not include the constraint on the prior probability of relevance. We would not include this constraint if we felt that we had no reason, a priori, to distinguish between relevant and non-relevant documents in any way other than that which is incorporated in the constraints, 5.17 and 5.18, regarding term occurrences. If after studying the characteristics of the

resulting probability distribution, we feel comfortable with what MAXENT is telling us, there would be no motivation for including other constraints.

In the model with prior probability of relevance unconstrained, the prior odds of relevance would be

$$O(rel) = \prod_{i=1}^{s} \frac{(e^{\lambda_i} + 1)}{(e^{\bar{\lambda}_i} + 1)} = \prod_{i=1}^{s} \frac{1 - \bar{\xi}_i}{1 - \xi_i} \tag{5.32}$$

for the maximum entropy distribution. This might cause little consternation. It does not, on the surface, seem to conflict with any preconceived notions we have concerning the relevance of documents. At first glance, a need for constraining $p(rel)$, thereby constraining $O(rel)$, is not apparent.

We would also notice, however, that in the model without the $p(rel)$ constraint, the odds of relevance for a document with none of the query terms occurring is:

$$O(rel|0, \ldots, 0) = \frac{e^{\lambda_0}}{e^{\lambda_0}} = 1 \tag{5.33}$$

This is nettlesome. The system designer will likely feel that the probability of a document in which none of the query terms are to be found is very far below $\frac{1}{2}$. This discrepancy is indicative of an under-constrained distribution. MAXENT is signaling that some pertinent knowledge has not been incorporated into the model. If the goal is for the system to present its probability of relevance to the user and the system's belief system is to mirror the designer's belief system, then some constraint must be added.

One obvious way to accomplish this, given that the weakness of the model has become apparent in the value it gives for $O(rel|x_1, \ldots, x_s)$, would be to constrain $p(rel|0, \ldots, 0)$ directly. This can be done, but it may not be the best approach. In typical IR system design situations most people would assign a very small value for

$p(rel|0,\ldots,0)$. The problem is that humans are notoriously poor at dealing with very small ($p(\ldots) \approx 0$) and very large ($p(\ldots) \approx 1$) probabilities.

Alternatively, an empirical approach might be taken. By studying a large number of queries, the value given to the conditional probability, $p(rel|0,\ldots,0)$, can be based on statistics of the data. Unfortunately, the extremely small probability that a document with no query terms would be found to be relevant comes to haunt us again. For such a small probability a very large sample would be needed. If the sample is not large enough we would not have much confidence in the resulting value of the statistic. For example, even for a reasonably large sample of queries against a large collection, there may well be no instance of a document containing none of the query terms having been judged relevant.

A preferable approach is to estimate the prior probability of relevance and utilize this as a constraint on the distribution as was done in BIM-MAXENT with constraint, 5.19. In the version of BIM-MAXENT with all three constraints, this problem does not arise, since the odds of relevance given no query terms is given by:

$$O(rel|0,\ldots,0) \quad = \quad e^{\lambda_R} = \frac{\rho}{1-\rho}\prod_{i=1}^{s}\frac{e^{\bar{\lambda}_i}+1}{e^{\lambda_i}+1} = \frac{\rho}{1-\rho}\prod_{i=1}^{s}\frac{1-\xi_i}{1-\bar{\xi}_i} \tag{5.34}$$

Implicit in constraining $p(rel)$ is a constraint on $O(rel|0,\ldots,0)$. Presumably ρ, and hence $\frac{\rho}{1-\rho}$ will have been constrained to be small. We also expect that, for each i, the constraints, $\xi_i = p(x_i \mid rel)$ and $\bar{\xi}_i = p(x_i \mid \overline{rel})$ will be in the relation, $\xi_i > \bar{\xi}_i$, which would mean that $\frac{1-\xi_i}{1-\bar{\xi}_i} < 1$, making $O(rel|0,\ldots,0)$ smaller still. This conforms to the prior knowledge that we desire to incorporate in our retrieval system. The system designer may depend on her own subjective judgment, empirical study, or some combination of the two. However it is done, constraining the prior probability of relevance will be a better approach to incorporating the knowledge that is felt to be missing in the two-constraint version of the model, when we come to realize that this version would entail even odds for a document with no query terms.

5.4 The CM-MAXENT Retrieval Model

In this section we apply the same approach to a modified set of constraints in order to derive the CMM.

5.4.1 Basic CM-MAXENT Model

For this model the constraints are:

$$p(X_i = 1|R = 0) = \bar{\xi}_i \qquad i = 1, \ldots, s \tag{5.35}$$

$$E(X_\#|R = 1) = \zeta \qquad \text{where: } X_\# = \sum_{i-1}^{s} X_i \tag{5.36}$$

$$p(R = 1) = \rho \tag{5.37}$$

The second constraint restricts the probability distributions under consideration to those with a given value for the expected number of query terms occurring in a relevant document. It will not be necessary that a value for this expectation be explicitly specified, however. The constraint will result in the inclusion of a parameter in the distribution and, as we will see, a number of alternatives for determining a value for this parameter will be available.

The following features of the elementary events:

$$\bar{g}_i(\omega) = \left\{ \begin{array}{ll} 1 & \text{if } X_i(\omega) = 1 \wedge R(\omega) = 0 \\ 0 & \text{otherwise} \end{array} \right\} \qquad i = 1, \ldots, s \tag{5.38}$$

$$g_\#(\omega) = \left\{ \begin{array}{ll} (X_1 + \cdots + X_s) & \text{if } R(\omega) = 1 \\ 0 & \text{otherwise} \end{array} \right\} \tag{5.39}$$

$$g_R(\omega) = R(\omega) \tag{5.40}$$

are be constrained by:

$$E[\bar{g}_i(\omega)] = \bar{G}_i = \bar{\xi}_i \cdot (1 - \rho) \qquad i = 1, \ldots, s \tag{5.41}$$

$$E[g_{\#}(\omega)] = G_{\#} = \zeta \cdot \rho \tag{5.42}$$

$$E[g_R(\omega)] = G_R = \rho \tag{5.43}$$

In eq. 5.42, we are effectively constraining $E(X_1 + \cdots + X_s | R = 1)$ to ζ.

By introducing Lagrange multipliers, setting partial derivatives to zero and solving for $p(\omega)$, we get:

$$p(\omega) = e^{\left[\lambda_0 + \left(\sum_{i=1}^{s} \bar{\lambda}_i \bar{r} x_i\right) + \lambda_{\#} x_{\#} r + \lambda_R r\right]} \tag{5.44}$$

where $x_{\#} = \sum_{i=1}^{s} x_i$.

Based on the probability distribution, eq. 5.44, we can determine the odds of relevance given a specific occurrence pattern:

$$O(rel|x_1, \ldots, x_s) = \frac{e^{\left[\lambda_0 + \lambda_{\#} x_{\#} + \lambda_R\right]}}{e^{\left[\lambda_0 + \left(\sum_{i=1}^{s} \bar{\lambda}_i x_i\right)\right]}} = e^{\lambda_{\#} x_{\#} + \lambda_R - \sum_{i=1}^{s} \bar{\lambda}_i x_i} \tag{5.45}$$

and, therefore, a ranking formula based on the conditional log-odds of relevance is:

$$\log O(rel|x_1, \ldots, x_s) = \lambda_{\#} x_{\#} - \sum_{x_i=1} \bar{\lambda}_i + \lambda_R \tag{5.46}$$

Here again λ_R is simply the log-odds of relevance conditioned on all query terms being absent, $\log O(rel|0, \ldots, 0)$, which is a constant and can be dropped for the purposes of ranking.

5.4.2 Characteristics of the CM-MAXENT Distribution

As with the BIM-MAXENT model, we can derive closed form solutions for the odds of certain events.

$$O(X_i = 1|rel) = e^{\lambda_{\#}} \tag{5.47}$$

$$O(X_i = 1|\overline{rel}) = e^{\bar{\lambda}_i} \tag{5.48}$$

$$O(rel) = e^{\lambda_R} \cdot \prod_{i=1}^{s} \frac{(e^{\lambda_\#} + 1)}{(e^{\bar{\lambda}_i} + 1)} \tag{5.49}$$

We note here that $e^{\lambda_\#}$ is independent of the values of the X_i, and so the probability of occurrence given relevance is the same for all query terms. Equal probabilities are assumed in the CMM. But, as with linked dependence for BIM, it appears as a property of the CM-MAXENT distribution as a consequence of maximizing the entropy.

From eq. 5.48 and the constraint given in eq. 5.35, we have that $\bar{\lambda}_i = \log \frac{\bar{\xi}_i}{1-\bar{\xi}_i}$. If, following Croft and Harper, we use $\frac{n_i}{N}$ for $\bar{\xi}_i$, where N is the total number of documents in the collection, and n_i is the number of documents in which query term i occurs, we have:

$$\bar{\lambda}_i = \log \frac{\bar{\xi}_i}{1-\bar{\xi}_i} = \log \frac{\frac{n_i}{N}}{1-\frac{n_i}{N}} = \log \frac{n_i}{N-n_i}$$

Using this in eq. 5.46 gives the formula:

$$\log O(rel|x_1, \ldots, x_s) = \lambda_\# \cdot x_\# + \left(\sum_{x_i=1} \log \frac{N-n_i}{n_i} \right) + \lambda_R \tag{5.50}$$

The first term is just a constant, $\lambda_\#$, multiplied by the number of terms that occur in the document. Taking into consideration that the last term, λ_R, is independent of the term occurrence variables and can be ignored for the purposes of ranking, we have the equivalent of the Combination Match Model formula.

5.4.3 Discussion of the CM-MAXENT Model

By exchanging the constraints on the probabilities of occurrence in relevant documents for a single constraint on the expected number of terms appearing in relevant documents, we obtain the CMM for document ranking,

$$\sum_{x_i=1}^{n} \left(\lambda_\# + \log \tfrac{N-n_i}{n_i} \right)$$

Croft and Harper point out that CMM is a generalization of the inverse document frequency weighting scheme originally proposed by Sparck Jones. It is interesting to note what happens if we ease the constraints on our probabilities in the CM-MAXENT model. In this section we will show how we can get a pure *idf* ranking formula by eliminating the constraint with respect to term occurrence in relevant documents. We will also show that elimination of the constraint on term occurrence in the non-relevant documents can be compared to the observation made by Croft and Harper that, in essence, a coordination match formula results from assuming, in their model, that the probability of a term occurring in a relevant document is very large. We continue in this section with a discussion of how assumptions in the CMM are properties of the CM-MAXENT model. This is analogous to the situation in the Binary Independence Model, where the linked dependence assumption turns out to be a property of the BIM-MAXENT version of the model. Finally, we discuss the constraint in CM-MAXENT on the expected number of query terms for relevant documents and approaches to associating a value with the constraint.

5.4.4 A MAXENT Version of *idf* Weighting

If we eliminate constraint eq. 5.42 respecting the expected value of the number of terms to be found in a relevant document, the ranking formula assumes the form:

$$\log O(rel|x_1,\ldots,x_s) \;=\; \left(\sum_{x_i=1} \log \frac{1-\bar{\xi}_i}{\bar{\xi}_i}\right) + \left(\sum_{i=1}^{s} \log \frac{2}{1-\bar{\xi}_i}\right) + \log \frac{\rho}{1+\rho} \quad (5.51)$$

Since the two terms at the right are constant over all documents, eq. 5.51 is equivalent to ranking by summing weights associated with each of the occurring query terms. If, as above, $\frac{n_i}{N}$ is used for $\bar{\xi}_i$, this is equivalent to the weighting scheme originally proposed by Sparck Jones with the minor difference that $\log \frac{N-n_i}{n_i}$ is used in place of $\log \frac{N}{n_i}$ for the term weights. The Sparck Jones weighting formula can

therefore be interpreted as the maximum entropy distribution constrained only so that $p(x_i \mid \overline{rel}) = \bar{\xi}_i$.

5.4.5 A MAXENT Version of Coordination Matching

In a similar fashion, we can consider a model in which knowledge concerning term occurrences in the collection as a whole is not used to constrain the distribution. In the absence of the constraints specified in eq. 5.41 the conditional log-odds of relevance would be:

$$\log O(rel|x_1, \ldots, x_s) = \lambda_{\#} x_{\#} + \lambda_R$$

In this formula, both $\lambda_{\#}$and λ_R are constant and both can be ignored for the purpose of ranking. The formula, a linear function of the number of query terms appearing in a document, is equivalent to coordination match ranking.

5.4.6 Assumptions of the Combination Match Model

As with the Binary Independence Model, no assumptions have been made in the MAXENT version of the combination match model. Neither the linked dependence assumption nor the Croft and Harper assumption of equal probabilities of occurrence in relevant documents is made in CM-MAXENT. Here, as before, the properties assumed in the classic models turn out to be true of the derived MAXENT probability distributions. The essence of the arguments given in favor of linked dependence in Section 5.3.5 hold for the CMM. Also, we have seen that the odds of occurrence in a relevant document is

$$O(X_i = 1|rel) = e^{\lambda_{\#}} \tag{5.52}$$

and hence is the same for all query terms. The property of equal probabilities of occurrence, assumed in the classical combination match model, is shown to be a

property, as well, of the maximum entropy distribution. The difference between a relation being assumed to hold and the relation arising as a property of a constrained maximum entropy distribution is an important one and is discussed in more detail in Section 5.5.

5.4.7 The $E[g_{\#}(\omega)]$ Constraint

In Section 5.3.4 we saw that the constraint on the probability of relevance was unnecessary for the purposes of ranking. We also discussed what steps might be taken if a ranking status value that can be interpreted as a probability is desired. The constraint on the expected value of the number of terms appearing in the relevant documents is somewhat different. Its value must be determined for ranking. Nonetheless, the constraint need not be specified explicitly. The Croft and Harper approach can be taken. The parameter $\lambda_{\#}$ can be left undetermined in the ranking formula and set as the result of empirical testing so as to yield the best possible retrieval results.

The MAXENT approach provides an interesting alternative. If there is data on which to base the setting of the constant, $\lambda_{\#}$, based on retrieval experiments, this same data could be used to estimate $E[X_{\#}|rel]$ directly. The same document collection, query set and relevance judgments that are used to analyze retrieval performance can be used to estimate the expected number of query terms appearing in relevant documents. An interesting option here is that $E[X_{\#}|rel]$ might be estimated as a function of query characteristics, yielding a query specific probability distribution on which conditional probabilities of relevance are calculated. A characteristic which comes immediately to mind in this regard is the number of terms in the query.

5.5 Discussion of Maximum Entropy Modeling

Both the Binary Independence Model and the Combination Match Model can be derived from the maximum entropy approach with appropriate constraints. In this section we analyze in further detail the difference between the maximum entropy approach and the classical approaches based on *a priori* assumptions. We attempt to signal both the philosophical and practical importance of this distinction to the conduct of IR research. We emphasize that constraining a distribution is not the same as making, possibly unwarranted, *a priori* assumptions. This becomes most clear in the case of the assumption of equal probabilities of occurrence in relevant documents made in the CMM. We assert in this section that thinking in terms of constraints results in greater adaptability when we encounter previously uncontemplated sources of knowledge that can be applied to document ranking. A unifying thread running through all of the following discussion is the notion that the probabilities manipulated by probabilistic retrieval systems can not reasonably be construed as frequencies. We begin with a discussion of difficulties inherent in the interpretation of the Probability Ranking Principle.

5.5.1 Probability Ranking Principle

In [88], Robertson gives a formal statement of the Probability Ranking Principle as originally put forth in an unpublished memorandum by William Cooper:

> If a reference system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data.

Use of the phrase "probabilities are estimated as accurately as possible", as well as the nature of the arguments in the body of the paper, indicate that a frequentist interpretation of probability is intended. But then we are cautioned that the estima-

tion is to be made on the basis of "whatever data has been made available". This is problematic.

Let's recall momentarily the case of the die that has been tossed millions of times with an average of 5.0. This is certainly knowledge "available to the system"; it has a bearing on the probability of the next toss revealing a 4.

The situation is the same in IR. Suppose we have a substantial theory, based on the study of extensive retrieval data. Let us suppose furthermore that this theory permits us to produce a well calibrated [26, 79, 25] estimate of the probability of relevance of a document to a query containing a term as a function of collection and document statistics with respect to the term. Now what do we do if we have a two word query? Our theory provides us with two probability estimates. Both are *correct*. The Probability Ranking Principle counsels us to use all evidence.

The Probability Ranking Principle, doesn't, however, advise on how this is to be done. The problem that arises for two word queries, is exacerbated for three word queries, more so for four word queries, and more so for twenty word queries. Perhaps further study of retrieval data will result, at a later date, in a more sophisticated model that will offer guidance as to how best to assess the probability of relevance based on statistical characteristics of all the query terms collectively. In the meantime "as accurate as possible" a probability of relevance must be estimated in the absence of such a theory. We appear to be at an impasse.

We have two estimates; both are as accurate as possible; we are enjoined to use all of the data at our disposal; we have no estimate at all based on all of the data. Our conclusion is that:

1. if we are to exploit all of the data, we are obliged to abandon the frequentist notion that the objective is the estimation of a true physical probability;

2. the alternative is to view the objective as the generation of a subjective probability – the system's belief that a document is relevant;

3. a guiding principle must be adopted for the determination of this probability based on knowledge possessed by the system;

4. the Maximum Entropy Principle is a very reasonable candidate.

5.5.2 Constraints are not Assumptions

The Binary Independence Model assumes that occurrence of query terms is independent in both the relevant and non-relevant documents. Both intuition and experimental evidence imply that such an assumption is unwarranted. There is little reason to believe these assumptions are even approximately correct. Unfortunately, attempts to model term dependence have been disappointing [110, 54, 102]. The problem is generally attributed to the inability to produce accurate probability estimates due to insufficient sample sizes. So, we return to the independence assumptions. But, what is the justification for basing a model on assumptions in which we have so little faith?

We suggest that "independence" in the Binary Independence Model should not really be thought of as an assumption at all. Rather, incorporating independence is an attempt to make the most reasonable use of the information that is available, accepting that there is information that could be very useful if only we had access to it, but we don't. The MAXENT approach makes this explicit. In BIM-MAXENT, there is no assumption of independence. In place of assumptions, we have constraints.

A constraint is not an assumption. Nothing is being assumed to be "true". No "physically real" population is presumed to exist, so there is nothing we can "assume" about it. When we constrain the probability of term occurrence in a relevant document to ξ_i, we are not saying that this is an estimate of the proportion of relevant documents that contain the term in some super-population of documents. We are saying that based on the evidence we have, a probability distribution for which

$p(X_i = 1 \mid R = 1) = \xi_i$ is the most reasonable distribution for us to accept, given what we know.

The probability produced by the BIM-MAXENT model is not an estimate of a true physical probability. It is a subjective probability. It is the system's subjective probability that the document will be judged relevant by the user. Again we turn to the analogy of the dice. When, after learning that the average of a large number of tosses of the die is 5.0, MAXENT assigns a probability of 0.136 for a die coming up 4 on the next throw, it is not producing an estimate. An estimate of what could it be? Perhaps, an estimate of the fraction of tosses in the universe of dice with expected values of 5.0 that come up 4:

$$\frac{\#(\{t|t \text{ is toss of a die } d \wedge E[d] = 5.0 \wedge \text{value of } t \text{ is } 4\})}{\#(\{t|t \text{ is toss of a die } d \wedge E[d] = 5.0\})} \tag{5.53}$$

Even if we were willing to contemplate such a population, on what basis would we estimate the fraction involved?

The frequentist may complain that the interpretation that we give to the probability, 0.136 is unscientific, or even less charitably, meaningless. We are not unsympathetic with regard to this reaction. But, then it seems that the frequentist is forced to conclude that there is no basis at all on which to rank the dice. We prefer to forge ahead, in spite of the difficulties involved.

We assert that it is misleading to conceptualize as estimates the probabilities on which the Binary Independence Model is based. If the design objective is to produce an estimate, it becomes very difficult to understand why an assumption of something known not to hold, even approximately, would be used to improve the estimation procedure.

5.5.3 Equal Probabilities Assumption of CMM

Croft and Harper state that, "prior to relevance feedback, we have 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." They are certainly not assuming this is true in a frequentist sense. They clearly state that they can assume equal probabilities because they have "no information". They can't mean that the absence of information implies something concrete, and very specific, about the material universe.

We take the liberty here of speaking for them, rephrasing what they said based on our perception of what they had in mind: "we have no information and therefore we should adopt the probability distribution that best expresses our uncertainty". The Maximum Entropy Principle asserts that our uncertainty is best expressed by the distribution with greatest entropy subject to constraints embodying knowledge we feel we do possess. The development of the CM-MAXENT model presented here clarifies, we believe, the conceptual position of the original authors.

5.5.4 Flexibility of Constraints

An advantage of the MAXENT approach is that it naturally accommodates the introduction of added constraints. Assumptions such as linked independence in BIM, and the equal probabilities of term occurrences conditioned on relevance in CMM, have been shown to exist in the corresponding MAXENT versions in the form of properties of the constrained distribution. We may decide to bring more information to bear in the MAXENT models, and as a result, these properties may no longer hold.

For example, suppose that based on a study of retrieval data, we are able to develop a reliable model of the distributions of document length for both relevant and non-relevant documents. This is pertinent knowledge. Even though we have no

knowledge of these distribution for the particular query in question, knowledge, albeit general knowledge, can and should be brought to bear.

It is not immediately clear how knowledge such as this can be integrated into models such as BIM and CMM. The MAXENT approach, on the other hand, guides us as to how to proceed. What we would do is incorporate the information we had discerned concerning the two conditional distributions as further constraints on our overall probability distribution. While the mathematical difficulties that may be involved must not be minimized, the maximum entropy approach does provide a theoretical foundation for how best to proceed.

5.6 Maximum Entropy and Weights of Evidence

In this section we generalize the reasoning originally used in the development of the BIM-MAXENT and CM-MAXENT models. Also, we explicitly present the results in terms of weight of evidence, linking the Maximum Entropy Principle with the formalism introduced in Chapter 3. This will prepare the way for the derivation of the models to be discussed in the succeeding chapter.

5.6.1 Constraints

The theorem to be proved in this section, which we shall refer to as the MAXENT-WOE Theorem, asserts that the weight of evidence shall be additive for the maximum entropy distribution if the constraints are of a certain restricted form. For the theorem, we shall assume that the event space, Ω, can be factored as:

$$\Omega = \mathcal{H} \times \mathcal{E}_1 \times \ldots \times \mathcal{E}_m$$

where $\mathcal{H}$ is intended to be interpreted as the space of m_H hypotheses, $h_1, \ldots, h_{m_H}$, and each $\mathcal{E}_i$ is interpreted as the space corresponding to a separate *source of evidence*. In what follows, we will use h and h' for variables that range over possible hypotheses,

and e_i for variables that range over the possible values that may be observed for the i^{th} source of evidence. As before, ω will be used as a variable to range over all possible (compound) events.

We shall allow for three classes of constraints:

hypothesis constraints: A family of χ^H constraints on the prior probability distribution over the set of possible hypotheses, of the form:

$$E[\Theta_j^H(\omega)] = \xi_j^H \qquad (5.54)$$

where Θ_j^H depends only on $h = H(\omega)$

for $j = 1, \ldots, \chi^H$. Each member of this family places a constraint on the expected value of some function, Θ_j^H, of the hypothesis variable, $H(\omega)$.

evidence constraints: A family of χ^E constraints on the probability distribution over the sources of evidence:

$$E[\sum_{i=1}^{m} \eta_{ji}^E \cdot \Theta_{ji}^E(\omega)] = \xi_j^E \qquad (5.55)$$

where each Θ_{ji}^E depends only on $e_i = E_i(\omega)$

for $j = 1, \ldots, \chi^E$. Each member of this family places a constraint on the expected value of a linear combination of a sequence of functions, $\Theta_{j1}^E, \ldots, \Theta_{jm}^E$, of the individual source-of-evidence variables, $E_1(\omega), \ldots, E_m(\omega)$.

conditional constraints: A family of χ^C constraints on the probabilities of evidence conditioned on individual hypotheses:

$$\sum_{k=1}^{m_H} \alpha_{jk}^C \cdot E[\sum_{i=1}^{m} \beta_{jki}^C \Theta_{jki}^C(\omega) | h_k] = \xi_j^C \qquad (5.56)$$

where each Θ^C_{jki} depends only on $e_i = E_i(\omega)$

for $j = 1, \ldots, \chi^C$. Each of these constraints places a restriction on a linear combination of conditional expectations. Each expectation is of some function of the evidence conditioned on one of the possible values for the hypothesis variable. Each of these functions, in turn, is a linear combination of functions of the individual sources of evidence.

5.6.2 The MAXENT-WOE Theorem

Theorem 5.1: *For the probability distribution over the event space, $\Omega = \mathcal{H} \times \mathcal{E}_1 \times \ldots \times \mathcal{E}_m$, subject to a set of constraints, weight of evidence is additive:*

$$woe(h/h' : e_1, \ldots, e_m) \;=\; \sum_{i=1}^{m} woe(h/h' : e_i)$$

if the constraints are in the form of hypothesis constraints, evidence constraints, and conditional constraints, as defined above.

proof: For any of the conditional constraints:

$$\begin{aligned} \sum_{k=1}^{m_H} \alpha^C_{jk} \cdot E[\sum_{i=1}^{m} \beta^C_{jki} \Theta^C_{jki}(\omega) | h_k] &= \sum_{k=1}^{m_H} \sum_{i=1}^{m} \alpha^C_{jk} \beta^C_{jki} E[\Theta^C_{jki}(\omega) | h_k] \\ &= \sum_{k=1}^{m_H} \sum_{i=1}^{m} \alpha^C_{jk} \beta^C_{jki} E[\Theta^{C'}_{jki}(\omega)] / p(h_k) \qquad (5.57) \end{aligned}$$

$$\text{where:} \qquad \Theta^{C'}_{jki}(\omega) = \begin{cases} \Theta^C_{jki}(\omega) & \text{if } h = h_k \\ 0 & \text{otherwise} \end{cases}$$

For $k = 1, \ldots, m_H$, let $\pi^*_k = p(h_k)$, the probability of the hypothesis, h_k, corresponding to the distribution of maximum entropy. Then, for any distribution with these marginal probabilities, $p(h_1), \ldots, p(h_m)$, it follows that:

$$\sum_{k=1}^{m_H} \alpha^C_{jk} \cdot \sum_{i=1}^{m} \beta^C_{jki} E[\Theta^C_{jki}(\omega) | h_k] = \xi^C_j \quad \textit{iff} \quad \sum_{k=1}^{m_H} \sum_{i=1}^{m} \alpha^C_{jk} \beta^C_{jki} E[\Theta^{C'}_{jki}(\omega)] / \pi^*_k = \xi^C_j$$

$$\text{iff} \quad \sum_{k=1}^{m_H}\sum_{i=1}^{m}\eta_{jki}^{C'}E[\Theta_{jki}^{C'}(\omega)] = \xi_j^C$$

$$\text{iff} \quad E[\sum_{k=1}^{m_H}\sum_{i=1}^{m}\eta_{jki}^{C'}\Theta_{jki}^{C'}(\omega)] = \xi_j^C \tag{5.58}$$

where: $\eta_{jki}^{C'} = \beta_{jki}^C \cdot \alpha_{jk}^C/\pi_k^* \qquad j = 1,\ldots,\chi^C;\ k = 1,\ldots,m_H;\ i = 1,\ldots,m$

Based on this, the conditional constraints given in eq. 5.56 can be replaced by constraints of the form:

$$E[\sum_{k=1}^{m_H}\sum_{i=1}^{m}\eta_{jki}^{C'}\Theta_{jki}^{C'}(\omega)] = \xi_j^C \tag{5.59}$$

where each $\Theta_{jki}^{C'}$ depends only on $e_i = E_i(\omega)$

without affecting the maximum entropy solution that will be obtained.

So, we continue assuming that all constraints are of the form given in eq. 5.54, eq. 5.55, and eq. 5.59. As we saw in Section 5.3.2, we can apply the Lagrange method of undetermined multipliers in order to maximize the entropy subject to these constraints. By introducing the multipliers, λ_0'; λ_j^H for $j = 1,\ldots,\chi^H$; λ_j^E for $j = 1,\ldots,\chi^E$; and λ_j^C for $j = 1,\ldots,\chi^C$; the problem of maximizing the constrained function:

$$H = -\sum_{\omega\in\Omega} p(\omega)\log p(\omega)$$

becomes the problem of maximizing the unconstrained function:

$$\begin{aligned} H'(p) = \ & -\sum_{\omega\in\Omega} p(\omega)\log p(\omega) \\ & + \lambda_0' \cdot [(\sum_{\omega\in\Omega} p(\omega)) - 1] \\ & + \sum_{j=1}^{\chi^H}\lambda_j^H \cdot [\ (\sum_{\omega\in\Omega} p(\omega)\Theta_j^H(\omega)) - \xi_j^H\ ] \\ & + \sum_{j=1}^{\chi^E}\lambda_j^E \cdot [\ (\sum_{\omega\in\Omega} p(\omega)\sum_{i=1}^{m}\eta_{ji}^E\Theta_{ji}^E(\omega)) - \xi_j^E\ ] \\ & + \sum_{j=1}^{\chi^C}\lambda_j^C \cdot [\ (\sum_{\omega\in\Omega} p(\omega)\sum_{k=1}^{m_H}\sum_{i=1}^{m}\eta_{jki}^{C'}\Theta_{jki}^{C'}(\omega)) - \xi_j^C\ ] \end{aligned} \tag{5.60}$$

where the term, $\lambda_0'(1 - \Sigma p(\omega))$, corresponds to the constraint that the probabilities, $p(\omega)$, must sum to 1.

Taking the partial derivative with respect to $p(\omega)$, for a specific event, ω, gives:

$$\begin{aligned}\frac{\partial}{\partial p(\omega)}H' &= -1-\log p(\omega)+\lambda_0' &+ \sum_{j=1}^{\chi^H}\lambda_j^H\Theta_j^H(\omega)\\ &&+ \sum_{j=1}^{\chi^E}\lambda_j^C\sum_{i=1}^{m}\eta_{ji}^E\Theta_{ji}^E(\omega)\\ &&+ \sum_{j=1}^{\chi^C}\lambda_j^C\sum_{k=1}^{m_H}\sum_{i=1}^{m}\eta_{jki}^{C'}\Theta_{jki}^{C'}(\omega)\end{aligned}$$

Using λ_0 for $\lambda_0' - 1$ and setting the derivatives (one for each ω) equal to zero, we get:

$$\log p(\omega) = \lambda_0 + \sum_{j=1}^{\chi^H}\lambda_j^H\Theta_j^H(\omega) + \sum_{j=1}^{\chi^E}\lambda_j^E\sum_{i=1}^{m}\eta_{ji}^E\Theta_{ji}^E(\omega) + \sum_{j=1}^{\chi^C}\lambda_j^C\sum_{k=1}^{m_H}\sum_{i=1}^{m}\eta_{jki}^{C'}\Theta_{jki}^{C'}(\omega)$$

and hence,

$$p(\omega) = e^{\left[\lambda_0+\sum_{j=1}^{\chi^H}\lambda_j^H\Theta_j^H(\omega)+\sum_{j=1}^{\chi^E}\lambda_j^E\sum_{i=1}^{m}\eta_{ji}^E\Theta_{ji}^E(\omega)+\sum_{j=1}^{\chi^C}\lambda_j^C\sum_{k=1}^{m_H}\sum_{i=1}^{m}\eta_{jki}^{C'}\Theta_{jki}^{C'}(\omega)\right]}$$

Taking advantage of the restricted dependencies of Θ_j^H, Θ_{ji}^E and $\Theta_{jki}^{C'}$ this can be written as:

$$\begin{aligned}p(\omega) &= e^{\left[\lambda_0+\sum_{j=1}^{\chi^H}\lambda_j^H\Theta_j^H(h)+\sum_{j=1}^{\chi^E}\lambda_j^E\sum_{i=1}^{m}\eta_{ji}^E\Theta_{ji}^E(e_i)+\sum_{j=1}^{\chi^C}\lambda_j^C\sum_{k=1}^{m_H}\sum_{i=1}^{m}\eta_{jki}^{C'}\Theta_{jki}^{C'}(h,e_i)\right]}\\ &= e^{\left[\lambda_0+\sum_{j=1}^{\chi^H}\lambda_j^H\Theta_j^H(h)+\sum_{i=1}^{m}\left(\sum_{j=1}^{\chi^E}\lambda_j^E\eta_{ji}^E\Theta_{ji}^E(e_i)+\sum_{j=1}^{\chi^C}\sum_{k=1}^{m_H}\lambda_j^C\eta_{jki}^{C'}\Theta_{jki}^{C'}(h,e_i)\right)\right]}\\ &= e^{\lambda_0}\cdot e^{\left[\sum_{j=1}^{\chi^H}\lambda_j^H\Theta_j^H(h)\right]}\cdot\prod_{i=1}^{m}e^{\left[\sum_{j=1}^{\chi^E}\lambda_j^E\eta_{ji}^E\Theta_{ji}^E(e_i)+\sum_{j=1}^{\chi^C}\sum_{k=1}^{m_H}\lambda_j^C\eta_{jki}^{C'}\Theta_{jki}^{C'}(h,e_i)\right]}\\ &= e^{\lambda_0}\cdot\Phi_H(h)\cdot\prod_{i=1}^{m}\Phi_i(h,e_i)\end{aligned}$$

$$
\begin{aligned}
\text{where:}\quad h &= H(\omega) \\
e_i &= E_i(\omega) \\
\text{and:}\quad \Phi_H(h) &= e^{\left[\sum_{j=1}^{\chi^H} \lambda_j^H \Theta_j^H(h)\right]} \\
\Phi_i(h, e_i) &= e^{\left[\sum_{j=1}^{\chi^E} \lambda_j^E \eta_{ji}^E \Theta_{ji}^E(e_i) + \sum_{j=1}^{\chi^C} \sum_{k=1}^{m_H} \lambda_j^C \eta_{jki}^{C'} \Theta_{jki}^{C'}(h, e_i)\right]}
\end{aligned}
$$

By summing up over all possible combinations of values for the sources of evidence,

$$< e_1, \ldots, e_m > \in \mathcal{E}_1 \times \ldots \times \mathcal{E}_m$$

we get an expression for the probability of an hypothesis, $p(h)$:

$$
\begin{aligned}
p(h) &= \sum_{\mathcal{E}_1 \times \ldots \times \mathcal{E}_m} p(h, e_1, \ldots, e_m) \\
&= \sum_{\mathcal{E}_1 \times \ldots \times \mathcal{E}_m} e^{\lambda_0} \cdot \Phi_H(h) \cdot \prod_{i=1}^{m} \Phi_i(h, e_i) \\
&= e^{\lambda_0} \cdot \Phi_H(h) \cdot \sum_{\mathcal{E}_1 \times \ldots \times \mathcal{E}_m} \prod_{i=1}^{m} \Phi_i(h, e_i) \\
&= e^{\lambda_0} \cdot \Phi_H(h) \cdot \prod_{i=1}^{m} \sum_{e_i \in \mathcal{E}_1} \Phi_i(h, e_i) \qquad (5.61)
\end{aligned}
$$

Similarly, we can sum up over all possible combinations of values for the sources of evidence, $< e_2, \ldots, e_m > \in \mathcal{E}_2 \times \ldots \times \mathcal{E}_m$, to get $p(h, e_1)$:

$$
\begin{aligned}
p(h, e_1) &= \sum_{\mathcal{E}_2 \times \ldots \times \mathcal{E}_m} p(h, e_1, e_2, \ldots, e_m) \\
&= \sum_{\mathcal{E}_2 \times \ldots \times \mathcal{E}_m} e^{\lambda_0} \cdot \Phi_H(h) \cdot \prod_{i=1}^{m} \Phi_i(h, e_i) \\
&= \sum_{\mathcal{E}_2 \times \ldots \times \mathcal{E}_m} e^{\lambda_0} \cdot \Phi_H(h) \cdot \Phi_1(h, e_1) \cdot \prod_{i=2}^{m} \Phi_i(h, e_i) \\
&= e^{\lambda_0} \cdot \Phi_H(h) \cdot \Phi_1(h, e_1) \cdot \sum_{\mathcal{E}_2 \times \ldots \times \mathcal{E}_m} \prod_{i=2}^{m} \Phi_i(h, e_i)
\end{aligned}
$$

$$= \quad e^{\lambda_0} \cdot \Phi_H(h) \cdot \Phi_1(h, e_1) \cdot \prod_{i=2}^{m} \sum_{e'_i \in \mathcal{E}_i} \Phi_i(h, e'_i) \tag{5.62}$$

Using eq. 5.61 and eq. 5.62 we can determine the probabilities for the first source of evidence conditioned on a hypothesis, $p(e_1 \mid h)$:

$$\begin{aligned} p(e_1 \mid h) = \frac{p(h, e_1)}{p(h)} &= \frac{e^{\lambda_0} \cdot \Phi_H(h) \cdot \Phi_1(h, e_1) \cdot \prod_{i=2}^{m} \sum_{e'_i \in \mathcal{E}_1} \Phi_i(h, e'_i)}{e^{\lambda_0} \cdot \Phi_H(h) \cdot \prod_{i=1}^{m} \sum_{e'_i \in \mathcal{E}_1} \Phi_i(h, e'_i)} \\ &= \frac{\Phi_1(h, e_1)}{\sum_{e'_1 \in \mathcal{E}_1} \Phi_1(h, e'_1)} \end{aligned} \tag{5.63}$$

For the sake of concreteness, we have focused on the conditional probabilities for observations of the first source of evidence. However, eq. 5.63 generalizes to an arbitrary source of evidence:

$$p(e_i \mid h) \quad = \quad \frac{\Phi_i(h, e_i)}{\sum_{e'_i \in \mathcal{E}_i} \Phi_i(h, e'_i)} \tag{5.64}$$

Now, for the probability of a complete set of observations conditioned on a given hypothesis, we have:

$$\begin{aligned} p(e_1, \ldots, e_m \mid h) &= \frac{p(h, e_1, \ldots, e_m)}{p(h)} \\ &= \frac{e^{\lambda_0} \cdot \Phi_H(h) \cdot \prod_{i=1}^{m} \Phi_i(h, e_i)}{e^{\lambda_0} \cdot \Phi_H(h) \cdot \prod_{i=1}^{m} \sum_{e_i \in \mathcal{E}_1} \Phi_i(h, e_i)} = \frac{\prod_{i=1}^{m} \Phi_i(h, e_i)}{\prod_{i=1}^{m} \sum_{e'_i \in \mathcal{E}_1} \Phi_i(h, e'_i)} \\ &= \prod_{i=1}^{m} \frac{\Phi_i(h, e_i)}{\sum_{e'_i \in \mathcal{E}_1} \Phi_i(h, e'_i)} = \prod_{i=1}^{m} p(e_i \mid h) \end{aligned} \tag{5.65}$$

Combining the results given in eq. 5.63 and eq. 5.65, we get the desired results.

$$\begin{aligned} woe(h/h' : e_1, \ldots, e_m) &= \log \frac{p(e_1, \ldots, e_m \mid h)}{p(e_1, \ldots, e_m \mid h')} = \log \frac{\prod_{i=1}^{m} p(e_i \mid h)}{\prod_{i=1}^{m} p(e_i \mid h')} = \log \prod_{i=1}^{m} \frac{p(e_i \mid h)}{p(e_i \mid h')} \\ &= \sum_{i=1}^{m} \log \frac{p(e_i \mid h)}{p(e_i \mid h')} = \sum_{i=1}^{m} woe(h/h' : e_i) \quad \square \end{aligned}$$

5.6.3 Application of the MAXENT-WOE Theorem

The principal motivation for the development of this theorem is to support the development of the models that will be studied in the next chapter. In what follows, we will rely on the fact that if all constraints are on the weights of evidence of individual sources, then the weights of evidence are additive. Formally stated,

Corollary 5.1.1: *If a probability distribution over the event space $\Omega = \mathcal{H} \times \mathcal{E}_1 \times \ldots \times \mathcal{E}_m$, is constrained such that all constraints are of the form*

$$woe(h/h' : e_i) = \xi_{ij} \qquad j = 1, \ldots, \chi_i^W$$

for $i = 1, \ldots, m$, then for the maximum entropy distribution:

$$woe(h/h' : e_1, \ldots, e_m) = \sum_{i=1}^{m} woe(h/h' : e_i) \tag{5.66}$$

proof: This follows from the theorem because the weight-of-evidence constraints can be viewed as conditional constraints, in the sense of eq. 5.56 since:

$$\begin{aligned} woe(h/h' : e_i) = \xi_{ij} \quad & \textit{iff} \quad \frac{p(e_i \mid h)}{p(e_i \mid h')} = \xi_{ij} \\ & \textit{iff} \quad p(e_i \mid h) - \xi_{ij} \cdot p(e_i \mid h') = 0 \\ & \textit{iff} \quad 1 \cdot E[I_{[E_i(\omega)=e_i]}(\omega)|h] - \xi_{ij} \cdot E[I_{[E_i(\omega)=e_i]}(\omega)|h'] = 0 \end{aligned} \tag{5.67}$$

$$\text{where:} \quad I_{[\psi]}(\omega) = \begin{cases} 1 & \text{if } \psi \text{ is true of omega} \\ 0 & \text{otherwise} \end{cases}$$
$$\text{is the indicator function}$$

Therefore, eq. 5.67 satisfies the linearity condition imposed on the conditional constraints. □.

Clearly, if there are only 2 hypotheses, h and $\bar{h}$, then eq. 5.66 reduces to:

$$woe(h : e_1, \ldots, e_m) = \sum_{i=1}^{m} woe(h : e_i) \tag{5.68}$$

It is interesting to point out that the general constraints allowed for in the MAXENT-WOE Theorem cover the constraints used in the BIM-MAXENT and CM-MAXENT models. Specifically, the constraints on the conditional probabilities of occurrence,

$$\begin{aligned} p(occ_i \mid rel) &= \xi_i, \\ p(occ_i \mid \overline{rel}) &= \bar{\xi}_i \end{aligned}$$

are consistent with the restrictions given for conditional constraints. Constraining the probability of relevance

$$p(rel) = \rho$$

is a hypothesis constraint. And, the

$$E(X_{\#}|R = 1) = \zeta \qquad \text{where: } X_{\#} = \sum_{i-1}^{s} X_i$$

constraint used in the CM-MAXENT model (eq. 5.36) is also a conditional constraint, since $X_{\#}$ is a linear combination of functions of the individual sources of evidence.

5.7 Summary

We have seen in this chapter how the BIM-MAXENT and CM-MAXENT models can be derived from the Maximum Entropy Principle with suitable constraints. In Section 5.6, the reasoning used with respect to these two models was generalized. The generalization was supported by conceptualizing the problem explicitly in terms of the concept of weight of evidence as formulated in Chapter 3. The result is Theorem 5.1.

Corollary 5.1.1 of this theorem will play an important role in the following chapter. The ranking formula developed there will include a component corresponding to the evidence given by the total set of *tf-idf* values appearing in the document. The formal results that have been derived here provide justification for calculating this contribution to the ranking value as a sum of the contributions due to the *tf-idf* values for individual terms appearing in the document.

CHAPTER 6

PROBABILISTIC MODELING OF MULTIPLE SOURCES OF EVIDENCE

> *. . . we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.*
>
> Claude Shannon in *Coding Theorems for a Discrete Source with a Fidelity Criterion*, [98, p. 126].

This chapter 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 ultimate 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.

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 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 [52] – will serve as the vehicle for exposition of the principal techniques involved.

The following sections will show how query/document features can be studied, how a model in terms of this evidence can be formulated, and how 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.

Results will be presented suggesting that the modeling framework, and more important the general approach to the analysis of evidence, developed in this study may lead to a ranking formula that performs as well as state-of-the-art retrieval formulas that have evolved over the years.

6.1 Data Preparation

In this section we review the query/document features that will be considered; the format of the data to be analyzed; and how the queries were prepared.

6.1.1 Query/Document Characteristics

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,

$$-\log \frac{df}{N}$$

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.

document length: The number of words the document contains. Although document length is considered, as will be seen, it will not play a role in the final ranking formula.

query	doc.	term	coord	idf	tf	rel
151	10383	Clinton	2	2.3	1	0
151	10383	proceedings	2	3.1	3	0
151	10674	proceedings	1	3.1	2	0
151	10992	impeachment	1	3.5	1	0
151	11005	Clinton	1	2.3	1	0
151	11013	Clinton	3	2.3	2	1
151	11013	impeachment	3	3.5	7	1
151	11013	proceedings	3	3.1	3	1
151	11089	proceedings	1	3.1	1	0
...						
152	10046	dairy	2	4.2	3	1
152	10046	industry	2	1.9	2	1
152	10572	industry	1	1.9	1	0
...						

Table 6.1. Format of Retrieval Data

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.

6.1.2 The Data to be Analyzed

Table 6.1 shows the format of the data, the analysis of which is discussed in this chapter[1]. There are entries in this table only for query/document pairs for which at least one of the query terms appear in the document. IR systems generally do not process documents for which no query terms are present, so the goal of the analysis will be to build a model conditioned on the occurrence of at least one query term (i.e. a non-zero coordination level).

[1]The terms shown for query #151 are not those of the actual query extracted from TREC topic #151. The query shown corresponds to the example used earlier in the text, and the associated data is fictitious.

For each of these query/document pairs, there is an entry for each of the terms that appear in the document. For each entry, the following fields depend only on the query/document pair:

query: a number that identifies the query;

doc: a number that identifies the document;

coord: coordination level – the number of query terms appearing in the document;

rel: the TREC relevance judgment for this document with respect to the query, where 1 indicates relevance and 0 indicates non-relevance;

and the following two fields that depend as well on the individual term:

idf: inverse document frequency;

tf: term frequency.

6.1.3 Query Preparation

In this chapter, 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 run against the AP88 collection is used to exemplify the process. Queries were taken from the titles of the 50 TREC topics. In order to convert the TREC title field to a query, stopwords were removed, duplicates were eliminated, and words were stemmed, as was done for the testing described in Chapter 4.

6.2 Overview of the Modeling Strategy

The objective of the analysis is to develop a model for the weight of evidence in favor of relevance given by the query/document features under consideration:

$$woe(rel : e_1, \ldots, e_n \mid q, *) \quad = \quad \log \frac{O(rel|e_1, \ldots, e_n, q, *)}{O(rel|q, *)} \tag{6.1}$$

where the weight of evidence is conditioned on the query being evaluated. To be more precise, the weight of evidence that will be modeled is restricted as well to the subspace corresponding to those query/document pairs for which at least one of the query terms appears in the document. This is indicated by the $*$ in the condition of the weight of evidence given in eq. 6.1. In general, IR systems do not evaluate documents that do not include at least one of the query terms. For that reason, all probabilities and weights of evidence considered in this chapter will be conditioned on the occurrence of at least one term. We will explicitly include the $*$ in the formulas appearing in this section, but for the sake of reducing clutter in the notation, it will be left implicit in the sections that follow.

6.2.1 Four Models

The data analysis will result in the development of four models, which will be denoted by: $\mathbf{M_0}$, $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$. The focus of the modeling effort for each of these models will be:

$$\mathbf{M_0}: \quad \log O(rel|Qry = q, *) \tag{6.2}$$

$$\mathbf{M_1}: \quad woe(rel : Coord = co \mid Qry = q, *) \tag{6.3}$$

$$\mathbf{M_2}: \quad woe(rel : Idf = idf \mid Coord = co, Qry = q, *) \tag{6.4}$$

$$\mathbf{M_3}: \quad woe(rel : Tf = tf \mid Idf = idf, Coord = co, Qry = q, *) \tag{6.5}$$

with each model extending the previous one, in terms of the constraints imposed on the probability distribution. Although, the direct objective at each of the last three steps is to model weight of evidence conditioned on the query, the log-odds in favor of relevance, and hence the probability of relevance, for the query is modeled as well. This is true because for each of the models concerned:

$$\begin{aligned}
\mathbf{M_1}: \quad & \log O(rel|co,q,*) && = \log O(rel|q,*) + woe(rel:co \mid q,*) \\
\mathbf{M_2}: \quad & \log O(rel|idf,co,q,*) && = \log O(rel|co,q,*) + woe(rel:idf \mid co,q,*) \\
\mathbf{M_3}: \quad & \log O(rel|tf,idf,co,q,*) && = \log O(rel|idf,co,q,*) + woe(rel:tf \mid idf,co,q,*)
\end{aligned}$$

In each case, the conditional probability can be derived from the conditional log-odds by:

$$p(rel \mid \ldots) = \frac{e^{\alpha}}{1-e^{\alpha}} \quad \text{where } \alpha = \log O(rel|\ldots)$$

To begin the process, model $\mathbf{M_0}$ is developed by simply estimating, for each query, the probability that an arbitrary document will be found to be relevant to that query. This is described in more detail in Section 6.3. We proceed, in Section 6.4, to analyze evidence corresponding to coordination level. This results in the $\mathbf{M_1}$ model of relevance conditioned on the query being evaluated and the number of query terms occurring in the document. In Section 6.5, we see that inverse document frequency is correlated with residual log-odds of relevance, relative to the $\mathbf{M_1}$ model. Extension of the model to include idf_i for each of the query terms, $i = 1, 2, \ldots$, produces the $\mathbf{M_2}$ model. Finally, analysis of the role of term frequency, discussed in Section 6.6, results in the $\mathbf{M_3}$ model. It is this model on which a ranking formula will be based.

6.3 Base Model

A modeling assumption, derived from the Maximum Entropy Principle, is that the weight of the query/document evidence, $\vec{\mathbf{e}}$, shall be considered independent of the query, q. The ranking status value to be used will then be this weight of evidence, which can be assigned without knowledge of the prior probability of relevance for the query, $p(rel|q)$. However, as discussed in Section 6.3.1, modeling is best accomplished by including the probability of relevance conditioned only on the query, q, in order to eliminate potential problems due to confounding.

The basis for the analysis is established by first developing a model of the probability of relevance conditioned only on the query being evaluated, $p(rel \mid q)$. We shall loosely refer to the probability as the *prior* probability of relevance for query q in the sense that it is the probability that a randomly selected document will be found relevant to the query before any evidence is observed, that is before the contents of the document are known.

6.3.1 Confounding and the *Prior* Probability of Relevance

An important aspect of the modeling effort is that at each stage, $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$, it is the log-odds and weight of evidence *conditioned on the query* that is modeled. This can be done effectively since the prior probability of relevance can be estimated from the available data (resulting in the $\mathbf{M_0}$ model). However, since the prior probability is not expected to be available at *retrieval time*, the derived ranking formula will not depend on the query specific prior probability. Formally, this is possible because, in the model on which the ranking formula is based, the weight of evidence conditioned on the query is independent of the query itself. This being the case, one might ask why the modeling procedure involves the query specific priors when they are only to be factored out in the resulting ranking formula. The answer has to do with the issue of *confounding*.

In the words of Sahai and Khurshid, "Confounding exists when the association between two variables is altered after accounting or controlling for the effect of a third variable" [91, p. 55]. Suppose, for the sake of example, that the conditional weight of evidence is a linear function of coordination level:

$$woe(rel : Coord = co \mid Qry = q) = \beta_0^{\text{CO}} + \beta_1^{\text{CO}} \cdot co$$

and that coordination level is the only factor involved so that empirical data reflect the linearity relation perfectly; that is, the log-odds of relevance can be measured

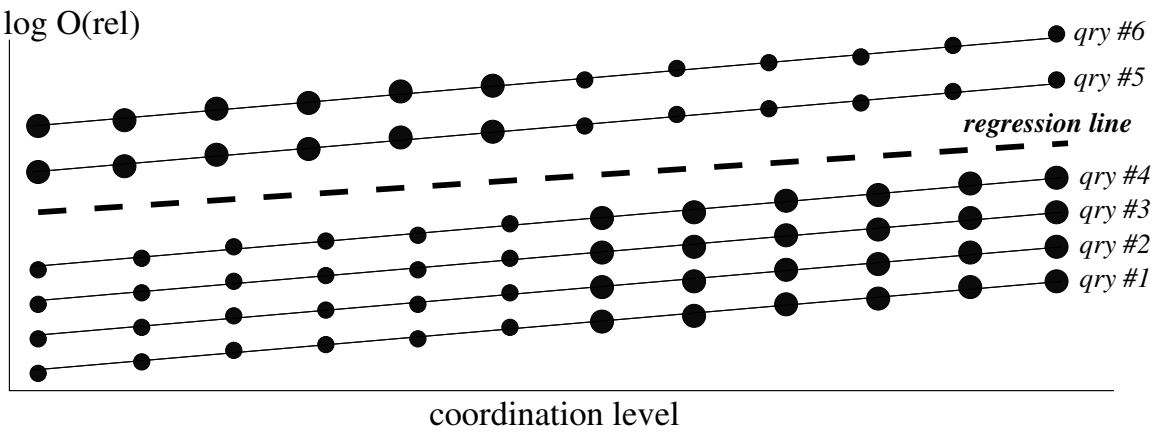

Figure 6.1. Data for six queries resulting in confounding

without any *noise*. Then, we would have:

$$\begin{aligned} \log O(rel|co,q) &= \log O(rel|q) + woe(rel:co \mid q) \\ &= \beta_q^{\mathrm{Q}} + \beta_0^{\mathrm{CO}} + \beta_1^{\mathrm{CO}} \cdot co \end{aligned}$$

where β_q^{Q} is the prior log-odds of relevance, $\log O(rel|q)$, for query, q. A scatterplot of $\log O(rel|co)$ vs. co for six queries would look something like that shown in Figure 6.1. For each query, the log-odds of relevance is also a linear function of coordination level; the slope, β_1^{CO}, is constant across queries, resulting in parallel lines; and the intercept values, $\beta_q^{\mathrm{Q}} + \beta_0^{\mathrm{CO}}$, vary across queries, resulting in a different line for each query.

Suppose further that there is a tendency for most of the data to correspond to higher coordination levels for queries with lower β_q^{Q} (queries 1-4 in the figure), and for most of the data to correspond to lower values of co for queries with higher β_q^{Q} (queries 5 & 6). This is suggested in Figure 6.1 by the size of the circles, with larger circles intended to indicate greater quantities of data for the given point.

If the modeling process does not take into consideration the difference in prior log-odds from one query to another, data will be analyzed independent of the query from whence they came. The result will be that the log-odds for low values of the

query	# rels	# docs	$\mathbf{p(rel \mid q)}$	$\log \mathbf{O(rel \mid q)}$
151	7	4488	0.0015	-2.80
152	2	2646	0.0007	-3.12
153	1	523	0.0019	-2.71
...				
173	11	514	0.0214	-1.66
...				

Table 6.2. Reduction of Data for Query Analysis

predictor value will be overestimated relative to the estimate of the log-odds at higher values of the predictor variable. The estimate of the slope, $\hat{\beta}_1^{\text{CO}}$, will then be lower than it should be, $\hat{\beta}_1^{\text{CO}} < \beta_1^{\text{CO}}$, as is shown by the dashed line in the figure. This underestimate is due to the relation existing between the *distribution* of the *co* values and prior log-odds. The relation between log-odds of relevance and coordination level is said to be *confounded* [58].

In order to avoid the problems of confounding, the prior odds of relevance are included in models $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$. They are then factored out in order to model the weight of evidence conditioned on the query, and from that produce a ranking function.

6.3.2 Estimating the Prior Probability of Relevance

The estimation of the probability of relevance conditioned on the query (and the occurrence of at least one query term) is straightforward. The data shown in Table 6.2 are grouped by the *query id*, a unique number identifying the TREC topic from which the query has been taken. The total number of documents and the total number of those documents that were judged relevant are counted, as shown at the left of Table 6.2. From these counts the probability of relevance, $p(rel|q)$, can be estimated as,

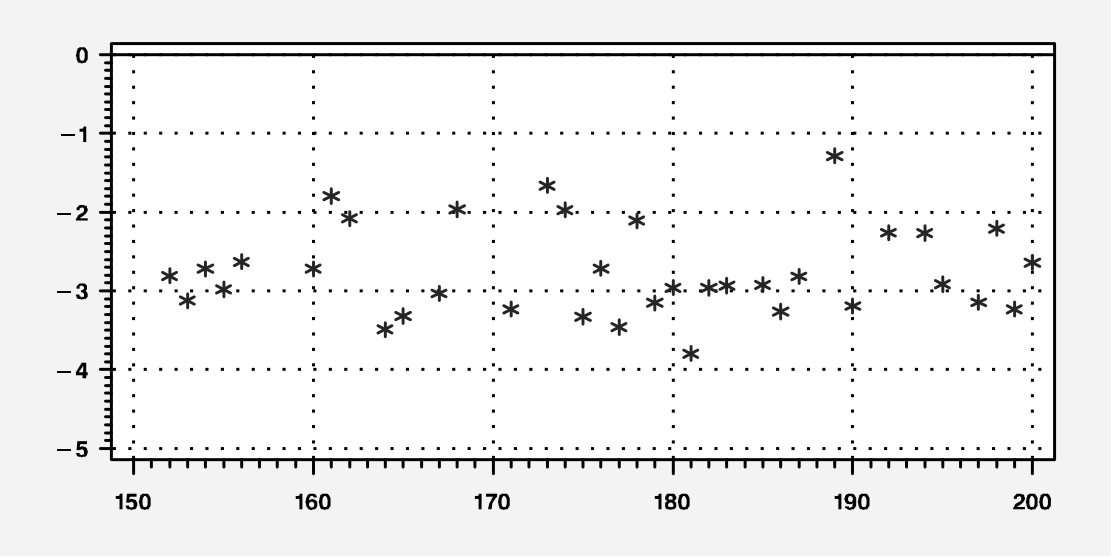

Figure 6.2. (Prior) log-odds of relevance for TREC-3 queries

$$\hat{p}(rel \mid q) = \#rels \ / \ \#docs$$

and this can be converted to log-odds as shown in the rightmost two columns of the table.

A graph of these data is shown in Figure 6.2. It can be seen in this graph that the prior probability of relevance ranges over almost three orders of magnitude, from a little less than one in ten (log-odds ≈ -1) for query #189, to slightly above one in ten thousand (log-odds ≈ -4), for query #181.

6.4 Modeling Coordination Level

Using $\mathbf{M_0}$ as a base, we model the weight of evidence offered by coordination level.

6.4.1 Calculation of Residual Log-odds

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

query	doc #	coord	rel	$\mathbf{p_0(rel|q)}$	docs
151	10674	1	0	.0027	1
151	10992	1	0	.0027	1
151	11005	1	0	.0027	1
151	11089	1	0	.0027	1
151	11203	1	1	.0027	1
...					
152	10572	1	0	.0004	1
152	10734	1	1	.0004	1
...					
			# rels	E_0(#rels)	# docs

Table 6.3. Reduction of Data for Coordination Level Analysis

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

$$\hat{r}_i = \sum_{Qry=q} n_{i,q} \cdot \hat{p}(rel|Qry = q) \tag{6.6}$$

where $n_{i,q}$ is the number of documents that contain i terms from query q. The probability, $\hat{p}(rel|Qry = q)$, is the conditional probability of relevance given by model, $\mathbf{M_0}$, that was 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. 6.6 can also be expressed as:

$$\hat{r}_i = \sum_{(q,d):Coord(q,d)=i} \hat{p}(rel|Qry = q) \tag{6.7}$$

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.

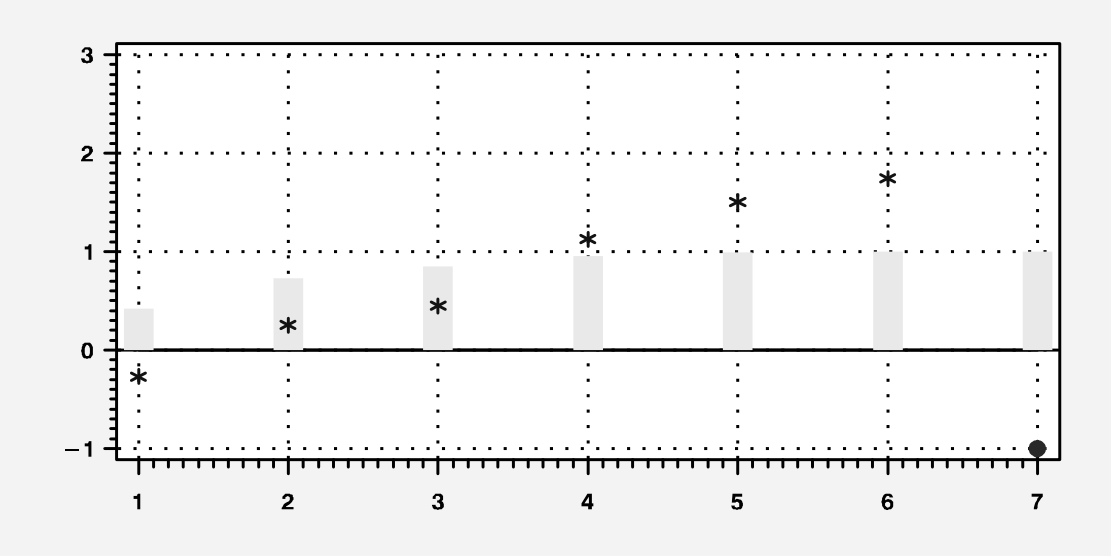

Figure 6.3. Residual log-odds as function of coordination level: unsmoothed

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:

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

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 6.3 shows the scatterplot 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.

6.4.2 Fitting a Regression Line

Motivated by the linearity suggested by Figure 6.3, a weighted linear regression was performed. The point at $Coord = 7$ represented by the small circle in Figure 6.3 has been ignored for the purposes of the regression. Since the point corresponds to very few documents, its effect on the overall regression would be negligible. When we come to the modeling of the *idf* variable, a more general approach to the processing of these infinite estimates will be described.

The linear regression was weighted because the variance associated with the points on the graph is not constant. There are two factors contributing to the differing variance, or *heteroskedasticity* [81, p. 170], both of which must be taken into consideration. The first is that each point on the graph corresponds to a different number of data points, as indicated by the histogram in the background. The greater the number of documents corresponding to a point on the graph, the smaller the variance will be. Second, even were there to be an equal number of documents entering into the calculation of each of the points on the graph, the variance for each point would be different due to the differences in the *true* log-odds of relevance at different values of the predictor variable. In order to compensate for the disparity in the variance as a function of the predictor, we use a *weighted* linear regression.

A weighted linear regression produces a fit that minimizes the weighted sum of squared residuals. Each point is weighted by the inverse of the variance of the response variable at that point [81, p. 170]. The greater the variance, the smaller the weight. For the problem here, the response variable is[2]:

[2]Natural log is used here to simplify the derivation. The change of scale of the log only affects the final calculation by a multiplicative constant, which will have no affect on the weights relative to one another.

$$\log \frac{r_i/(n_i - r_i)}{\hat{r}_i/(n_i - \hat{r}_i)} = \log \frac{r_i}{n_i - r_i} - \log \frac{\hat{r}_i}{n_i - \hat{r}_i}$$

where r_i is the observed number of relevant documents, $\hat{r}_i$ is the expected number of relevant documents, and n_i is the total number of documents for the i^{th} point. Assuming that the variability of the second term involving the expected value is small relative to the variability of the first, and using p_i for the observed proportion of documents, r_i/n_i, that are relevant, we can write:

$$\text{Var}\left[\log \frac{r_i/(n_i - r_i)}{\hat{r}_i/(n_i - \hat{r}_i)}\right] \approx \text{Var}\left[\log \frac{r_i}{n_i - r_i}\right] = \text{Var}\left[\log(\frac{p_i}{1 - p_i})\right]$$

Letting p_i^0 be the (unknown) true probability of relevance, and using a first order Taylor series expansion [30, Section 8.8] about p_i^0 to approximate $\log \frac{p_i}{1-p_i}$, we have:

$$\begin{aligned}
\log(\frac{p_i}{1 - p_i}) &\approx \log(\frac{p_i^0}{1 - p_i^0}) + (p_i - p_i^0)\left[\frac{d}{dp}\log(\frac{p_i}{1 - p_i})\right]_{p_i^0} \\
&= \log(\frac{p_i^0}{1 - p_i^0}) + (p_i - p_i^0)\left[\frac{-1}{p_i(1 - p_i)}\right]_{p_i^0} \\
&= \log(\frac{p_i^0}{1 - p_i^0}) + (p_i - p_i^0)\left(-\frac{1}{p_i^0(1 - p_i^0)}\right)
\end{aligned}$$

Using this approximation to calculate the variance, gives us:

$$\begin{aligned}
\text{Var}\left[\log(\frac{p_i}{1 - p_i})\right] &= \text{Var}\left[\log(\frac{p_i^0}{1 - p_i^0}) + (p_i - p_i^0)\left(-\frac{1}{p_i^0(1 - p_i^0)}\right)\right] \\
&= \left[\frac{1}{p_i^0(1 - p_i^0)}\right]^2 \cdot \text{Var}\left[p_i\right] \qquad (6.8) \\
&= \left[\frac{1}{p_i^0(1 - p_i^0)}\right]^2 \cdot \frac{p_i^0(1 - p_i^0)}{n_i} \qquad (6.9)
\end{aligned}$$

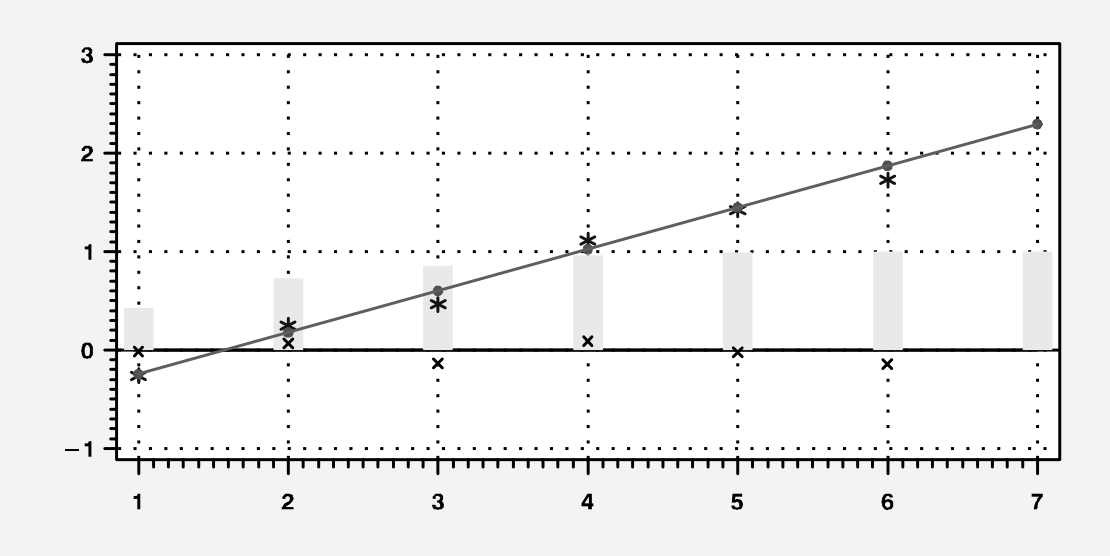

Figure 6.4. Residual log-odds as function of coordination level: smoothed with regression

$$= \frac{1}{n_i \cdot p_i^0 \cdot (1 - p_i^0)}$$

Eq. 6.8 follows since p_i^0, and hence $\log(\frac{p_i^0}{1-p_i^0})$ and also $-\frac{1}{p_i^0(1-p_i^0)}$, are constants; and eq. 6.9 follows because the variance of the proportion, p_i, of *successes* on n_i independent Bernoulli trials with probability of success, p_i^0, on each trial, is given by $\frac{p_i^0(1-p_i^0)}{n_i}$.

The true probability of relevance, p_0, is unknown, but we can estimate it by the observed proportion of relevant documents, $\hat{p}_i = r_i/n_i$, which gives as an estimate of the variance:

$$\frac{1}{n_i \hat{p}_i (1 - \hat{p}_i)} \tag{6.10}$$

The inverse of this variance, $n_i \hat{p}_i (1 - \hat{p}_i)$, can be used as the weight of the i^{th} point for the linear regression. The result of the weighted regression is shown in Figure 6.4, which is equivalent to the graph of Figure 6.3, with the regression line

$$res = \beta_0^{\mathrm{CO}} + \beta_1^{\mathrm{CO}} \cdot co \tag{6.11}$$

query	doc #	term	coord	rel	$\mathbf{p_1(rel|q, coord)}$	docs
151	10383	Clinton	2	0	.0039	0.50
151	11005	Clinton	1	0	.0016	1.00
151	11013	Clinton	3	0.33	.0051	0.33
...						
				# rels	E_1(#rels)	# docs

Table 6.4. Reduction of Data for *idf* Analysis

overlaid.

6.4.3 Producing the $\mathbf{M_1}$ Model

This regression line can be used to form a model, $\mathbf{M_1}$, 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, $\mathbf{M_1}$, 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:

$$\begin{aligned} woe(rel : co \mid q) &= \log \frac{O(rel|co, q)}{O(rel|q)} \\ &= \log O(rel|co, q) - \log O(rel|q) \end{aligned} \tag{6.12}$$

This $\mathbf{M_1}$ model advances the development 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.

6.5 Modeling Inverse Document Frequency

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 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 weighted for the purposes of model fitting.

6.5.1 Weighting of Data Points

In order that each document, and hence each relevance judgment, be treated equally, 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.

6.5.2 Evidence Respecting Specific Query Terms

In order to study the value of a term's inverse document frequency as a source of evidence, the evidence associated with learning that a specific query term, t, was one of the terms that occurred in the document, was studied. The weight of evidence tied to this event was estimated for each of the query terms over all of the queries.

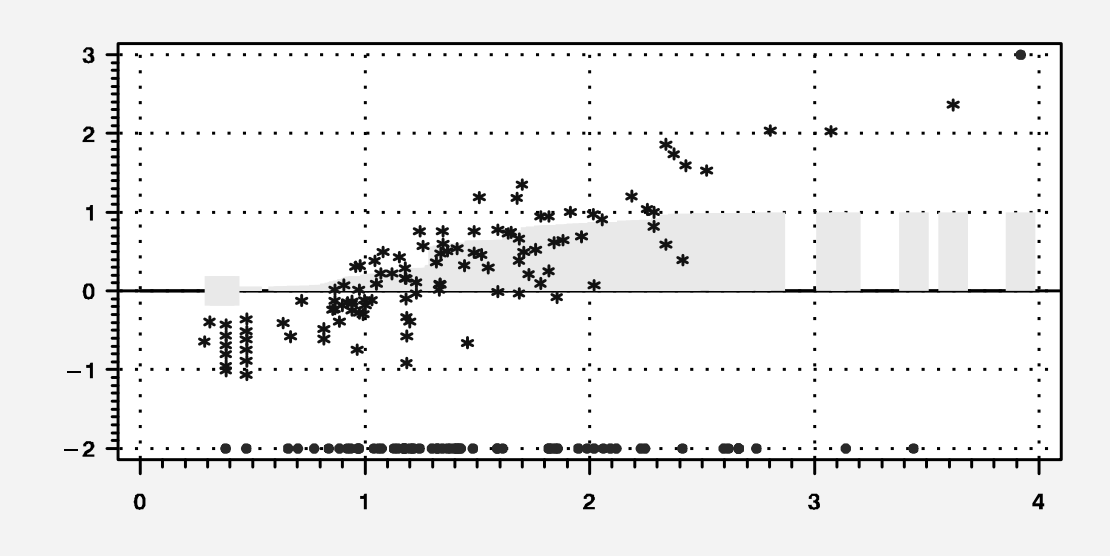

Figure 6.5. Residual log-odds as function of *idf*: unsmoothed

In order to carry out this estimation process, the data were first 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 two or more different queries are, for this purpose, 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. 6.7:

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

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

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

with the second term being the weight of evidence predicted by the $\mathbf{M_1}$ model; i.e. given by the regression line shown in Figure 6.4.

Figure 6.5 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. Also visible is a small circle in the upper right hand corner. This corresponds to a term that only appeared in relevant documents. For this term the estimated probability of relevance is 1, giving infinite odds, and hence infinite log-odds, of relevance. The vertical bars in the background give a cumulative histogram for the (weighted) data points.

6.5.3 Smoothing

There are three reasons for considering the application of smoothing to the data displayed in Figure 6.5. First is the problem of infinite estimates (both positive and negative). A benefit of using scatter plots as part of the exploratory data analysis process is to produce a visual impression of the behavior of the data. The difficulty with Figure 6.5 is that the eye is unable to integrate the information represented by the small circles with the information represented by the stars. By smoothing, estimates based on a single point are replaced by estimates that incorporate information garnered from neighboring points as well. The result is the production of more robust estimates and the integration of visual information that was previously presented separately.

A second reason to consider smoothing of the data pertains to the variance displayed for estimates of the weight of evidence for points corresponding to similar values of the explanatory variable, even when the points corresponding to infinite

estimates are ignored. For example, in Figure 6.5, there are seventeen points plotted at a value of *idf* between 0.9 and 1.0. (Actually some of these correspond to the same word, occurring as a term in different queries). Five of these are infinite. The estimates for the twelve remaining points vary from -0.7 to +0.3. This variance is in evidence in all parts of the graph, contributing to an overall fuzziness in the presentation, making it more difficult to discern an underlying pattern that may be present in the data.

The third reason is related to the first. The points at the right of the graph correspond to rare terms; terms that occur in a small fraction of documents in the collection and hence have a high *idf* value. The estimate for one of these terms is based on a small number of documents, since it is based on the relevance of only those documents in which the term appears. These estimates can be expected to be less robust. The points on the right side of the graph should be given less consideration than estimates corresponding to high frequency terms which are based on a greater number of documents. This, however is difficult for the eye to do without help. By smoothing the data in the manner described below, each point of the resulting graph can be considered to be more or less of equal value in arriving at a sense of underlying patterns existent in the data.

These same problems were encountered earlier in the analysis of *idf* weighting (see Section 4.1). However the solution adopted there is not adequate to our needs in this situation. In Chapter 4 we grouped terms into pseudoterms and weight of evidence was estimated for each pseudoterm. Now, in contrast, we need to estimate *residual* weight of evidence. We cannot simply group the data by *idf* value, because each data point corresponds to a different coordination level. Observed weight of evidence could be calculated as before, but not the difference between observed weight of evidence and the weight of evidence as predicted by $\mathbf{M_1}$, which is what is needed. To produce a smoothed plot, an alternative technique was developed.

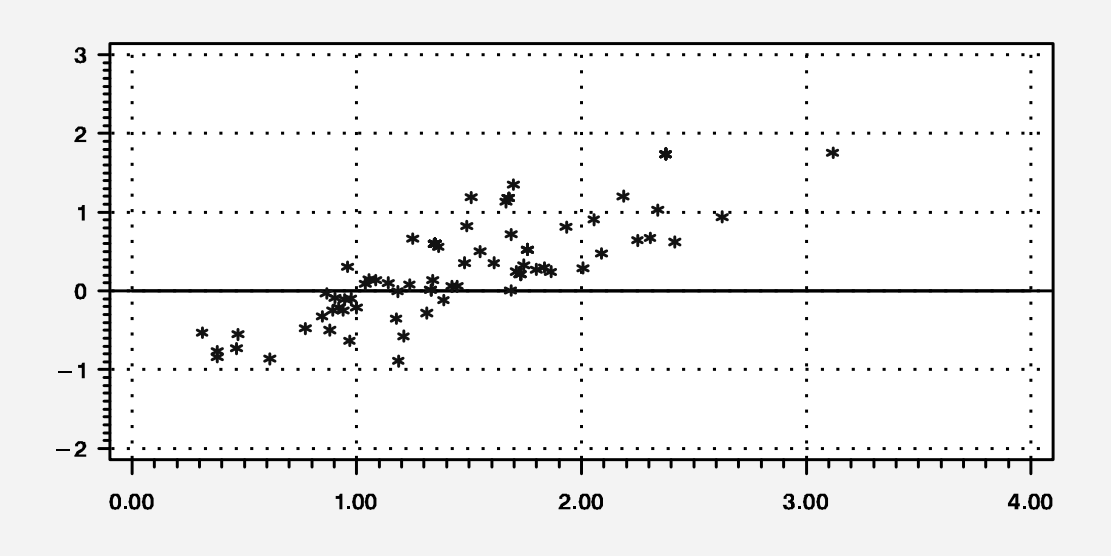

Figure 6.6. Residual log-odds as function of *idf*: smoothed

Figure 6.6 shows a smoothed version of the data displayed in Figure 6.5. To produce the smoothed plot, a number of *bins*, b, is fixed, with $b = 50$ for this graph. The points of Figure 6.5 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.

Finally, it should be mentioned that the kernel regression approach discussed in Chapter 4 can be readily adapted to the needs of residual analysis. Weighted averaging using (Gaussian) kernel functions can simply replace the division into bins. This approach was contemplated, but discarded in favor of binning for two reasons. The principal consideration was pragmatic. The SAS statistical analysis package was used to support this phase of the research. The kernel regression approach is computationally intensive and would not be practical if implemented in the SAS environment.

The second consideration involved the question of robustness of estimation. As discussed above, the binning approach responds to this issue by varying bin sizes based on the number of relevant documents included in the bin. In order to address the problem of robustness with the kernel regression approach, a *variable bandwidth* could be used. The bandwidth is essentially the *width* or *dispersion* of the function — for example, the variance of the distribution in the case of the Gaussian. A variable bandwidth means the use of a different bandwidth for different points. For our purposes, the bandwidth could be based on the proportion of relevant documents in the neighborhood of the point being estimated. Unfortunately, this would serve to exacerbate the problem of computational complexity.

6.5.4 Fitting a Three-piece Linear Function

As discussed in Chapter 4, the study of the weight of evidence provided by term *idf* values, suggests 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

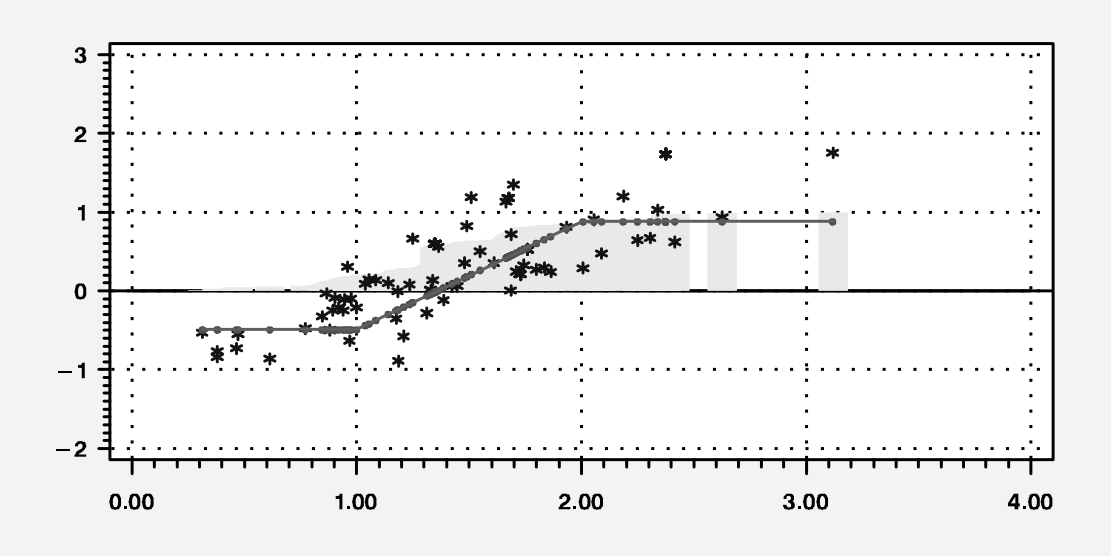

Figure 6.7. Residual log-odds as function of *idf*: smoothed with regression

shown in Figure 6.6, 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:

$$res = \beta_0^{\text{IDF}} + \beta_1^{\text{IDF}} \cdot \widetilde{idf}_1 \tag{6.13}$$

$$\text{where } \widetilde{idf}_1 = \begin{cases} 0 & \text{if } idf < 1 \\ idf - 1 & \text{if } 1 \leq idf \leq 2 \\ 1 & \text{if } idf > 2 \end{cases} \tag{6.14}$$

The resulting estimates for the parameters, β_0^{IDF} and β_1^{IDF}, yield the model, $\mathbf{M_2}$, 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-

query	coord	idf	tf	rel	$\mathbf{p_2(rel|q, coord, idf)}$	docs
151	2	2.3	1	0	.0041	0.50
151	1	3.5	1	0	.0021	1.00
151	1	2.3	1	0	.0018	1.00
151	1	3.1	1	0	.0023	1.00
...						
152	1	1.9	1	0	.0030	1.00
...						
				# rels	E_2(#rels)	# docs

Table 6.5. Reduction of Data for *tf* Analysis

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 6.7 shows the resulting model imposed on the scatterplot of smoothed residual values.

6.6 Modeling Term Frequency

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 $TF_1, TF_2, \ldots$, 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, $\mathbf{M_2}$:

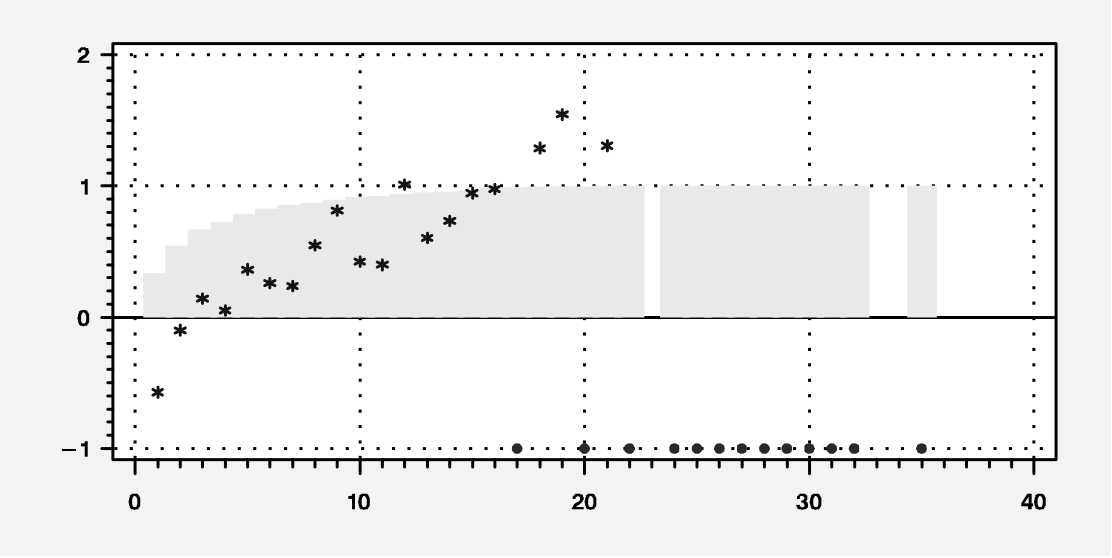

Figure 6.8. Residual log-odds as function of *tf*: unsmoothed

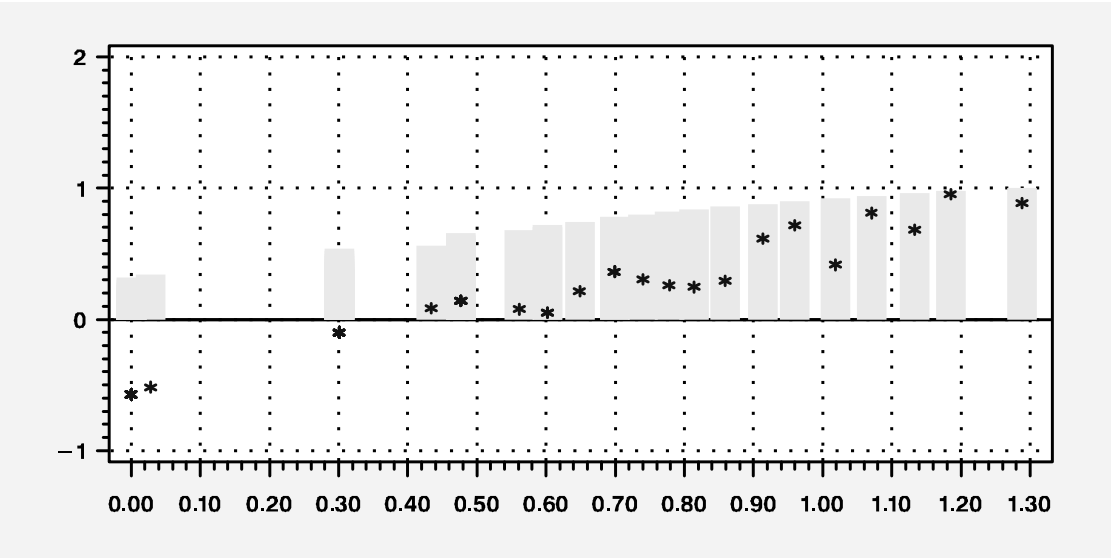

Figure 6.9. Residual log-odds as function of *tf*: smoothed with curve fit

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

A scatterplot of the resulting residuals is shown in Figure 6.8. A number of transformations of the variables involved were tried. Figure 6.9 shows a plot of

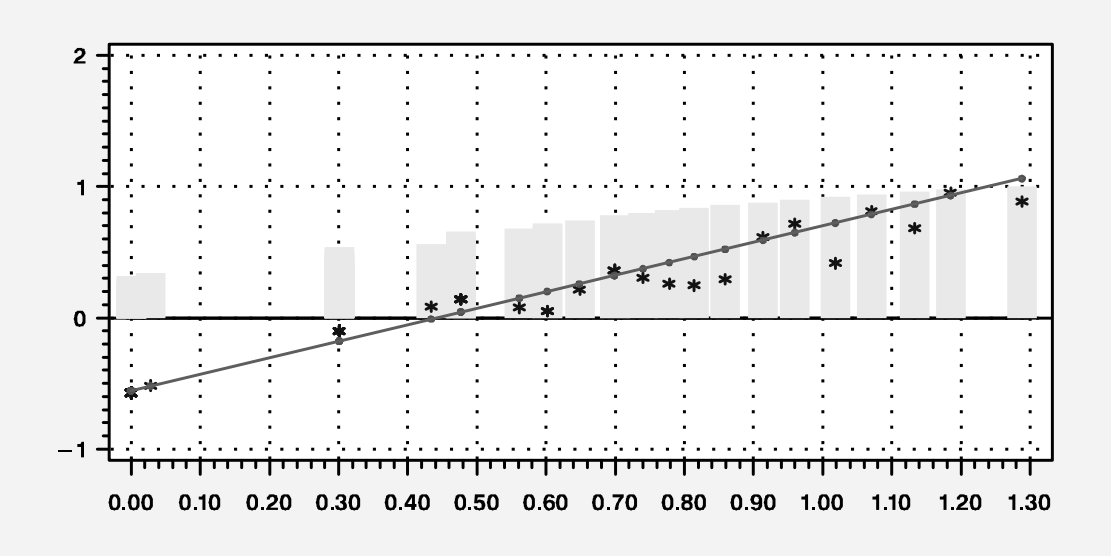

Figure 6.10. Residual log-odds as function of log(*tf*): smoothed with regression

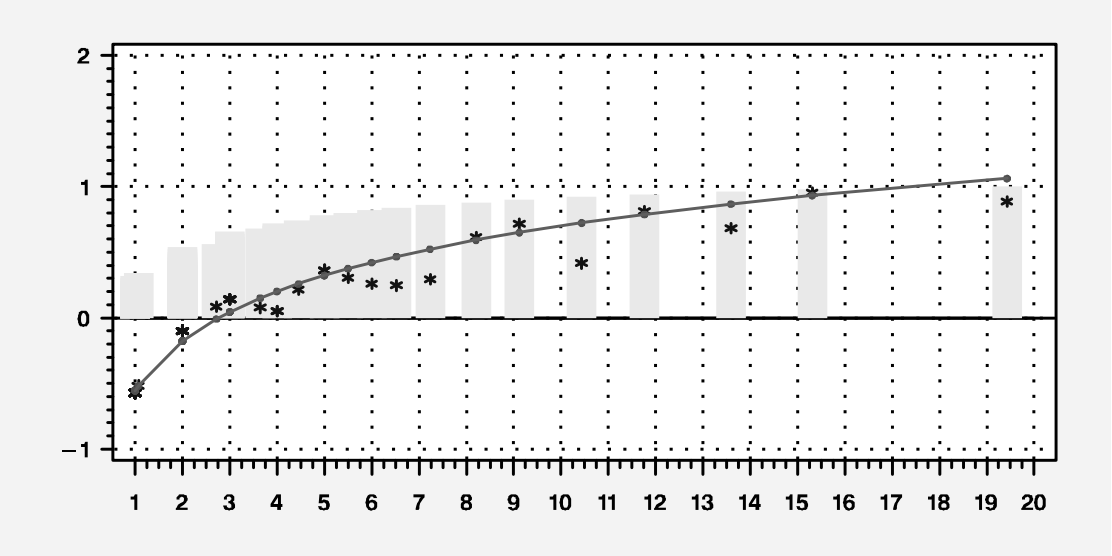

Figure 6.11. Residual log-odds as function of *tf*: smoothed with regression on original scale

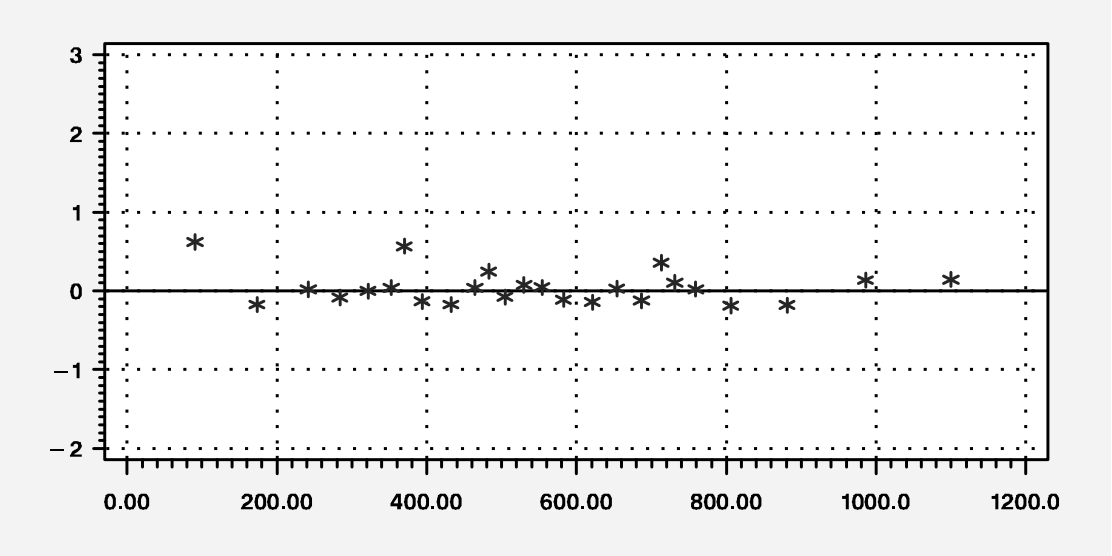

Figure 6.12. Residual log-odds as function of document length

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 regression results in a line given by the equation:

$$res = \beta_0^{\mathrm{TF}} + \beta_1^{\mathrm{TF}} \cdot \log(tf) \tag{6.15}$$

This line is overlaid on the smoothed scatterplot of Figure 6.10. The fit of the curve to the smoothed data on the more natural, unlogged *tf* scale can be seen in Figure 6.11.

6.7 Modeling Document Length

Many modern retrieval systems normalize term frequencies by document length in some way. The intuition is that a term is more likely to occur a larger number of times, the longer a document is. In the vector space model, term frequency is typically normalized by use of the cosine rule [112]. Here the score is normalized by

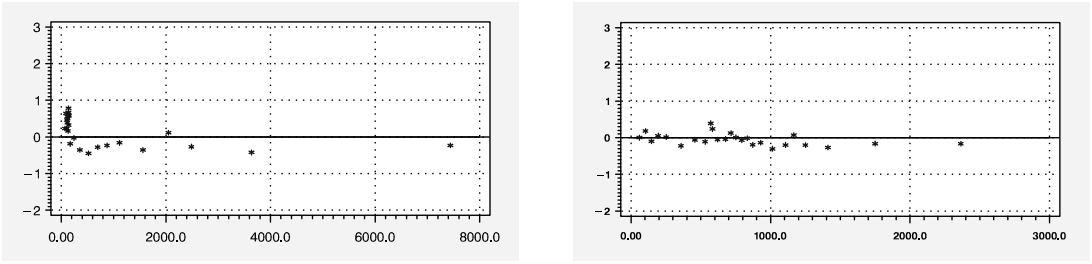

Figure 6.13. Residual log-odds as function of document length for ZIFF2/TREC3 and WSJ89/TREC1 datasets

the Euclidean length of the document vector, which will tend to be greater for longer documents. In INQUERY, the term formula component of the ranking formula is:

$$\frac{tf}{tf + 0.5 + 1.5 \times \frac{dl}{avg\_dl}} \tag{6.16}$$

where dl is document length and avg_dl is the average document length over the entire collection, and also incorporates a form of normalization based on document length.

In order to consider the role of document length in this work, two types of analysis were performed. First, a residual plot was produced with document length as predictor, as it was for the other variables. The data were grouped by document length, and for each group: an expected number of relevant documents was computed, based on $\mathbf{M_3}$; the actual number of relevants were counted; and from these a residual log-odds was calculated. A plot for 50 bins is shown in Figure 6.12. From the plot it does not appear that there is any predictive value associated with document length. Figure 6.13 shows graphs for two other data sets. For the WSJ89 collection with the TREC1 queries, little if any effect due to document length is observed, as was the case with the AP88 data shown in Figure 6.12. Some effect can be seen for the ZIFF2 collection with the TREC3 queries, but again the effect is relatively small.

A different way of viewing the problem is displayed in Figure 6.14. Each of the three graphs shows the (smoothed) residual weight of evidence as a function of term

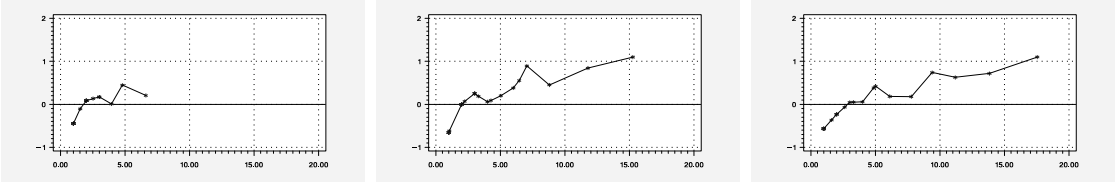

Figure 6.14. Residual plot for term-frequency for three ranges of document length

frequency. Each was produced in exactly the same way as the graph in Figure 6.8 of Section 6.6, except that only a fraction of the data were used. In each case, only data corresponding to a specified range of document lengths entered into the calculations. The ranges considered for Figure 6.8 were document lengths: below 300; between 300 and 500; and above 500 words. These ranges encompass approximately one third of the data entries each. Although fragmentation of the data is obviously reducing the robustness of the estimates, as compared to the estimates produced when all of the data is used, we can say that document length does not appear to have a significant affect on the weight of evidence provided by the term frequency variable.

Based on the above analyses, it was decided not to include document length as a source of evidence in the final model.

6.8 Discussion

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

6.8.1 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 6.15 shows plots of $\hat{woe}(rel : co \mid q)$ as a function of coordination level, co, for all of the TREC 3 queries

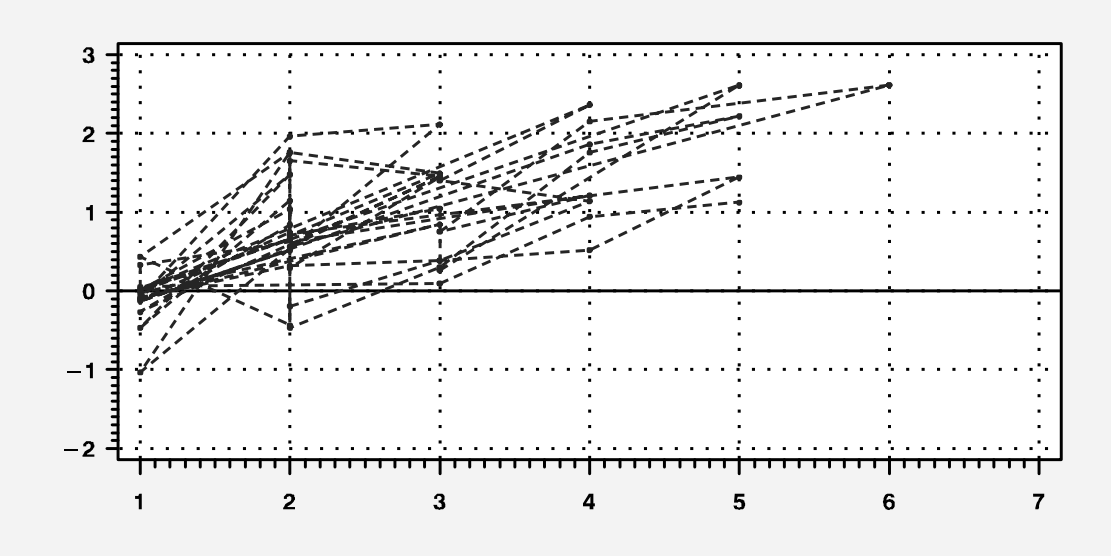

Figure 6.15. Residual log-odds as function of coordination level for individual queries

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.

6.8.2 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 6.5 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 6.6, 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.

6.8.3 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 6.8 strongly suggests a sharp decrease in the weight of evidence provided by one additional occurrence of a query term as term frequency 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:

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

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

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

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 6.16. suggest that the difference between $Tf = 1$ and $Tf = 2$ should not be treated as equal to the difference between

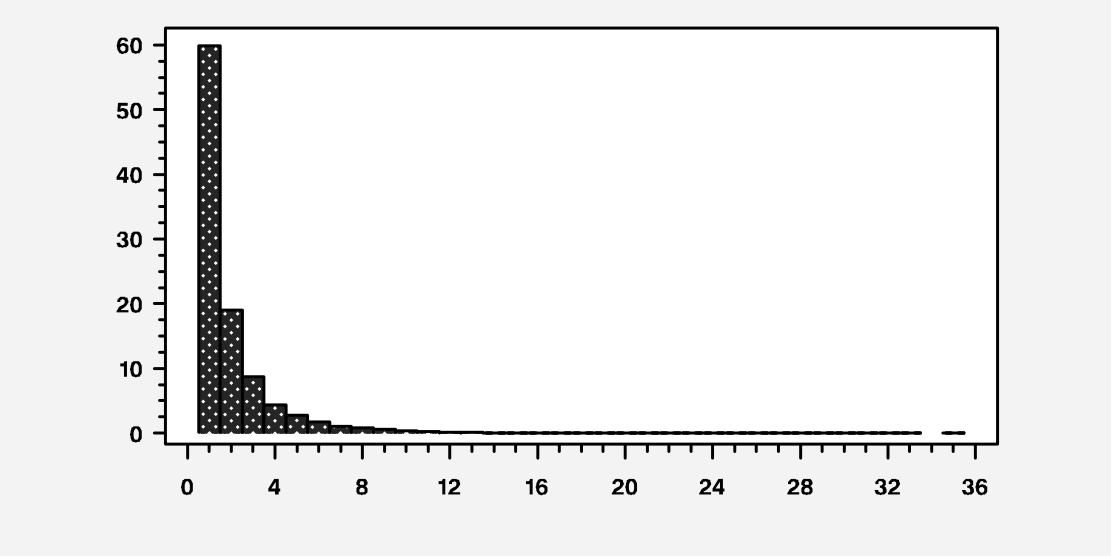

Figure 6.16. Histogram of *tf* values

$Tf = 21$ 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.

6.8.4 Document Length As Evidence

We have concluded that there is no convincing indication that document length should be included in our model. This goes against much evidence to the contrary in the experimentation literature. It should be kept in mind, however, that attention has been restricted to news articles. Two points can be made.

First, there may not be any effect due to document length when attention is restricted to a relatively homogeneous collection of documents such as a collection of news articles. The benefits of document length normalization in an environment such as TREC may be due to differences in behavior across the variety of sub-collections.

Second, the distribution of document lengths over a sub-collection is much more uniform than it is over an entire TREC collection. It is reasonable to conjecture that even if there is an effect due to document length, it may be too small to be detected,

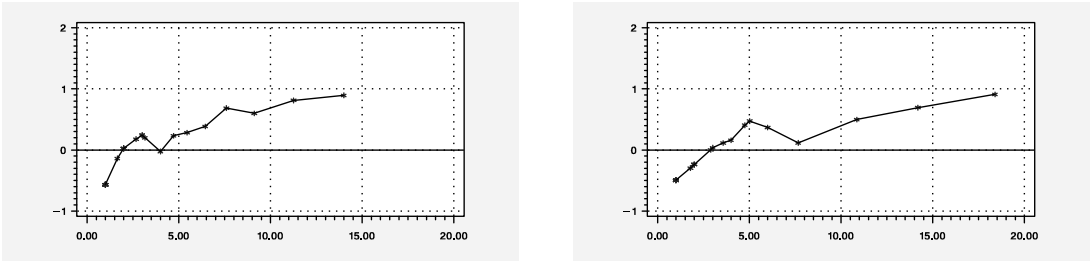

Figure 6.17. Residual plot for term-frequency for two ranges of document length over a combination of AP88 and FR88 documents

and perhaps too small to make a difference in retrieval effectiveness, when retrieval is limited to a sub-collection.

The above comments are corroborated to a degree by the graphs shown in Figure 6.17. The two graphs were produced from data corresponding to a combination of the Associated Press articles for 1988 (AP88) and Federal Register documents for the same year (FR88). The graph at the left in this figure shows weight of evidence as a function of term frequency for documents of 600 words or less, whereas the graph at the right shows the same curve for documents of 600 words or more. A comparison of the two graphs indicates that the rise in weight of evidence is more gradual for the longer documents than it is for the shorter ones.

6.9 Development of a Ranking Formula

In the section we take the $\mathbf{M_3}$ model and from it derive a probabilistic ranking formula in terms of weight of evidence.

6.9.1 Weight of Evidence as a Ranking Formula

The Probability Ranking Principle (see Chapter 5) councils us to rank documents by the probability of relevance. Equivalently, we may rank by the log-odds of relevance. The evidence we have considered to this point is: the coordination value; and,

for each of the query terms appearing in the document, the *idf* and *tf* values. In terms of this evidence, the log-odds of relevance is given by:

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

There are two steps that need to be taken to convert this to a ranking formula. First we note that we do not expect to have knowledge of the prior odds of relevance, $O(rel|q)$, for the query. However, this value is not needed. Since $O(rel|q)$ is constant for a given query, we can ignore it, and, for the purposes of ranking, simply use the weight of evidence as an RSV, in place of log-odds:

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

At this point, the theoretical results derived in the previous chapter can be applied. The (tf_i, idf_i) pairs in the second term can be viewed as separate sources of evidence. Based on the data analysis of sections, 6.5 and 6.6 we want a distribution such that:

$$w\hat{o}e(rel : idf_i, tf_i \mid co, q) \tag{6.17}$$

is constrained for each *tf-idf* pair. According to Corollary 5.1.1, for the maximum entropy distribution,

$$w\hat{o}e(rel : idf_1, tf_1, \ldots, idf_n, tf_n \mid co, q) = \sum_{i:\tau_i \in d} [w\hat{o}e(rel : idf_i, tf_i \mid co, q)] \tag{6.18}$$

Therefore, the ranking status value can be written as,

$$RSV = w\hat{o}e(rel : co \mid q) + \sum_{i:\tau_i \in d} [w\hat{o}e(rel : idf_i, tf_i \mid co, q)]$$

$$= \quad w\hat{o}e(rel : co \mid q) + \sum_{i:\tau_i \in d} [w\hat{o}e(rel : idf_i \mid co, q) + w\hat{o}e(rel : tf_i \mid idf_i, co, q)]$$

Using the regression equations, 6.11, 6.13, and 6.15, produced to model the weights of evidence in the above expression, yields a ranking status value in the form of:

$$RSV = \beta_0^{\text{CO}} + \beta_1^{\text{CO}} \cdot co + \sum_{i:\tau_i \in d} [\beta_0^{\text{IDF}} + \beta_1^{\text{IDF}} \cdot \widetilde{idf}_i + \beta_0^{\text{TF}} + \beta_1^{\text{TF}} \cdot \log(tf)]$$

$$\text{where: } \widetilde{idf}_i = \left\{ \begin{array}{ll} 0 & \text{if } idf_i < 1 \\ \beta_1^{\text{IDF}}(idf_i - 1) & \text{if } 1 \leq idf_i \leq 2 \\ \beta_1^{\text{IDF}} & \text{if } idf_i > 2 \end{array} \right\}$$

$$= \beta_0^{\text{CO}} + \sum_{i:\tau_i \in d} [\beta_1^* + \beta_1^{\text{IDF}} \cdot \widetilde{idf}_i + \beta_1^{\text{TF}} \cdot \log(tf)] \tag{6.19}$$

$$\text{where: } \beta_1^* = \beta_1^{\text{CO}} + \beta_0^{\text{IDF}} + \beta_0^{\text{TF}}$$

In the above formula β_0^{CO} can be ignored for the purposes of ranking.

The coefficient values corresponding to the regression which have been fit in the previous sections are given by:

$$\begin{aligned} (\beta_0^{\text{CO}}, \beta_1^{\text{CO}}) &= (-0.66, +0.42) \\ (\beta_0^{\text{IDF}}, \beta_1^{\text{IDF}}) &= (-0.49, +1.27) \\ (\beta_0^{\text{TF}}, \beta_1^{\text{TF}}) &= (-0.55, +1.25) \end{aligned}$$

giving the ranking formula:

$$RSV = -0.66 + \sum_{i:\tau_i \in d} [-0.62 + 1.27 \cdot \widetilde{idf}_i + 1.25 \cdot \log(tf)]$$

6.9.2 Discussion of the $\mathbf{M_3}$ Ranking Formula

It is instructive to compare the ranking formula given in eq. 6.19 with formulas commonly used in IR systems. First, the general form of the $\mathbf{M_3}$ formula is different

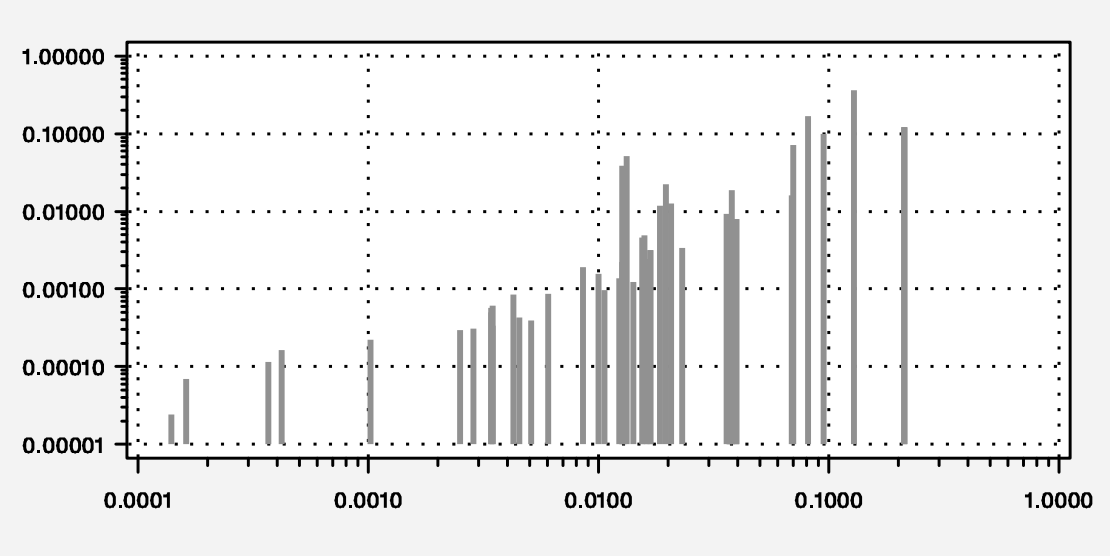

Figure 6.18. Observed $p(rel)$ vs. expected $p(rel)$ for 50 bins

from traditional *tf-idf* formulas. In classic versions of the formula, some function of the document frequency is multiplied by some function of the term frequency and these products are added over all terms appearing in the document.

$$\sum_{i:\tau_i \in d} \phi_1(tf) \cdot \phi_2(df)$$

In the $\mathbf{M_3}$ version, the *idf* and *tf* components are added.

A second difference is the introduction of a 3-piece linear function of $-\log \frac{df}{N}$ as part of the ranking formula. This form of the *idf* function is a direct result of the data analysis performed.

The *tf* formula in contrast is not novel. Intuition has suggested, and experimentation has validated, the idea that the impact of an increasing number of term occurrences should be dampened in retrieval formulas. Nonetheless, to our knowledge, the study of residual log-odds as a function of *tf* reported here is the first time direct evidence has been observed in support of this approach.

6.9.3 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 $\mathbf{M_3}$ 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 6.18 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 6.18 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.

6.9.4 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.

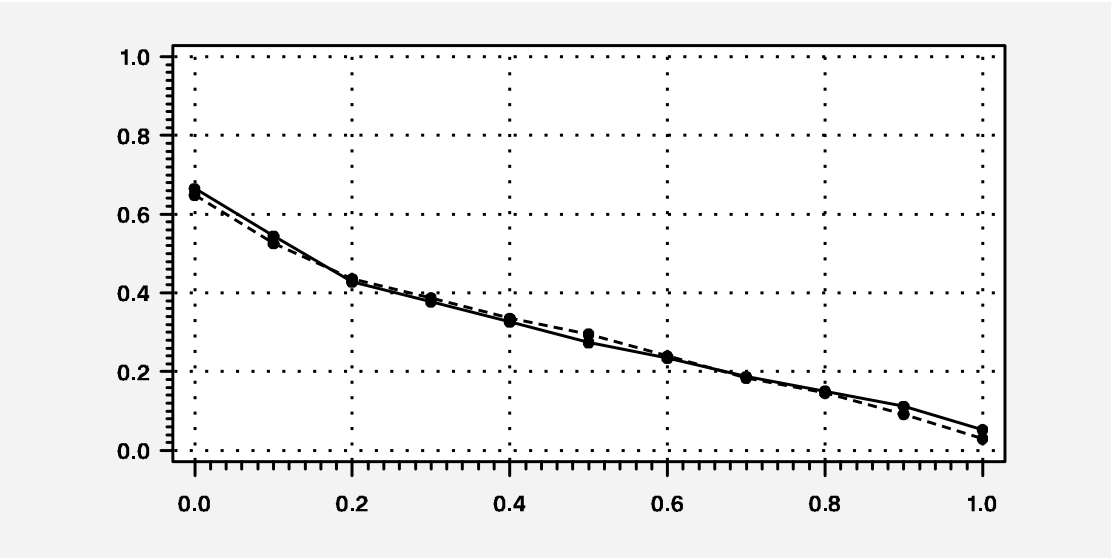

Figure 6.19. Recall-precision graph for TREC 3 queries on AP88

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 $\mathbf{M_3}$ model, the INQUERY information retrieval system [11] was modified to apply a formula based on eq. 6.19 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 settings. Figure 6.19 shows an 11-point recall-precision graph for the $\mathbf{M_3}$ 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 6.19 correspond to testing the $\mathbf{M_3}$ model on the same data set for which it was developed. However, test results on other data sets are promising.

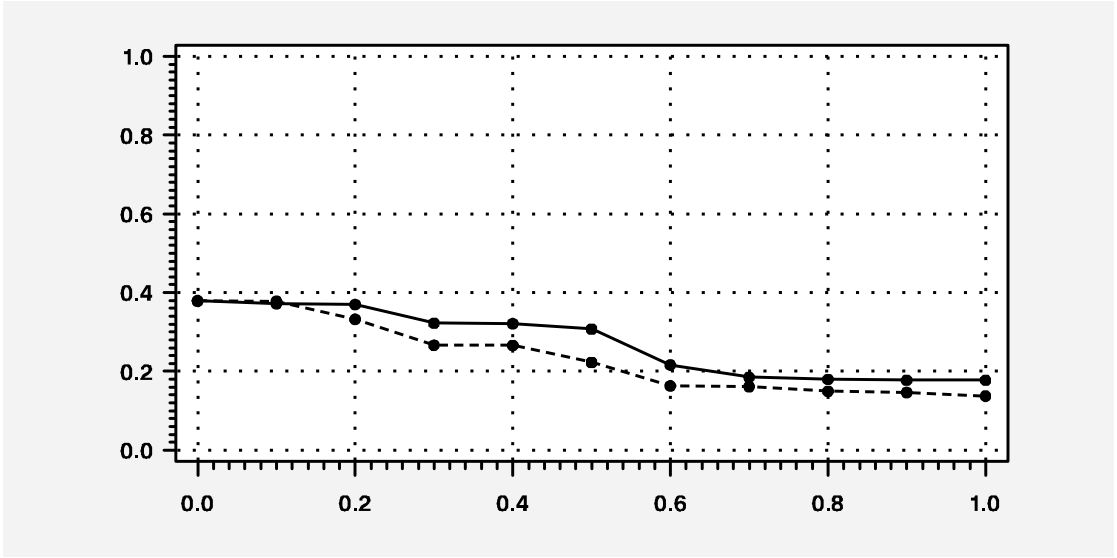

Figure 6.20. Recall-precision graphs for TREC 3 queries on ZIFF2

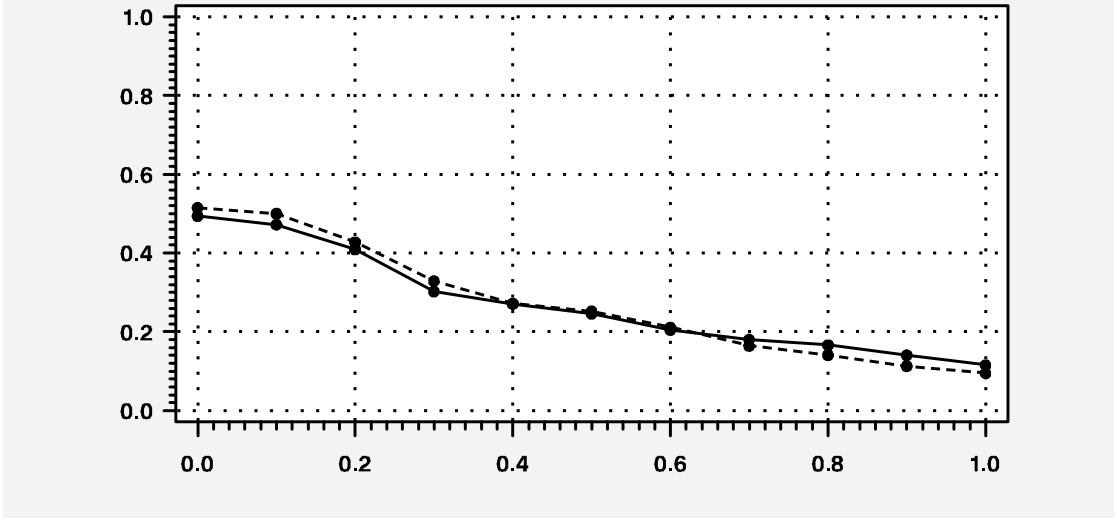

Figure 6.21. Recall-precision graphs for TREC 1 queries on WSJ89

Figures 6.20 and 6.21 show 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, 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.

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 bet-

ter 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.

CHAPTER 7

CONCLUSIONS

Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things.

Alan Agresti in *Categorical Data Analysis* [1, p. 28], says this is quote from 1887 paper by M.H. Doolittle, which he got from 1959 paper by Goodman and Kruskal.

This chapter reviews the principal contributions of this work and discusses a number of directions for further investigation.

7.1 Research Contributions

There are four general ways in which this dissertation can be considered to have contributed to the advance of information retrieval research.

Data driven methodology: Development of a methodology for the modeling of relevance as a direct product of the analysis of retrieval data.

Formal framework: Development of a formal framework based on weight of evidence and the Maximum Entropy Principle which guides the data exploration process and serves as a foundation for modeling.

Application of the methodology: Creation of a statistical model through the application of the methodology to available retrieval data.

Ranking formula: Derivation of a probabilistic ranking formula interpretable in terms of observable statistical regularities of retrieval data.

The following four sub-sections describe each of these general categories. In each case specific results of the research are listed giving a more detailed itemization of the contributions.

7.1.1 Methodology for the Study of Retrieval Evidence

The principle contribution of this research is the development of a methodology for the study of evidence used for the ranking of documents in response to the expression of a user's need for information. As discussed in the introduction, two general approaches to IR research can be identified. The engineering approach focuses on experimenting with a variety of techniques with the goal of increasing retrieval performance on some set of experimental data. The more theoretical approach starts from *a priori* assumptions and formally derives techniques that should improve retrieval. There is a rich interplay between these two approaches, as engineering techniques are used to convert theoretical results into practical improvements, and formal analysis is used to advance the understanding of progress due to engineering.

Rarely however are empirical methods used to directly assess or model retrieval strategies. As we have seen, statistical analysis was used to test the 2-Poisson hypothesis. The use of logistic regression by the group at Berkeley does probabilistic modeling, but not exploratory data analysis. A more direct precursor to this dissertation research is work on pivoted document length normalization [101, 100], where techniques very similar to those used here were used to develop a modification to the way document length is treated in ranking.

The research reported here, however, is the first time EDA has been applied to all sources of evidence that participate in the retrieval strategy. It can fairly be categorized as the first example of what we have termed the *data driven approach.*

The result is a complete probabilistic model from which a document ranking formula can be derived. Although, application of the methodology developed in this dissertation has been restricted to the study of features traditionally used in ranking formulas, it is a general methodology. It can be applied to other features and other issues which have been the subject of IR research for many years. Some of these possibilities will be discussed further in the section on future work.

Aspects of the methodology worthy of specific mention are:

Extensive use of graphical displays: Graphical representation of the data is the most important aspect of the methodology. To a degree, the extent of the reliance on graphical displays is evident from this thesis report itself. But it should also be kept in mind that much auxiliary graphical material is produced and studied before arriving at final graphs, such as those appearing in this document. Graphical displays play a key role in determining what it is that is worthy of study and how it can best be studied, before questions of scale and curve-fitting are even considered.

Analysis in terms of weight of evidence: We have argued that weight of evidence is particularly appropriate to the study and modeling of retrieval data because the goal is to predict a binary value in an environment where probabilities of interest tend to be small and vary over a range spanning a number of orders of magnitude. Although weight of evidence can be seen to have played a role in previous research, the explicit formal recognition of this role and the specific manner in which it has been incorporated into the overall methodology is a unique contribution of this work.

Study of *woe* residuals: A technique specific to the modeling of weight of evidence has been developed for the studying of residuals. This technique, which was applied for the development of each of the models, $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$, involved

comparison of observation with expectation: observation of the actual number of relevant documents in a subset of the data; expectation of the number of relevant documents based on an existing model.

woe **smoothing**: Techniques specific to weight of evidence have also been developed for the purpose of smoothing, motivation for which was given in Section 6.5.3. As described in the same section, the techniques involved smoothing over the counts that enter into the calculation of weight of evidence, rather than over the calculated values themselves. Due consideration is also given to insuring that the resulting smoothed estimates are robust.

Transformations resulting from EDA: Raw evidence used in document ranking is often transformed in some way. Prominent examples of this are the classical *idf* formula introduced by Sparck Jones [103], and the application of some dampening function to raw *tf* values, both of which are common in modern ranking formulas. The use of EDA to suggest promising transformations is, however, a contribution of this dissertation. This is made particularly clear in Chapter 4, where we see 1) how the classical form for *idf* could have arisen from the data analysis that was performed, and 2) how, as a result of the same data analysis, an improved form can emerge.

7.1.2 Formal Framework Based on WOE and MAXENT

As support for the data analysis and modeling, a formal framework based on the Maximum Entropy Principle has been produced. Initially, two classic probabilistic models were analyzed from the maximum entropy standpoint. This experience led to the formulation of a general theorem expressing conditions for *woe* additivity, which in turn, supported the modeling process that forms the core of this research.

In summary, specific contributions are:

The MAXENT-WOE Theorem: A theorem that gives very general conditions that, if satisfied by a set of constraints, will result in additive weights of evidence for the maximum entropy distribution.

Reformulation of the BIM and CMM models: As an immediate consequence of the MAXENT-WOE Theorem, both of these models can be understood from the perspective of maximum entropy.

Justification for the models developed: The MAXENT-WOE Theorem becomes a justification for additive aspects of the modeling approach adopted.

We have argued that adherence to the Maximum Entropy Principle carries with it a number of advantages which may be summarized as follows:

Generality: In theory at least, arbitrarily complex constraints can be considered, consistent with knowledge available to the modeling agent. This is in contrast to modeling limited to relatively simplistic forms of independence assumptions.

Principled approach: The approach may be considered principled in that it is founded on information theoretic notions. While one may not be disposed to accept the *principles* in question, they are clearly stated and open to more precise and rigorous analysis than more ad-hoc approaches permit.

Combination of evidence: Combinations of multiple sources of evidence can be processed in a consistent and intuitively plausible manner within the theoretic formalism.

New evidence: Existing models can be extended to incorporate new sources of evidence. Most probabilistic IR models that have been devised are tailor made to address specific issues or to explain specific empirical results. With these models it is generally unclear how they might be extended to address issues for which they were not originally designed.

7.1.3 Application of the Methodology to the Analysis of Retrieval Data

The methodology based on weight of evidence was applied to an available set of retrieval data. This resulted in the development of a model for weight of evidence in favor of relevance given by the query/document features that were studied.

7.1.4 Ranking Formula

A ranking formula has been derived directly from the weight-of-evidence model. The derived formula has two important characteristics:

Probabilistically interpretable RSV: The ranking status value has a strict probabilistic interpretation. It is the weight of evidence in favor of relevance given by the features of the query/document pair being evaluated.

Decomposable: The components of the formula can be decomposed. Both the form and the parameter values correspond to observable regularities of the retrieval data.

The formula that resulted from the modeling process differs from other probabilistic and non-probabilistic models in a number of interesting ways:

3-piece idf: The *idf* function corresponds to a 3-piece linear function of the fraction of documents in which the term appears. This is a novel contribution of this work.

additive *tf-idf* score: The ranking formula is a sum of *tf-idf* scores. In contrast with standard ranking formulas however, the *tf-idf* score is the sum of a function of term frequency and a function of document frequency. Typically, *tf-idf* scores are products.

constant term in *tf-idf* score: The *tf-idf* component scores also include an additive constant, corresponding to the sum of the β_0^{IDF}, β_0^{TF}, and β_1^{CO} coefficients.

7.2 Future Work

A central motivation for this dissertation is the development of a theoretical framework that allows for various forms of evidence to be incorporated in a general retrieval system in a systematic way. As mentioned above, it should be possible to apply the techniques developed as a result of this research to other sources of evidence. Alternatively, the same sources of evidence can be analyzed in different retrieval settings. The approach taken in this dissertation opens the door to a research paradigm that can be brought to bear on the study of all aspects of the information retrieval problem. A number of directions that are ripe for immediate exploration are outlined in this section.

7.2.1 Alternate Sources of Evidence

The dissertation has addressed the weight of evidence provided by the occurrence of query terms in documents when query terms are single words. However, other sources of evidence may be considered.

7.2.1.1 Alternative Query Terms

Information retrieval is not restricted to single words as terms. For example, phrases can be used as terms. The user may explicitly indicate the submission of a phrase [10], or the system may apply some heuristic technique to identify phrases in the query [66]. Systems such as INQUERY allow for the specification of proximity operators as part of a query [10]. Through the use of proximity operators, the user can form features based on the co-occurrence of more basic features (e.g. words, phrases) within a fixed *window* within a document. A window may be a component of natural language text, such as a sentence or paragraph, or simply an arbitrary fixed length sequence of words within a document.

Typically, features corresponding to phrases and proximity operators are, once identified, treated in the same way as single-word terms. It is reasonable to speculate,

however, that there is a systematic difference between single word terms, phrases and proximity operators with regard to the evidence their occurrence provides in favor of relevance. It follows that a weighting scheme tailored to the type of term should result in improved retrieval performance. The methodology employed in this thesis can be used to derive such a weighting formulation and test this hypothesis.

7.2.1.2 Query Term Sub-categories

At the same time that it might be beneficial to base weighting on the category (e.g. single word, phrase, proximity operator) of a term, it is reasonable to explore the way evidence may differ for different sub-categories. For single word terms a number of attributes come immediately to mind. Intuition suggests that a word's part of speech, or whether a term is a proper noun, or simply whether or not a word is capitalized, may well be correlated with differences in the weight of evidence associated with a query term. For automatic query formulation from topic descriptions, such as those developed for the TREC competitions, whether the term was derived from the title, description, or narrative fields, or some combination of them may also be useful for estimating weight of evidence.

In 1957 Luhn pronounced, "there are as yet many unanswered questions, such as whether nouns and adjectives or other portions of sentences furnish ... the most effective discriminating elements." [74]. Such questions still await a precise quantitative response.

7.2.1.3 Query Expansion

A technique frequently employed to improve retrieval performance is *query expansion*, whereby the query is extended by terms deemed to be related to the query, but not explicitly specified by the user. When relevance judgments are available, the query can be expanded with terms that appear frequently in documents that are known to be relevant. *Local feedback* is an expansion technique that does not rely on

relevance feedback [23]. In this approach, an initial search is realized using the query as given by the user. The high ranking documents from this initial query are then treated as if they were known to be relevant, and are used for query expansion.

Qui and Frei [86] represent terms as vectors, the components of which are given by a measure of the association between the term and each of the documents in the collection. A query is associated with a weighted sum of the query terms. The query can then be expanded by adding terms with vector representations close to that of the query vector.

Jing and Croft build an *association thesaurus* of noun phrases, and expand queries with related phrases from the thesaurus [66]. Each phrase is represented as a list of term/frequency pairs. The terms are those that co-occur with the phrase in the corpus and the frequency is the number of co-occurrences. This representation of phrases is the same as the representation used for documents. In order to determine expansion phrases, the query is applied to the phrase thesaurus as if it were a collection of documents. Top ranked phrases are added to the query.

Many other techniques have been proposed. Whatever the mechanism for expansion, the final query will be a combination of original terms and those that have been chosen as a result of expansion. Typically, once expansion terms are chosen, they are simply affixed to the query, with no distinction between the original terms and those that have been added. But, here again it is reasonable to investigate the possibility that, with respect to their weights in favor of relevance, there is a systematic difference between terms chosen initially by the user and those added by a given expansion strategy.

It is true that researchers have often experimented with schemes for weighting expansion terms differently. But these tend to be ad-hoc attempts, where some parameterized weighting is chosen and parameter values are set as a result of empirical testing. The methodology developed as a result of this thesis offers the opportu-

nity to develop a more principled approach to automatically incorporating new terms in a query. Furthermore, the behavior of terms resulting from different expansion techniques can be compared and contrasted with the goal of obtaining greater understanding of what these techniques are doing; why it is that they work; and how they might be improved.

The study of query expansion with terms produced as a result of *Local Context Analysis* [114] is a promising area for exploration. Local Context Analysis (LCA) is a variation on local feedback that does not assume that all top-ranked documents are relevant. It is based instead on the hypothesis that, "Query expansion using words co-occurring with all words in the top ranked documents will produce more effective retrieval" [113, p. 53]. LCA utilizes a formula for ranking the terms appearing in top-ranked documents as candidates for query expansion. The ranking formula takes into consideration the number of documents in the collection that contain the term and the number of times the term co-occurs with the query terms in the collection. Experiments have shown that local context analysis is very effective, not only for the ad-hoc retrieval task [114], but also in the context of cross-lingual retrieval [4] and topic segmentation [83].

A direct study of weight of evidence in favor of relevance can be beneficial in many ways with respect to LCA. As mentioned above, it can be used to determine how best to weight expansion terms, relative to terms appearing in the original query. A variable to be considered in this weighting is the score assigned to expansion terms as part of the process of ranking them.

The formula used to rank candidate expansion terms can, itself, be studied using techniques of exploratory analysis. The goal in this case would be to uncover correlations between the characteristics of a candidate term input to the Xu ranking formula and the quality of the expansion term. A formal definition of *expansion term quality* would be required for this purpose.

7.2.2 Alternative Retrieval Settings

An IR system evolves as a result of experimentation and experience with retrieval in a given setting. As information retrieval technology has matured and proven itself useful, systems developed for one purpose have been applied to new tasks and new environments. Unfortunately, the lack of solid theoretical foundations and established methodology makes it difficult, if not impossible, to attack these problems from first principles.

7.2.2.1 Differences across Languages

Often a retrieval system built for one language, say English, is used, essentially as is, for document searching in another language of the same general linguistic family – Spanish perhaps. Text pre-processing routines may need to be reworked to accommodate, for example, differences in the morphological rules used for stemming, and interface issues may need to be addressed. Typically, however, the same basic techniques and weighting formulas are used without modification. The implicit assumption is that the nature of retrieval in one language is the same as (or is similar enough to) the problem of retrieval for a different language. This approach appears to work, but the result may be sub-optimal. Is term weighting necessarily the same across languages?

In [36], Fujii expresses the view that, "the general research method has not yet been systematically explored, despite there being a number of studies of non-English retrieval" [p. 21]. He finds that existing studies "tended either to focus on the idiosyncratic characteristics of a specific target language (*often in an ad-hoc manner*), or to ignoring the effecting factor of the language(s) used in the experiments[1]" [36, p. 22]. The data driven approach provides a methodology for the study of this issue.

[1] italics added

Once sufficient data has been accumulated in a new language, such as Spanish, the adequacy of utilizing a weighting formula that had been originally developed for English could be tested directly. If the fit is good, fine. If it is not, exploratory techniques such as residual plots, should help suggest where to begin the process of modification.

7.2.2.2 Specialized Tasks

In the same way that retrieval algorithms may be used, as is, for new languages, weighting formulas are often used, unmodified, for new specialized tasks. For example, Allan, Papka & Lavrenko have reported on an algorithm for unsupervised event detection, in which documents are treated as queries [2]. Each document, as it is received (for example, over a newswire), is evaluated as a query against all documents previously encountered. The resulting ranking scores are then used as a measure of document similarity. The measure of similarity is used, in turn, to assess how likely it is that the document in question concerns a new event. It is doubtful, however, that a term weighting formula that is optimal when the objective is prediction of relevance to a query, will also be optimal when the objective is prediction of a new event. The degree to which this may be a problem for the event detection task can be directly appraised by way of a post-hoc analysis of the data.

Allan *et al.* have also shown that the time between arrival of two documents is a good source of evidence as to whether or not a document will be considered to encompass a new event. They utilize this information in their detection algorithm, but the incorporation of elapsed time in the algorithm is *ad-hoc*. A non-parametric regression with a binary new-event response variable, and elapsed-time and ranking-score as predictor variables, may be a better way to model the problem. Aside from the possibility of improved performance due to a better combination-of-evidence formula, a score resulting from the formula would have semantic content as an estimate of the

probability that a document would be considered to correspond to a new event. Furthermore, a probabilistic interpretation would allow decision theoretic principles to be employed for the setting of a threshold value, given a user specified cost function.

Another example of how the formula used for standard retrieval is used for a very different purpose was given in Section 7.2.1.3. When the association thesaurus of Jing and Croft is used for term expansion, the query is run against the thesaurus as if it were a document collection. This approach works well, but there is good reason to think it might be made to work better, since the ranking formula used was developed for a different purpose. More important, once again, is that a study of this nature may lead to a better understanding of characteristics of information retrieval; in this case with regard to the role phrases can play in the expansion of queries.

Also mentioned in Section 7.2.1.3 was the use of local context analysis for applications such as cross-lingual retrieval and text segmentation. Here again a technique derived with one application in mind is called into service in a novel setting. In the context of specific alternate applications, the study of how LCA terms are chosen, and what weights might best be associated with the terms once they are chosen, may prove both informative and beneficial.

7.2.2.3 Specialized Collections

Even if both the language and the essential character of the task are constant across applications, differences in the nature of collections may require variations in the ranking formula used, if optimal retrieval is to be achieved. It is certainly conceivable that the relationship of term characteristics to weight of evidence for queries used for searching a collection of legal documents may be different from the relationship for terms of queries made against a collection of newspaper articles. Queries against a collection of abstracts may differ from queries against a collection of full documents. Queries submitted to retrieval engines on the world wide web are

likely to display very different behavioral characteristics from information searches in more restricted environments.

It is hard to see how these questions can be addressed in a principled way with the methods currently employed in information retrieval research and system development. The data driven approach, used to study TREC queries in this thesis, could be used to study data collected in different environments. Differences in the behavior of query terms, or other sources of evidence, should not be difficult to uncover if they exist. Where discrepancies are observed, the theory developed as a result of the study of TREC data should serve to guide the investigation of variations across collections. As an additional benefit, the study of differences across collections may well provide insights into the general retrieval problem that might not be apparent from the study of individual collections in isolation.

7.2.3 Boolean Queries

As explained in Section 2.5, the original motivation for this research was previous work on the modeling of Boolean queries in the context of the INQUERY inference network. With reliable estimates of probabilities of relevance, generated by a model derived from inspection and analysis of extensive retrieval data, a more principled attempt at the modeling of Boolean queries can be made. Probabilistic intuitions should be a more trustworthy guide to modeling when the inputs to Boolean operators correspond to reliable estimates. Surprising or disappointing results are more likely to succumb to intuitive analysis.

More important, and more in concert with the overall philosophy of this research, is that the engineering approach relying solely on intuition and trial-and-error search, which characterizes earlier work such as that reported in [45], can be replaced by data driven analysis. As with the phrases and proximity operators discussed in Section 7.2.1.1, Boolean combinations of query terms can be treated as sources of evi-

dence in their own right. Post-hoc determination of the weight of evidence provided by components of Boolean queries can be computed, and multivariate non-parametric regression techniques can be applied. The relationship between univariate weights of evidence and relevance can be studied for a number of queries, and a model can be developed for the weight of evidence provided by a combination of components as a function of the weights of evidence given by each component individually.

Also, the data driven approach can help determine how best to take advantage of other aspects of the inference network framework. In the same way that Boolean combinations can be studied, combinations of multiple query formulations, or evidence provided by multiple document representations, can be investigated.

BIBLIOGRAPHY

[1] Agresti, A. *Categorical Data Analysis.* John Wiley & Sons, New York, 1990.

[2] Allan, J., Papka, R., and Lavrenko, V. On line new event detection using single pass clustering. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia, Aug. 1998), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds., ACM Press.

[3] Andrews, D. F. Data analysis, exploratory. In *International Encyclopedia of Statistics*, W. H. Kruskal and J. M. Tanur, Eds., vol. 7. Free Press, New York, 1978, pp. 210–218.

[4] Ballesteros, L., and Croft, W. B. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Philadelphia, Pennsylvania, July 1997), N. J. Belkin, A. D. Narasimhalu, and P. Willett, Eds., ACM Press, pp. 84–91.

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

[6] Beniger, J. R., and Brown, D. L. Quantitative graphics in statistics: A brief history. *The American Statistician 32*, 1 (1978), 1–9.

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

[8] Bretthorst, G. L. Excerpts from bayesian spectrum analysis and parameter estimation. In *Maximum Entropy and Bayesian Methods in Science and Engineering* (Norwell, MA, 1988), G. J. Erickson and C. R. Smith, Eds., Kluwer Academic Publishers, pp. 75–146.

[9] Brillouin, L. Observation, information and imagination. In *Information and Prediction in Science*, S. Dockx and P. Bernays, Eds. Academic Press, New York, 1965, pp. 1–14.

[10] Callan, J. P., Croft, W. B., and Broglio, J. TREC and TIPSTER experiments with INQUERY. *Information Processing & Management 31*, 3 (1995), 327–343.

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

[12] Carlin, B. P., and Louis, T. A. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, London, 1996.

[13] Charniak, E. Bayesian networks without tears. *AI Magazine 12*, 4 (Apr. 1991), 50–63.

[14] Chiang, A. C. *Fundamental Methods of Mathematical Economics.* McGraw-Hill, New York, 1967.

[15] Church, K., Gale, W., Hanks, P., and Hindle, D. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon* (Hillsdale, NJ, 1991), U. Zernik, Ed., Lawrence Erlbaum Associates, pp. 115–164.

[16] Cooper, W. S. Exploiting the maximum entropy principle to increase retrieval effectiveness. *Journal of the American Society for Information Science 34*, 1 (1983), 31–39.

[17] Cooper, W. S. Some inconsistentencies and misnomers in probabilistic information retrieval. In *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (Chicago, Illinois, USA, Oct. 1991), A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan, Eds., pp. 57–61.

[18] Cooper, W. S. The formalism of probability theory in IR: A foundation or an encumbrance. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland, July 1994), W. B. Croft and C. J. van Rijsbergen, Eds., pp. 242–248.

[19] Cooper, W. S., Chen, A., and Gey, F. C. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In *The Second Text REtreival Conference (TREC-2)* (Gaithersburg, Md., Mar. 1994), D. K. Harman, Ed., NIST Special Publication 500-215, pp. 57–66.

[20] Cooper, W. S., Chen, A., and Gey, F. C. Experiments in the probabilistic retrieval of full text documents. In *The Third Text REtreival Conference (TREC-3)* (Gaithersburg, Md., Apr. 1995), D. K. Harman, Ed., NIST Special Publication 500-225, pp. 127–134.

[21] Cooper, W. S., Dabney, D., and Gey, F. Probabilistic retrieval based on staged logistic regression. In *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Copenhagen, Denmark, June 1992), N. Belkin, P. Ingwersen, and A. M. Mejtersen, Eds., pp. 198–210.

[22] Cooper, W. S., and Huizinga, P. The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology, Research & Development 1* (1982), 99–112.

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

[24] Croft, W. B., and Xu, J. Corpus-specific stemming using word form co-occurence. In *Proceedings for the Fourth Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, Nevada, Apr. 1995), pp. 147–159.

[25] Dawid, A. P. Probability forecasting. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson, Eds., vol. 7. Wiley, New York, 1989, pp. 210–218.

[26] DeGroot, M., and Feinberg, S. The comparison and evaluation of forecasters. *The Statistician 32* (1982), 12–22.

[27] Devroye, L. *A Course in Density Estimation.* Birkhauser, Boston, 1987.

[28] Durant, W. *The Story Of Philosophy: The Lives And Opinions Of The Great Philosophers*, 2 ed. Simon and Schuster, New York, 1938.

[29] Erickson, G. J., and Smith, C. R. *Maximum Entropy and Bayesian Methods in Science and Engineering.* Kluwer Academic Publishers, Norwell, MA, 1988.

[30] Faires, J. D., and Faires, B. T. *Calculus*, 2 ed. Random House, New York, 1988.

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

[32] Fine, T. L. *Theories of Probability: An Examination of Foundations.* Academic Press, New York, 1973.

[33] Fox, E., Betrabet, S., and Koushik, M. Extended boolean models. In *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Englewood Cliffs, NJ, 1992, pp. 393–418.

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

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

[36] Fujii, H. *An Investigation of the Linguistic Characteristics of Japanese Information Retrieval.* PhD thesis, University of Massachusetts, Amherst, Massachusetts, Feb. 1998.

[37] Gey, F. C. Inferring probability of relevance using the method of logistic regression. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland, July 1994), W. B. Croft and C. J. van Rijsbergen, Eds., pp. 222–231.

[38] Gey, F. C., Chen, A., He, J., and Meggs, J. Logistic regression at TREC4: Probabilistic retrieval from full text document collections. In *The Fourth Text REtreival Conference (TREC-4)* (Gaithersburg, Md., Oct. 1996), D. K. Harman, Ed., NIST Special Publication 500-236, pp. 65–72.

[39] Gey, F. C., Chen, A., He, J., Xu, L., and Meggs, J. Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms at TREC-5. In *The Fifth Text REtreival Conference (TREC-5)* (Gaithersburg, Md. 500-238, Nov. 1997), E. M. Voorhees and D. K. Harman, Eds., NIST Special Publication 500-238, pp. 181–190.

[40] Golan, A., Judge, G. G., and Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* John Wiley & Sons, New York, 1996.

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

[42] Good, I. J. *Good Thinking: The Foundations of Probability and its Applications.* University of Minnesota Press, Minneapolis, 1983.

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

[44] Good, I. J. Statistical evidence. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson, Eds. Wiley, 1989, pp. 651–656.

[45] Greiff, W. R., Croft, W. B., and Turtle, H. Computationally tractable probabilistic modeling of boolean operators. In *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Philadelphia, Pennsylvania, July 1997), N. J. Belkin, A. D. Narasimhalu, and P. Willett, Eds., ACM Press, pp. 119–128.

[46] Greiff, W. R., Croft, W. B., and Turtle, H. PIC matrices: A computationally tractable class of probabilistic query operators. To appear in *ACM Transactions on Information Systems* (1999).

[47] Greiff, W. R., and Ponte, J. The maximum entropy approach and probabilistic IR models. To appear in *ACM Transactions on Information Systems* (1999).

[48] Gull, S. F., and Daniell, G. J. Image reconstruction from incomplete and noisy data. *Nature 272* (1978), 686–690.

[49] Hacking, I. *Logic of Statistical Inference.* Cambridge University Press, Cambridge, 1965.

[50] Haines, D., and Croft, W. B. Relevance feedback and inference networks. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pittsburgh, Pa. USA, June 1993), R. Korfhage, E. Rasmussen, and P. Willett, Eds., pp. 191–203.

[51] Harman, D. Overview of the first Text REtrieval Conference (TREC-1). In *The First Text REtrieval Conference (TREC1)* (Gaithersburg, Md., Feb. 1993), D. K. Harman, Ed., NIST Special Publication 500-207, pp. 1–20.

[52] Harman, D. Overview of the third Text REtrieval Conference (TREC-3). In *The Third Text REtreival Conference (TREC-3)* (Gaithersburg, Md., Apr. 1995), D. K. Harman, Ed., NIST Special Publication 500-225, pp. 1–20.

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

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

[55] Harter, S. P. A probabilistic approach to automatic keyword indexing, Part I: On the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science 26* (1975), 197–206.

[56] Harter, S. P. A probabilistic approach to automatic keyword indexing, Part II: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science 26* (1975), 280–289.

[57] Hartwig, F., and Dearing, B. E. *Exploratory Data Analysis.* No. 07-016 in Sage university paper series: Quantitative applications in the social sciences. Sage Publications, Beverly Hills, 1979.

[58] Hosmer, Jr, D. W., and Lemeshow, S. *Applied Logistic Regression.* John Wiley & Sons, New York, 1989.

[59] Härdle, W. *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, 1990.

[60] JAYNES, E. T. Information theory and statistical mechanics: Part I. *Physical Review 106* (1957), 620–630.

[61] JAYNES, E. T. Information theory and statistical mechanics: Part II. *Physical Review 108* (1957), 171.

[62] JAYNES, E. T. Information theory and statistical mechanics. In *Statistical Physics: Brandeis Summer Institute Lectures in Theoretical Physics*, G. E. Uhlenbeck, Ed., vol. 3 of *Brandeis Summer Institute Lectures in Theoretical Physics*. W. A. Benjamin, New York, 1963, pp. 182–218.

[63] JAYNES, E. T. Where do we stand on maximum entropy. In *The Maximum Entropy Formalism* (Cambridge, Massachusetts, May 1979), R. D. Levine and M. Tribus, Eds., MIT Press, pp. 15–118.

[64] JAYNES, E. T. Probability theory: The logic of science. available via ftp://bayes.wustl.edu/pub/Jaynes/book.probability.theory/, 1994.

[65] JEFFREYS, H. *Theory of Probability*, 3 ed. Oxford University Press, Oxford, 1961.

[66] JING, Y., AND CROFT, W. B. An association thesaurus for information retrieval. In *RIAO 94 Conference Proceedings* (1994), Lawrence Erlbaum Associates, pp. 146–160.

[67] KAC, M. *Enigmas of Chance: An Autobiography.* Harper & Row Press, New York, 1985.

[68] KANTOR, P. B. Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology, Research & Development 3*, 2 (Apr. 1984), 88–94.

[69] KANTOR, P. B., AND LEE, J. J. The maximum entropy principle in information retrieval. In *Proceedings of the 9th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy, Sept. 1986), F. Rabitti, Ed., pp. 269–274.

[70] KANTOR, P. B., AND LEE, J. J. Testing the maximum entropy principle for information retrieval. *Journal of the American Society for Information Science 49*, 6 (1998), 557–566.

[71] KIM, J. H., AND PEARL, J. A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence* (Karlsruhe, West Germany, August 1983), pp. 190–193.

[72] KROVETZ, R. Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pittsburgh, Pa. USA, June 1993), R. Korfhage, E. Rasmussen, and P. Willett, Eds., pp. 191–203.

[73] Lee, J. J., and Kantor, P. B. A study of probabilistic information retrieval in the case of inconsistent expert judgments. *Journal of the American Society for Information Science 42*, 3 (1991), 166–172.

[74] Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. *The IBM Journal of Research and Development 1*, 4 (Oct. 1957), 309–317.

[75] Luhn, H. P. The automatic creation of literature abstracts. *The IBM Journal of Research and Development 2*, 2 (Apr. 1958), 159–164.

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

[77] Maritz, J. S. *Empirical Bayes Methods*. Chapman and Hall, London, 1989.

[78] Maron, M. E., and Kuhns, J. L. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM 7*, 3 (July 1960), 216–243.

[79] Marshall, K. T., and Oliver, R. M. *Decision Making and Forecasting: with Emphasis on Model Building and Policy Analysis*. McGraw-Hill, New York, 1995.

[80] Minsky, M., and Selfridge, O. G. Learning in random nets. In *Information Theory: Fourth London Symposium* (London, 1961), C. Cherry, Ed., Butterworths, pp. 335–347.

[81] Neter, J., Wasserman, W., and Kutner, M. H. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*, 2 ed. R. D. Irwin, Homewood, Ill., 1985.

[82] Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[83] Ponte, J., and Croft, W. B. Text segmentation by topic. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries* (1997), pp. 120–129.

[84] Ponte, J. M. *Probabilistic Language Models for Topic Segmentation and Information Retrieval*. PhD thesis, University of Massachusetts, Amherst, Massachusetts, May 1998.

[85] Ponte, J. M., and Croft, W. B. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia, Aug. 1998), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds., ACM Press, pp. 275–281.

[86] Qui, Y., and Frei, H. P. Concept based query expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pittsburgh, Pa. USA, June 1993), R. Korfhage, E. Rasmussen, and P. Willett, Eds., pp. 160–169.

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

[88] Robertson, S. E. The probability ranking principle in IR. *Journal of Documentation 33* (1977), 294–304.

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

[90] Robertson, S. E., and Walker, S. On relevance weights with little relevance information. In *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Philadelphia, Pennsylvania, July 1997), N. J. Belkin, A. D. Narasimhalu, and P. Willett, Eds., pp. 16–24.

[91] Sahai, H., and Khurshid, A. *Statistics in Epidemiology.* CRC Press, Boca Raton, 1996.

[92] Salton, G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Publishing Company, Reading, MA, 1989.

[93] Salton, G., Buckley, C., and Fox, E. A. Automatic query formulations in information retrieval. *Journal of the American Society for Information Science 34*, 4 (July 1983), 262–280.

[94] Salton, G., Fox, E. A., and Wu, H. Extended Boolean information retrieval. *Communications of the ACM 26*, 12 (Dec. 1983), 1022–1036.

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

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

[97] Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal 27* (1948), 379–423 & 623–656.

[98] Shannon, C. E. Coding theorems for a discrete source with a fidelity criterion. In *Information and Decision Processes*, R. E. Machol, Ed. McGraw-Hill, New York, 1960, pp. 93–126.

[99] Silverman, B. W. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[100] Singhal, A., Buckley, C., and Mitra, M. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Zurich, Switzerland, aug 1996), H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, Eds., ACM Press, pp. 21–29.

[101] Singhal, A., Salton, G., Mitra, M., and Buckley, C. Document length normalization. *Information Processing & Management 32*, 5 (Sept. 1996), 619–633.

[102] Smeaton, A. F., and van Rijsbergen, C. J. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal 25*, 3 (1983), 239–246.

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

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

[105] Tribus, M. *Rational Descriptions, Decisions, and Designs.* Pergamon-Hall, New York, 1969.

[106] Tribus, M. Thirty years of information theory. In *The Maximum Entropy Formalism* (Cambridge, Massachusetts, May 1979), R. D. Levine and M. Tribus, Eds., MIT Press, pp. 1–14.

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

[108] Turtle, H., and Croft, W. B. Inference networks for document retrieval. In *Proceedings of the 13th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Brussels, Belgium, Sept. 1990), J.-L. Vidick, Ed., pp. 1–24.

[109] Turtle, H. R. *Inference Networks for Document Retrieval.* PhD thesis, University of Massachusetts, 1990.

[110] van Rijsbergen, C. J. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation 33* (1977), 106–119.

[111] van Rijsbergen, C. J. *Information Retrieval*, 2 ed. Butterworths, London, 1979.

[112] Witten, I. H., Moffat, A., and Bell, T. C. *Managing Gigabytes: Compressing and Indexing Documents and Images.* van Nostrand Reinhold, New York, 1994.

[113] Xu, J. *Solving the Word Mismatch Problem Through Automatic Text Analysis.* PhD thesis, University of Massachusetts, Amherst, Massachusetts, May 1997.

[114] Xu, J., and Croft, W. B. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Zurich, Switzerland, aug 1996), H.-P. Frei, D. Harman, P. Schäube, and R. Wilkinson, Eds., pp. 4–11.

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

[116] Yu, C. T., and Mizuno, I. Two learning schemes in information retrieval. In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval* (Grenoble, France, June 1988), Y. Chiaramella, Ed., pp. 201–215.