IR-145

The Maximum Entropy Approach and Probabilistic IR Models

Warren R. Greiff and Jay M. Ponte

Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts, Amherst

(October 5, 1998)

Abstract

The Principle of Maximum Entropy is discussed and two classic probabilistic models of information retrieval, the Binary Independence Model of Robertson and Sparck Jones and the Combination Match Model of Croft and Harper are derived using the maximum entropy approach. The assumptions on which the classical models are based are not made. In their place, the probability distribution of maximum entropy consistent with a set of constraints is determined. It is argued that this subjectivist approach is more philosophically coherent than the frequentist conceptualization of probability that is often assumed as the basis of probabilistic modeling and that this philosophical stance has important practical consequences with respect to the realization of information retrieval research.

1 Introduction

This paper takes a fresh look at modeling approaches to information retrieval that have been the basis of much of the probabilistically motivated IR research over the last twenty years. We shall adopt a subjectivist Bayesian view of probabilities and argue that classical work on probabilistic retrieval is best understood from this perspective. The main focus of the paper will be the ranking formulas corresponding to the Binary Independence Model (BIM), presented originally by Robertson and Sparck Jones [RS77] and the Combination Match Model (CMM), developed shortly thereafter by Croft and Harper [CH79]. We will show how these same ranking formulas can result from a probabilistic methodology commonly known as Maximum Entropy (MAXENT).

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. In this paper we will show how these formal models can be derived from the Principle of Maximum Entropy, which counsels us to select the probability distribution with maximum entropy of all those that conform to

1.1 Maximum Entropy

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:

$$p(H|E,K) = \frac{p(E|H,K) \cdot p(H|K)}{p(E|K)} \tag{1}$$

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 has 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:

$$p(E = e \mid K) \quad = \quad \sum_{i=1}^{n} p(E = e \wedge H = h_i \mid K) = \sum_{i=1}^{n} p(E = e \mid H = h_i, K) p(H = h_i \mid K)$$

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 [Fin73, Hac65] for more in depth discussions of these issues.

The Principle of Maximum Entropy: At the end of the 19th century, primarily as a result of the work of Maxwell, Boltzmann and Gibbs [Jay79], 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" [Sha48]. He showed that for a probability distribution, $\mathbf{p} = (p_1, \ldots, p_k)$, over k possible elementary events, the quantity:

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" [Tri79]. In a pair of seminal articles, [Jay57a, Jay57b], 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" [Jay57a]. His method of inference is based on what has come to be known as the Principle of Maximum Entropy. 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. [Jay57a, 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 the greatest entropy subject to constraints on the expectations of a given set of functions of the variable. That is, the distribution that maximizes (2) subject to a set of constraints:

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

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

An example: The example given here is an adaptation of the "Brandeis Dice Problem" originally presented as an illustration of the maximum entropy approach in [Jay63].

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 [Rob77], 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?

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,

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

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 MAXENT 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 which 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 [GJM96, 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.

1.2 The MAXENT Approach and Probabilistic IR Modeling

In two papers in the early '80s, Cooper and Huizinga [CH82] and Cooper [Coo83], 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" [Coo83]. 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]" [CH82, 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 [SWY76] 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 [Kan84, KL86], Kantor and Lee extend the analysis of the Principle of Maximum Entropy in the context of information retrieval. In [LK91] 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, [KL98], they have conducted experiments to test the performance of the PME as a method of document retrieval. While they outperform two simpler methods on small collections, they report discouraging results on large document sets and conclude that the PME, in general, does not appear to present advantages over more "naive" methods.

In contrast to the work of Kantor and Lee in [KL98], our interest is not in the development of an alternative retrieval algorithm based on the PME. 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 Principle of Maximum Entropy. The primary objectives of this paper 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 uses mathematically derived forms of the maximum

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" [Jay94, 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 paper, the following view of a probabilistic retrieval system is adopted. The rank 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.

1.3 Contents of Paper

In the next section we 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. With this, we will be prepared for the main purpose of the paper.

In Section 3, we show how the essence of the Binary Independence Model (BIM) can be derived from the Principle of Maximum Entropy. 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 MAXENT model in that it is a characteristic of the resulting probability distribution. In section 4 we go on to show how the work of Croft and Harper can also be reproduced from the maximum entropy standpoint. Again we compare the approach taken by the original authors to that adopted with MAXENT.

In Section 5, we discuss the two models we have developed with the MAXENT approach. More specifically, we will discuss the differences between

2 Background

2.1 Binary Independence Model

The Binary Independence Model (BIM), developed by Robertson and Sparck Jones [RS77, van79], 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:

$$O(rel|x_1,\ldots,x_s) = \frac{p(x_1,\ldots,x_s|rel)}{p(x_1,\ldots,x_s|\overline{rel})} \cdot O(rel) \tag{3}$$

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

$$\forall(x_1,\ldots,x_s) \in \{0,1\}^s : \quad p(x_1,\ldots,x_s|rel) = \prod_{i=1}^{s} p(x_i|rel) \tag{4}$$

$$p(x_1,\ldots,x_s|\overline{rel}) = \prod_{i=1}^{s} p(x_i|\overline{rel}) \tag{5}$$

From which, it immediately follows that:

$$\forall(x_1,\ldots,x_s) \in \{0,1\}^s : \frac{p(x_1,\ldots,x_s|rel)}{p(x_1,\ldots,x_s|\overline{rel})} = \prod_{i=1}^{s} \frac{p(x_i|rel)}{p(x_i|\overline{rel})} \tag{6}$$

William Cooper later emphasized that equation (6) is all that really needs to be assumed [Coo91]. This, "linked dependence assumption" is weaker than the pair of conditional independence assumptions, (4) and (5), 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.

The product on the right of (6) can be reorganized into two separate products: one over terms that do occur in the document and one over terms that don't:

$$\prod_{i=1}^{s} \frac{p(x_i|rel)}{p(x_i|\overline{rel})} = \prod_{x_i=0} \frac{p(x_i|rel)}{p(x_i|\overline{rel})} \cdot \prod_{x_i=1} \frac{p(x_i|rel)}{p(x_i|\overline{rel})}$$

By extracting $\frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})}$ from each of the factors, we obtain a formula that involves the multiplication of a value which is independent of the term occurrence pattern by a term-dependent coefficient for each of the terms occurring in a document,

$$\prod_{i=1}^{s} \frac{p(x_i|rel)}{p(x_i|\overline{rel})} = \prod_{x_i=0} \frac{p(x_i|rel)}{p(x_i|\overline{rel})} \Big/ \frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})} \cdot \prod_{x_i=1} \frac{p(x_i|rel)}{p(x_i|\overline{rel})} \Big/ \frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})} \cdot \prod_{i=1}^{s} \frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})}$$

$$\prod_{i=1}^{s} \frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})} \cdot \prod_{x_i=0} \frac{p(x_i|rel)}{p(x_i|\overline{rel})} \Big/ \frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})} \cdot \prod_{x_i=1} \frac{p(x_i|rel)}{p(x_i|\overline{rel})} \Big/ \frac{p(X_i=1|rel)}{p(X_i=1|\overline{rel})} \cdot \prod_{i=1}^{s} \frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})}$$

Under the linked dependence assumption (6), the expression (7) may be substituted for the fraction in (3), giving:

$$O(rel|x_1,\ldots,x_s) = \prod_{x_i=1}^{s} \frac{p(X_i=1|rel)}{p(X_i=1|\overline{rel})}\frac{p(X_i=0|\overline{rel})}{p(X_i=0|rel)} \cdot \prod_{i=1}^{s}\frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})} \cdot O(rel) \quad (8)$$

Taking the log of both sides yields:

$$\log O(rel|x_1,\ldots,x_s) = \sum_{x_i=1} \log \frac{p(X_i=1|rel)}{p(X_i=1|\overline{rel})}\frac{p(X_i=0|\overline{rel})}{p(X_i=0|rel)} + \sum_{i=1}^{s}\log\frac{p(X_i=0|rel)}{p(X_i=0|\overline{rel})} + \log O(rel) \quad (9)$$

The Binary Independence Model supposes that relevance feedback information is available and that the probabilities in (9) can be estimated from the set of documents judged relevant and non-relevant:

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

giving,

$$\log O(rel|x_1,\ldots,x_s) = \sum_{x_i=1} \log \frac{\xi_i(1-\bar{\xi}_i)}{\bar{\xi}_i(1-\xi_i)} + \sum_{i=1}^{s}\log\frac{1-\xi_i}{1-\bar{\xi}_i} + \log O(rel) \quad (10)$$

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, has been called "weight-of-evidence" by Good [Goo60, Goo50]. 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:

$$\log \frac{\xi_i(1-\bar{\xi}_i)}{\bar{\xi}_i(1-\xi_i)} \quad (11)$$

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.

2.2 Croft & Harper 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, (11), as:

$$\sum_{x_i=1}^{s} \log \frac{p(X_i=1|rel)(1-p(X_i=1|rel))}{p(X_i=1|\overline{rel})(1-p(X_i=1|\overline{rel}))} = \sum_{x_i=1}^{s}\log\frac{p(X_i=1|rel)}{1-p(X_i=1|rel)} + \sum_{x_i=1}^{s}\log\frac{1-p(X_i=1|\overline{rel})}{p(X_i=1|\overline{rel})} \quad (12)$$

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 (12) is 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:

$$C \cdot \sum_{x_i=1}^{s} 1 + \sum_{x_i=1}^{s}\log\frac{N-n_i}{n_i} \quad (13)$$

3 The BIM-MAXENT Retrieval Model

In this section, we derive a retrieval model based on the Principle of Maximum Entropy. 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.

3.1 Basic BIM-MAXENT Model

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

$$H(p) = \sum_{w \in \Omega} p(\omega) \log p(\omega) \tag{14}$$

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 (14) is taken over all possible (binary) assignments $(x_1, \ldots, x_s, r)$ to $(X_1, \ldots, X_s, R)$.

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:

$$p(X_i = 1 | R = 0) = \bar{\xi}_i \qquad i = 1, \ldots, s \tag{15}$$

$$p(X_i = 1 | R = 1) = \xi_i \qquad i = 1, \ldots, s \tag{16}$$

$$p(R = 1) = \rho \tag{17}$$

The constraints given in (15) and (16) 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 (8). However, it is not needed for the purposes of ranking. We include constraint (17) 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 f_{xr}'s of the elementary

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

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

$$
\begin{aligned}
E[\bar{g}_i(\omega)] &= \bar{G}_i \equiv \bar{\xi}_i \cdot (1 - \rho) \qquad i = 1, \ldots, s \qquad (18) \\
E[g_i(\omega)] &= G_i \equiv \xi_i \cdot \rho \qquad i = 1, \ldots, s \qquad (19) \\
E[g_R(\omega)] &= G_R \equiv \rho \qquad (20)
\end{aligned}
$$

In (20) we constrain the probability $p(R = 1)$ to ρ directly,

$$
E[g_R(\omega)] = \rho \quad \textit{iff} \quad E[R] = \rho \quad \textit{iff} \quad p(R = 1) = \rho
$$

since the expected value of a binary variable is simply the probability that the variable equals 1. In (18), we are effectively constraining $p(X_i = 1|R = 0)$ to $\bar{\xi}_i$, since

$$
\begin{aligned}
E[\bar{g}_i(\omega)] = \bar{\xi}_i(1 - \rho) \quad & \textit{iff} \quad p(X_i = 1, R = 0) &&= \bar{\xi}_i \cdot (1 - \rho) \\
& \textit{iff} \quad p(X_i = 1|R = 0) \cdot p(R = 0) &&= \bar{\xi}_i \cdot (1 - \rho) \\
& \textit{iff} \quad p(X_i = 1|R = 0) &&= \bar{\xi}_i
\end{aligned}
$$

The last step follows because, having constrained $p(R = 1)$ to ρ, we have constrained $p(R = 0)$ to $1 - \rho$. Similarly, (19) effects the desired constraint on $p(X_i = 1|R = 1)$:

$$
\begin{aligned}
E[g_i(\omega)] = \xi\rho \quad & \textit{iff} \quad p(X_i = 1, R = 1) &&= \xi_i \cdot \rho \\
& \textit{iff} \quad p(X_i = 1|R = 1) \cdot p(R = 1) &&= \xi_i \cdot \rho \\
& \textit{iff} \quad p(X_i = 1|R = 1) &&= \xi_i
\end{aligned}
$$

probability of an arbitrary event: To maximize the entropy subject to these constraints, we apply the Lagrange method of undetermined multipliers [Chi67]. 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, (18)–(20), is transformed into the maximization of the unconstrained function:

$$
\begin{aligned}
H'(p) = \sum_{\omega \in \Omega} p(\omega) \log p(\omega) \quad & + \quad \lambda'_0 (1 - \sum_{\omega \in \Omega} p(\omega)) \\
& + \quad \bar{\lambda}_1 (\bar{G}_1 - \sum_{\omega \in \Omega} p(\omega) \bar{g}_1(\omega)) \; + \; \ldots \; + \; \bar{\lambda}_s (\bar{G}_s - \sum_{\omega \in \Omega} p(\omega) \bar{g}_s(\omega)) \\
& + \quad \lambda_1 (G_1 - \sum_{\omega \in \Omega} p(\omega) g_1(\omega)) \; + \; \ldots \; + \; \lambda_s (G_s - \sum_{\omega \in \Omega} p(\omega) g_s(\omega)) \\
& + \quad \lambda_R (G_R - \sum_{\omega \in \Omega} p(\omega) g_R(\omega))
\end{aligned} \qquad (21)
$$

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, gives:

$$
\begin{aligned}
\tfrac{\partial}{\partial p(\omega)} H' \quad = \quad 1 + \log p(\omega) \quad - \quad \lambda'_0 \quad & - \quad \bar{\lambda}_1 \bar{g}_1(\omega) - \ldots - \bar{\lambda}_s \bar{g}_s(\omega) \\
& - \quad \lambda_1 g_1(\omega) - \ldots - \lambda_s g_s(\omega) \\
& - \quad \lambda_R
\end{aligned}
$$

Using λ_0 for $\lambda'_0 - 1$ and setting the derivatives (one for each ω) equal to zero, we get:

where 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 [Tri69]) that this solution will always be, not only a maximum, but a global maximum for the entropy.

3.2 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:

$$\begin{aligned} O(rel|x_1, \ldots, x_s) &= \frac{p(rel|x_1, \ldots, x_s)}{p(\overline{rel}|x_1, \ldots, x_s)} = \frac{p(x_1, \ldots, x_s, rel)/p(x_1, \ldots, x_s)}{p(x_1, \ldots, x_s, \overline{rel})/p(x_1, \ldots, x_s)} = \frac{p(x_1, \ldots, x_s, rel)}{p(x_1, \ldots, x_s, \overline{rel})} \\ &= \frac{e^{\left(\lambda_0 + \sum_{i=1}^{s} \lambda_i x_i + \lambda_R\right)}}{e^{\left(\lambda_0 + \sum_{i=1}^{s} \bar{\lambda}_i x_i\right)}} = e^{\left[\left(\sum_{i=1}^{s} (\lambda_i - \bar{\lambda}_i) x_i\right) + \lambda_R\right]} \end{aligned}$$

Therefore, the log-odds of relevance is given by:

$$\log O(rel|x_1, \ldots, x_s) = \left(\sum_{i=1}^{s} (\lambda_i - \bar{\lambda}_i) x_i\right) + \lambda_R = \left(\sum_{x_i=1} (\lambda_i - \bar{\lambda}_i)\right) + \lambda_R \tag{23}$$

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.

3.3 Characteristics of the BIM-MAXENT Distribution

To prepare for the determination of the values for the Lagrange multipliers, we shall find it convenient to derive closed form solutions for the following probabilities:

- $O(X_i = 1|R = 1)$: the odds of term occurrence given relevance.
- $O(X_i = 1|R = 0)$: the odds of term occurrence given non-relevance.
- $O(R = 1)$: the prior odds of relevance.

odds of term occurrences given relevance: Using the formulation developed for each term in the previous section, the odds of occurrence conditioned on relevance can be determined. For the sake of concreteness we develop the odds for the occurrence of the first term. For an arbitrary assignment, $x_2, \ldots, x_s$, of values to $X_2, \ldots, X_s$:

$$\frac{p(1, x_2, \ldots, x_s, rel)}{p(0, x_2, \ldots, x_s, rel)} = \frac{e^{\left[\lambda_0 + \lambda_1 + \left(\sum_{i=2}^{s} \lambda_i x_i\right) + \lambda_R\right]}}{e^{\left[\lambda_0 + \left(\sum_{i=2}^{s} \lambda_i x_i\right) + \lambda_R\right]}} = e^{\lambda_1}$$

We see that e^{λ_1} expresses how many times more likely we are to find an occurrence, as opposed to a non-occurrence, of the first term, in a relevant document that has an occurrence configuration for the remaining terms of $(x_2, \ldots, x_s)$. That is,

Since this is the case for an arbitrary configuration, $(x_2, \ldots, x_s) \in \{0,1\}^{s-1}$, we may sum over all possible configurations:

$$\begin{aligned} \sum_{x_2,\ldots,x_s} p(1, x_2, \ldots, x_s, rel) &= \sum_{x_2,\ldots,x_s} e^{\lambda_1}\, p(0, x_2, \ldots, x_s, rel) \\ &= e^{\lambda_1} \sum_{x_2,\ldots,x_s} p(0, x_2, \ldots, x_s, rel) \end{aligned}$$

which is to say:

$$p(X_1 = 1, R = rel) = e^{\lambda_1}\, p(X_1 = 0, R = rel)$$

and therefore:

$$\begin{aligned} O(X_1 = 1|rel) &= \frac{p(X_1 = 1|R = rel)}{p(X_1 = 0|R = rel)} = \frac{p(X_1 = 1, R = rel)/p(R = rel)}{p(X_1 = 0, R = rel)/p(R = rel)} = \frac{p(X_1 = 1, R = rel)}{p(X_1 = 0, R = rel)} \\ &= e^{\lambda_1} \end{aligned}$$

What we have shown for $O(X_1 = 1|rel)$ holds equally well for all X_i:

$$O(X_i = 1|rel) = \frac{p(X_i = 1, R = rel)}{p(X_i = 0, R = rel)} = e^{\lambda_i} \qquad i = 1, \ldots, s \tag{24}$$

odds of term occurrences given non-relevance: The analysis of term occurrences given non-relevance is very similar to that for relevance:

$$\frac{p(1, x_2, \ldots, x_s, \overline{rel})}{p(0, x_2, \ldots, x_s, \overline{rel})} = \frac{e^{\left[\lambda_0 + \bar{\lambda}_1 + \left(\sum_{i=2}^{s} \bar{\lambda}_i x_i\right)\right]}}{e^{\left[\lambda_0 + \left(\sum_{i=2}^{s} \bar{\lambda}_i x_i\right)\right]}} = e^{\bar{\lambda}_1}$$

which leads to:

$$O(X_i = 1|\overline{rel}) = e^{\bar{\lambda}_i} \qquad i = 1, \ldots, s \tag{25}$$

prior odds of relevance: We will derive the prior odds of relevance in terms of the probabilities of relevance and non-relevance. In order to derive a closed form for the probability of relevance from the formula given in (22), we can sum over all elementary events for which $R = 1$:

$$\begin{aligned} p(rel) &= \sum_{x_1,\ldots,x_s} p(x_1, \ldots, x_s, rel) \\ &= \sum_{x_1,\ldots,x_s} e^{\left[\lambda_0 + \left(\sum_{i=1}^{s} (\bar{\lambda}_i \bar{r} + \lambda_i r) x_i\right) + \lambda_R r\right]} = e^{(\lambda_0 + \lambda_R)} \sum_{x_1,\ldots,x_s} e^{\left[\sum_{i=1}^{s} \lambda_i x_i\right]} \end{aligned}$$

Considering that for each configuration $(x_2, \ldots, x_s)$ there are two terms in the summation, one with $x_1 = 0$ and one with $x_1 = 1$, we can write:

$$\begin{aligned} p(rel) &= e^{(\lambda_0 + \lambda_R)} \sum_{x_2,\ldots,x_s} \left(e^{\left[\lambda_1 + \sum_{i=2}^{s} \lambda_i x_i\right]} + e^{\left[\sum_{i=2}^{s} \lambda_i x_i\right]} \right) \\ &= e^{(\lambda_0 + \lambda_R)} (e^{\lambda_1} + 1) \sum_{x_2,\ldots,x_s} e^{\left[\sum_{i=2}^{s} \lambda_i x_i\right]} \end{aligned}$$

Applying this same reasoning to each of the remaining x_i, in turn:

$$\begin{aligned} p(rel) &= e^{(\lambda_0 + \lambda_R)} (e^{\lambda_1} + 1)(e^{\lambda_2} + 1) \sum_{x_3,\ldots,x_s} e^{\left[\sum_{i=3}^{s} \lambda_i x_i\right]} \\ &\vdots \\ &= e^{(\lambda_0 + \lambda_R)} \prod_{i=1}^{s} (e^{\lambda_i} + 1) \end{aligned} \tag{26}$$

Similar analysis for the probability of non-relevance yields:

$$\begin{aligned} p(\overline{rel}) &= \sum_{x_1,\ldots,x_s} p(x_1,\ldots,x_s,\overline{rel}) = e^{\lambda_0} \sum_{x_1,\ldots,x_s} e^{\left[\sum_{i=1}^{s} \bar{\lambda}_i x_i\right]} \\ &= e^{\lambda_0} \prod_{i=1}^{s} (e^{\bar{\lambda}_i} + 1) \end{aligned} \tag{27}$$

The probabilities of relevance and non-relevance given in (26) & (27) can now be combined to give the odds of relevance:

$$\begin{aligned} O(rel) &= \frac{p(rel)}{p(\overline{rel})} = \frac{e^{(\lambda_0+\lambda_R)} \prod_{i=1}^{s} (e^{\lambda_i} + 1)}{e^{\lambda_0} \prod_{i=1}^{s} (e^{\bar{\lambda}_i} + 1)} \\ &= e^{\lambda_R} \prod_{i=1}^{s} \frac{(e^{\lambda_i} + 1)}{(e^{\bar{\lambda}_i} + 1)} \end{aligned} \tag{28}$$

3.4 Parameter Values for BIM-MAXENT

The Lagrange multipliers introduced in (21) become parameters of the probability distribution derived in (22). If we are to derive a specific distribution, we must determine the values of these parameters. While the derivation of the distribution was based on the form of the expressions to be constrained, as given in (18)–(20), the actual values to which they must be constrained have yet to play their role. We now turn our attention to these values.

the value of $\lambda_1 = \ldots = \lambda_s$: From (24), we have $\lambda_i = \log O(X_i = 1|rel)$, but $p(X_i = 1|rel)$ has been constrained to ξ. Equivalently, $p(X = 0|rel)$ has been constrained to $1 - \xi$, and therefore:

$$\begin{aligned} \lambda_i &= \log O(X_i = 1|rel) = \log \frac{p(X_i = 1|rel)}{p(X_i = 0|rel)} \\ &= \log \frac{\xi_i}{1 - \xi_i} \end{aligned} \tag{29}$$

the value of $\bar{\lambda}_1, \ldots, \bar{\lambda}_s$: Similarly, from (25),

$$\begin{aligned} \bar{\lambda}_i &= \log O(X_i = 1|\overline{rel}) = \log \frac{p(X_i = 1|\overline{rel})}{p(X_i = 0|\overline{rel})} \\ &= \log \frac{\bar{\xi}_i}{1 - \bar{\xi}_i} \end{aligned} \tag{30}$$

the value of λ_R: The prior, $p(rel)$, has been constrained to ρ, which is equivalent to $O(rel)$ being constrained to $\frac{\rho}{1-\rho}$. Combining this with the expression for $O(rel)$ given in (28):

$$\begin{aligned} e^{\lambda_R} \prod_{i=1}^{s} \frac{e^{\lambda_i} + 1}{e^{\bar{\lambda}_i} + 1} &= \frac{\rho}{1 - \rho} \\ e^{\lambda_R} &= \frac{\rho}{1 - \rho} \prod_{i=1}^{s} \frac{e^{\bar{\lambda}_i} + 1}{e^{\lambda_i} + 1} \\ \lambda_R &= \log \left(\frac{\rho}{1 - \rho} \prod_{i=1}^{s} \frac{e^{\bar{\lambda}_i} + 1}{e^{\lambda_i} + 1} \right) = \log \frac{\rho}{1 - \rho} + \sum_{i=1}^{s} \log \frac{e^{\bar{\lambda}_i} + 1}{e^{\lambda_i} + 1} \end{aligned}$$

This together with the expressions for λ_i and $\bar{\lambda}_i$, derived in (29) and (30), gives:

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:

$$\log O(rel|0,\ldots,0) = \log \frac{p(0,\ldots,0,rel)}{p(0,\ldots,0,\overline{rel})} = \log \frac{e^{[\lambda_0+\lambda_R]}}{e^{\lambda_0}} = \lambda_R \tag{32}$$

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

3.5 BIM-MAXENT Ranking Formula – reprise

Substituting the values of the parameters derived in (29), (30), (31) for the conditional log-odds of relevance given in (23), we have:

$$\begin{aligned}
\log O(rel|x_1,\ldots,x_s) &= \left(\sum_{x_i=1}(\lambda_i - \bar{\lambda}_i)\right) + \lambda_R \\
&= \left(\sum_{x_i=1}\left((\log\frac{\xi_i}{1-\xi_i} - \log\frac{\bar{\xi}_i}{1-\bar{\xi}_i})\right)\right) + \log\frac{\rho}{1-\rho} + \sum_{i=1}^{s}\log\frac{1-\xi_i}{1-\bar{\xi}_i} \\
&= \left(\sum_{x_i=1}\log\frac{\xi_i(1-\bar{\xi}_i)}{\bar{\xi}_i(1-\xi_i)}\right) + \left(\sum_{i=1}^{s}\log\frac{1-\xi_i}{1-\bar{\xi}_i}\right) + \log\frac{\rho}{1-\rho}
\end{aligned} \tag{33}$$

This is the ranking formula, (10), of the Binary Independence Model.

3.6 Discussion of the BIM-MAXENT Model

To summarize the development presented in this section:

- we have imposed the set of constraints (18- 20),

$$\begin{aligned}
E[\bar{g}_i(\omega)] &= \bar{G}_i \equiv \bar{\xi}_i \cdot (1-\rho) \qquad i = 1,\ldots,s \\
E[g_i(\omega)] &= G_i \equiv \xi_i \cdot \rho \qquad i = 1,\ldots,s \\
E[g_R(\omega)] &= G_R \equiv \rho
\end{aligned}$$

- it was shown in (22) that for the maximum entropy distribution subject to these constraints, the probability of an arbitrary event, is given by:

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

- for this distribution, the log-odds of relevance conditioned on a given term occurrence pattern was found in (23) to be:

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

- by applying the constraint values ξ_i, $\bar{\xi}_i$, ρ, values for the parameters were determined in (29– 31) as:

$$\begin{aligned}
\lambda_i &= \log\frac{\xi_i}{1-\xi_i} \\
\bar{\lambda}_i &= \log\frac{\bar{\xi}_i}{1-\bar{\xi}_i} \\
\lambda_R &= \log\frac{\rho}{1-\rho} + \sum_{i=1}^{s}\log\frac{1-\xi_i}{1-\bar{\xi}_i}
\end{aligned}$$

- finally, in (33) we have the log-odds of relevance in terms of the constraint values:

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

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.

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 Principle of Maximum Entropy. 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.

For an arbitrary configuration $(x_1,\ldots,x_s) \in \{0,1\}^s$:

$$\begin{aligned}
\frac{p(x_1,\ldots,x_s|rel)}{p(x_1,\ldots,x_s|\overline{rel})} = \frac{p(x_1,\ldots,x_s,rel)/p(rel)}{p(x_1,\ldots,x_s,\overline{rel})/p(\overline{rel})} &= \frac{e^{\lambda_0+\left(\sum_{i=1}^{s}\lambda_i x_i\right)+\lambda_R}}{e^{\lambda_0+\left(\sum_{i=1}^{s}\bar{\lambda}_i x_i\right)}}/O(rel) \\
&= \frac{e^{\left(\sum_{i=1}^{s}\lambda_i x_i\right)-\left(\sum_{i=1}^{s}\bar{\lambda}_i x_i\right)+\lambda_R}}{O(rel)} \\
&= \frac{e^{\left(\sum_{i=1}^{s}(\lambda_i-\bar{\lambda}_i)x_i\right)}e^{\lambda_R}}{O(rel)} \qquad (34)
\end{aligned}$$

Recalling the expression for $O(rel)$ derived in (28), we have:

$$\frac{p(x_1,\ldots,x_s|rel)}{p(x_1,\ldots,x_s|\overline{rel})} = \frac{e^{\left(\sum_{i=1}^{s}(\lambda_i-\bar{\lambda}_i)x_i\right)}e^{\lambda_R}}{e^{\lambda_R}\prod_{i=1}^{s}\frac{(e^{\lambda_i}+1)}{(e^{\bar{\lambda}_i}+1)}} = e^{\left(\sum_{i=1}^{s}(\lambda_i-\bar{\lambda}_i)x_i\right)} \cdot \prod_{i=1}^{s}\frac{(e^{\bar{\lambda}_i}+1)}{(e^{\lambda_i}+1)} \qquad (35)$$

On the other hand, the odds of each individual term occurrence conditioned on relevance was found in (24) to be e^{λ_i}. Similarly, the odds of term occurrence conditioned on non-relevance was found in (25) to be $e^{\bar{\lambda}_i}$. Therefore,

$$\begin{aligned}
p(X_i=1|rel) &= \frac{O(X_i=1|rel)}{1+O(X_i=1|rel)} = \frac{e^{\lambda_i}}{1+e^{\lambda_i}} \\
p(X_i=0|rel) &= 1-p(X_i=1|rel) = \frac{1}{1+e^{\lambda_i}} \\
p(X_i=1|\overline{rel}) &= \frac{O(X_i=1|\overline{rel})}{1+O(X_i=1|\overline{rel})} = \frac{e^{\bar{\lambda}_i}}{1+e^{\bar{\lambda}_i}} \\
p(X_i=0|\overline{rel}) &= 1-p(X_i=1|\overline{rel}) = \frac{1}{1+e^{\bar{\lambda}_i}}
\end{aligned}$$

These equations may be summarized as:

$$p(X_i = x_i|rel) = \frac{e^{\lambda_i x_i}}{1 + e^{\lambda_i}}$$

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

and so,

$$\begin{aligned}
\prod_{i=1}^{s} \frac{p(X_i = x_i|rel)}{p(X_i = x_i|\overline{rel})} &= \prod_{i=1}^{s} \left(\frac{e^{\lambda_i x_i}}{1 + e^{\lambda_i}} / \frac{e^{\bar{\lambda}_i x_i}}{1 + e^{\bar{\lambda}_i}} \right) \\
&= \left(\prod_{i=1}^{s} e^{(\lambda_i - \bar{\lambda}_i)x_i} \right) \left(\prod_{i=1}^{s} \frac{1 + e^{\bar{\lambda}_i}}{1 + e^{\lambda_i}} \right) \\
&= e^{\left(\sum_{i=1}^{s} (\lambda_i - \bar{\lambda}_i)x_i\right)} \left(\prod_{i=1}^{s} \frac{1 + e^{\bar{\lambda}_i}}{1 + e^{\lambda_i}} \right) \qquad (36)
\end{aligned}$$

This is equivalent to the expression given in (35), from which we may conclude 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 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.

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 (for all intents and purposes non-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, (18) & (19), 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:

$$O(rel) = \prod_{i=1}^{s} \frac{(e^{\lambda_i} + 1)}{(e^{\bar{\lambda}_i} + 1)} = \prod_{i=1}^{s} \frac{(e^{\log \frac{\xi_i}{1-\xi_i}} + 1)}{(e^{\log \frac{\bar{\xi}_i}{1-\bar{\xi}_i}} + 1)} = \prod_{i=1}^{s} \frac{\frac{\xi_i}{1-\xi_i} + 1}{\frac{\bar{\xi}_i}{1-\bar{\xi}_i} + 1} = \prod_{i=1}^{s} \frac{1 - \bar{\xi}_i}{1 - \xi_i} \qquad (37)$$

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:

$$O(rel|0, \ldots, 0) = \frac{e^{\lambda_0}}{e^{\lambda_0}} = 1 \qquad (38)$$

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 (20). 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:

$$O(rel|0,\ldots,0) \quad = \quad e^{\lambda_R} = \frac{\rho}{1-\rho}\prod_{i=1}^{s}\frac{e^{\bar{\lambda}_i}+1}{e^{\lambda_i}+1} = \frac{\rho}{1-\rho}\prod_{i=1}^{s}\frac{1-\xi_i}{1-\bar{\xi}_i} \tag{39}$$

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.

4 The CM-MAXENT Retrieval Model

In the previous section, we developed a model based on the PME from which we were able to derive the same ranking formula that results from the Binary Independence Model. In this section, we derive a maximum entropy retrieval model that will be constrained in such a way as to be consistent with the assumptions made in the Combination Match Model (CMM) of Croft and Harper. CMM adapts the Binary Independence Model to situations where no relevance information is available. Our goal, here, is to reproduce the CMM ranking formula.

4.1 Basic CM MAXENT Model

The development of this model will be very similar to that of the BIM-MAXENT model. The first and third BIM-MAXENT constraints, concerning the probability of term occurrence conditioned on non-relevance and the prior probability of relevance will be left as they were. The second constraint concerning the probability of term occurrence in relevant documents will be eliminated. In its place, the number of query terms expected to appear in a relevant document will be constrained.

Formally, the constraints will be:

$$p(X_i = 1|R = 0) = \bar{\xi}_i \qquad i = 1, \ldots, s \tag{40}$$

$$E(X_\#|R = 1) = \zeta \qquad \text{where: } X_\# = \sum_{i-1}^{s} X_i \tag{41}$$

$$p(R = 1) = \rho \tag{42}$$

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.

For this model, we will be concerned with the following features of the elementary events:

$$\bar{g}_i(\omega) = \left\{ \begin{array}{ll} 1 & \text{if } X_i(\omega) = 1 \wedge R(\omega) = 0 \\ 0 & \text{otherwise} \end{array} \right\} \qquad i = 1, \ldots, s \tag{43}$$

$$g_\#(\omega) = \left\{ \begin{array}{ll} (X_1 + \cdots + X_s) & \text{if } R(\omega) = 1 \\ 0 & \text{otherwise} \end{array} \right\} \tag{44}$$

$$g_R(\omega) = R(\omega) \tag{45}$$

which will be constrained by:

$$E[\bar{g}_i(\omega)] = \bar{G}_i = \bar{\xi}_i \cdot (1 - \rho) \qquad i = 1, \ldots, s \tag{46}$$

$$E[g_\#(\omega)] = G_\# = \zeta \cdot \rho \tag{47}$$

$$E[g_R(\omega)] = G_R = \rho \tag{48}$$

In (47), we are effectively constraining $E(X_1 + \cdots + X_s|R = 1)$ to ζ, since:

$$\begin{aligned} E[g_\#(\omega)] = \sum_{\omega \in \Omega} p(\omega) g_\#(\omega) &= \sum_{R(\omega)=1} p(\omega)(X_1(\omega) + \ldots + X_s(\omega)) \\ &= \sum_{R(\omega)=1} p(R = 1) \cdot p(\omega|R = 1)(X_1(\omega) + \ldots + X_s(\omega)) \\ &= p(R = 1) \cdot E[X_1 + \ldots + X_s|R = 1] \end{aligned}$$

and, because this has been constrained to $\zeta \cdot \rho$ and $p(R = 1) = E[R]$ has been constrained to ρ,

$$E[X_1 + \ldots + X_s|R = 1] = \zeta$$

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

$$p(\omega) = e^{\left[\lambda_0+\left(\sum_{i=1}^{s}\bar{\lambda}_i\bar{r}x_i\right)+\lambda_{\#}x_{\#}r+\lambda_R r\right]} \tag{49}$$

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

Based on the probability distribution, (49), we can determine the odds of relevance given a specific occurrence pattern:

$$O(rel|x_1,\ldots,x_s) = \frac{e^{\left[\lambda_0+\lambda_{\#}x_{\#}+\lambda_R\right]}}{e^{\left[\lambda_0+\left(\sum_{i=1}^{s}\bar{\lambda}_i x_i\right)\right]}} = e^{\lambda_{\#}x_{\#}+\lambda_R-\sum_{i=1}^{s}\bar{\lambda}_i x_i} \tag{50}$$

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

$$\begin{aligned}\log O(rel|x_1,\ldots,x_s) &= \lambda_{\#}x_{\#} - \sum_{i=1}^{s}\bar{\lambda}_i x_i + \lambda_R \\ &= \lambda_{\#}x_{\#} - \sum_{x_i=1}\bar{\lambda}_i + \lambda_R \end{aligned} \tag{51}$$

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

4.2 Characteristics of the CM-MAXENT Distribution

As with the BIM-MAXENT model, we find it convenient to derive closed form solutions for the odds of certain events. The reasoning employed here closely follows that of Section 3.

odds of term occurrences given relevance: For arbitrary values $x_2,\ldots,x_s$:

$$\begin{aligned} p(1,x_2,\ldots,x_s,rel) &= e^{\left[\lambda_0+\left(\sum_{i=1}^{s}\bar{\lambda}_i\bar{r}x_i\right)+\lambda_{\#}\left(\sum_{i=1}^{s}x_i\right)r+\lambda_R r\right]} \\ &= e^{\left[\lambda_0+\lambda_{\#}\left(1+\sum_{i=2}^{s}x_i\right)+\lambda_R\right]} \\ &= e^{\lambda_{\#}}\cdot e^{\left[\lambda_0+\lambda_{\#}\left(\sum_{i=2}^{s}x_i\right)+\lambda_R\right]} \\ &= e^{\lambda_{\#}}\cdot p(0,x_2,\ldots,x_s,rel) \end{aligned}$$

Summing over all possible values for $(x_2,\ldots,x_s)\in\{0,1\}^{s-1}$:

$$\begin{aligned} p(X_1=1,R=1) &= \sum_{x_2,\ldots,x_s} p(1,x_2,\ldots,x_s,rel) = \sum_{x_2,\ldots,x_s} e^{\lambda_{\#}}\cdot p(0,x_2,\ldots,x_s,rel) \\ &= e^{\lambda_{\#}}\cdot p(X_1=0,R=1) \end{aligned}$$

which generalizes to arbitrary query terms, giving the conditional odds of occurrence for term i as:

$$O(X_i=1|rel) = e^{\lambda_{\#}} \tag{52}$$

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.

odds of term occurrences given non-relevance: The odds of a query term occurring in a non-relevant document are the same as for BIM-MAXENT. For term 1, we have,

$$O(X_1=1|\overline{rel}) = \frac{p(1,x_2,\ldots,x_s,\overline{rel})}{p(0,x_2,\ldots,x_s,\overline{rel})} = \frac{e^{\left[\lambda_0+\bar{\lambda}_1+\left(\sum_{i=2}^{s}\bar{\lambda}_i x_i\right)\right]}}{e^{\left[\lambda_0+\left(\sum_{i=2}^{s}\bar{\lambda}_i x_i\right)\right]}} = e^{\bar{\lambda}_1} \tag{53}$$

which generalizes to arbitrary query terms.

prior odds of relevance: We will derive the prior odds of relevance in terms of the probabilities of relevance and non-relevance. In order to derive a closed form for the probability of relevance from the formula given in (49), we can sum over all elementary events for which $R = 1$:

$$\begin{aligned} p(rel) &= \sum_{x_1,\ldots,x_s} p(x_1,\ldots,x_s,rel) \\ &= \sum_{x_1,\ldots,x_s} e^{\left[\lambda_0+\left(\sum_{i=1}^{s}(\bar{\lambda}_i \bar{r} x_i)\right)+\lambda_{\#} x_{\#} r+\lambda_R r\right]} = e^{(\lambda_0+\lambda_R)} \sum_{x_1,\ldots,x_s} e^{\lambda_{\#} x_{\#}} \end{aligned}$$

This can be written as:

$$\begin{aligned} p(rel) &= e^{(\lambda_0+\lambda_R)} \sum_{x_1,\ldots,x_s} e^{\left[\sum_{i=1}^{s}\lambda_{\#} x_i\right]} \\ &= e^{(\lambda_0+\lambda_R)} \sum_{x_2,\ldots,x_s} \left(e^{\left[\lambda_{\#}+\sum_{i=2}^{s}\lambda_{\#} x_i\right]} + e^{\left[\sum_{i=2}^{s}\lambda_{\#} x_i\right]}\right) \\ &= e^{(\lambda_0+\lambda_R)}(e^{\lambda_{\#}}+1) \sum_{x_2,\ldots,x_s} e^{\left[\sum_{i=2}^{s}\lambda_{\#} x_i\right]} \end{aligned}$$

Applying the same reasoning for each x_i:

$$\begin{aligned} p(rel) &= e^{(\lambda_0+\lambda_R)}(e^{\lambda_{\#}}+1)^2 \sum_{x_3,\ldots,x_s} e^{\left[\sum_{i=3}^{s}\lambda_{\#} x_i\right]} \\ &\vdots \\ &= e^{(\lambda_0+\lambda_R)}(e^{\lambda_{\#}}+1)^s \end{aligned} \tag{54}$$

The probability of non-relevance is as before:

$$p(\overline{rel}) = e^{\lambda_0} \prod_{i=1}^{s}(e^{\bar{\lambda}_i}+1) \tag{55}$$

giving, for the odds of relevance:

$$O(rel) = e^{\lambda_R} \prod_{i=1}^{s} \frac{(e^{\lambda_{\#}}+1)}{(e^{\bar{\lambda}_i}+1)} \tag{56}$$

4.3 CH MAXENT Ranking Formula – reprise

In (51) above, the following expression for the log-odds of relevance was derived:

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

From (53) and constraint (40), 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:

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

giving the formula:

$$\begin{aligned}\log O(rel|x_1,\ldots,x_s) &= \lambda_{\#}\cdot x_{\#} - \left(\sum_{x_i=1}\log\frac{n_i}{N-n_i}\right)+\lambda_R \\ &= \lambda_{\#}\cdot x_{\#} + \left(\sum_{x_i=1}\log\frac{N-n_i}{n_i}\right)+\lambda_R \end{aligned} \quad (57)$$

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.

4.4 Discussion of the CM-MAXENT Model

In the foregoing sections we have seen that 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 derive the probability distribution (49),

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

which leads to the CMM for document ranking,

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

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.

A MAXENT Version of *idf* Weighting: If we eliminate constraint (47) respecting the expected value of the number of terms to be found in a relevant document, the probability of an arbitrary event would be:

$$p(\omega) = e^{\left[\lambda_0+\left(\sum_{i=1}^{s}\bar{\lambda}_i\bar{x}_i\right)+\lambda_R r\right]} \quad (58)$$

and the conditional log-odds of relevance would be:

$$\log O(rel|x_1,\ldots,x_s) = \left(\sum_{x_i=1}-\bar{\lambda}_i\right)+\lambda_R$$

For all occurrence patterns $(x_2,\ldots,x_s)\in\{0,1\}^{s-1}$:

$$p(1,x_2,\ldots,x_s,rel) = p(0,x_2,\ldots,x_s,rel) = e^{[\lambda_0+\lambda_R]}$$

and therefore the odds of occurrence of the first term given relevance are even, which can be generalized to arbitrary terms:

$$O(X_i=1|rel)=1 \quad (59)$$

The odds of term occurrence given non-relevance would be the same as before:

$$O(X_i = 1|\overline{rel}) = e^{\bar{\lambda}_i} \tag{60}$$

and the prior probability of relevance would be:

$$p(rel) = \sum_{x_1,\ldots,x_s} e^{[\lambda_0+\lambda_R]} = 2^s \cdot e^{[\lambda_0+\lambda_R]}$$

Combining this with the prior probability of non-relevance, which is the same as before (55), gives:

$$O(rel) = \frac{2^s \cdot e^{[\lambda_0+\lambda_R]}}{e^{\lambda_0}\prod_{i=1}^{s}(e^{\bar{\lambda}_i}+1)} = e^{\lambda_R} \cdot \prod_{i=1}^{s}\frac{2}{(e^{\bar{\lambda}_i}+1)}$$

This leads to parameter values of:

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

$$\lambda_R = \log\frac{\rho}{1-\rho} + \sum_{i=1}^{s}\log\frac{2}{1-\bar{\xi}_i}$$

And finally the ranking formula:

$$\log O(rel|x_1,\ldots,x_s) = \left(\sum_{x_i=1}\log\frac{1-\bar{\xi}_i}{\bar{\xi}_i}\right) + \left(\sum_{i=1}^{s}\log\frac{2}{1-\bar{\xi}_i}\right) + \log\frac{\rho}{1+\rho} \tag{61}$$

Since the two terms at the right are constant over all documents, (61) 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 interpretted as the maximum entropy distribution constrained only so that $p(x_i \mid \overline{rel}) = \bar{\xi}_i$.

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 (46) the following properties of the MAXENT distribution would hold:

$$p(\omega) = e^{[\lambda_0+\lambda_{\#}x_{\#}r+\lambda_R r]} \tag{62}$$

giving conditional log-odds of relevance:

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

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.

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 3.6 hold for the CMM. Also, we have seen that the odds of occurrence in a relevant document is

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

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.

The $E[g_{\#}(\omega)]$ Constraint: In Section 3.6 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 interpretted 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 Discussion

We have shown that 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 Probabilistic Ranking Principle.

Probability Ranking Principle: In [Rob77], 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 estimation 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 [DF82, MO95, Daw89] 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 possesed by the system; 4) the Maximum Entropy Principle is a very reasonable candidate.

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 [vR77, HvR78, SvR83]. 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:

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

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.

Our goal in this paper has not been to develop a new model. The goal has explicitly been to reproduce BIM and CMM. We propose the MAXENT formalism as a clearer conceptualization of the thought processes and research posture underlying these models.

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*

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 Principle of Maximum Entropy 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 in this paper clarifies, we believe, the conceptual position of the original authors.

Flexibility of Constraints: An advantage of the MAXENT approach is that it naturally accomodates 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 knowlege 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.

6 Summary

In this paper, we have adopted a probabilistic attitude with respect to information retrieval, where *probability* is understood as the system's judgment that a document will be relevant based on all information it has available to it. We have argued that previous work is best conceptualized in this way and that a frequentist view of probability as a fraction of an existing population is untenable. If a system is to rank document according to the probability that the document is relevant to the query, it must adopt a probability distribution of relevance conditioned on the evidence it considers. Probabilities will have to be determined in the absence of total knowledge concerning all aspects of the distribution. Available knowledge constrains the distribution, but does not leave it fully determined. The Principle of Maximum Entropy provides, in our opinion, the most reasonable methodology for fully determining the otherwise underconstrained distribution.

In support of our position, we have shown how both the Binary Independence Model and Combination Match Model may be understood in terms of the PME. Although no assumptions are made, we have shown that the linked dependence assumption in the case of BIM and the assumption of equal probability of term occurrence in relevant documents in the case of CMM are consequences of the Principle of Maximum Entropy. We have seen how the PME can guide us as to how best to assign a prior probability of relevance and how both pure coordination match ranking and pure *idf* weighting can result from different ways of constraining the probability distribution.

The difference between constraints to be applied to a subjective probability distribution and assumptions concerning the characteristics of frequencies in a supposedly existing population has been emphasized, and we have argued that this difference has important philosophical and practical ramifications for research in information retrieval.

Acknowledgments

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. This material is also based on work supported in part by United States Patent and Trademark Office and Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract number F19628-95-C-0235. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsors.

References

[Bre88] G. Larry Bretthorst. Excerpts from bayesian spectrum analysis and parameter estimation. In Gary J. Erickson and C. Ray Smith, editors, *Maximum Entropy and Bayesian Methods in Science and Engineering*, pages 75–146, Norwell, MA, 1988. Kluwer Academic Publishers.

[CH79] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, December 1979.

[CH82] W. S. Cooper and P. Huizinga. The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology, Research & Development*, 1:99–112, 1982.

[Chi67] Alpha C. Chiang. *Fundamental Methods of Mathematical Economics.* McGraw-Hill, New York, 1967.

[Coo83] William S. Cooper. Exploiting the maximum entropy principle to increase retrieval effectiveness. *Journal of the American Society for Information Science*, 34(1):31–39, 1983.

[Coo91] William S. Cooper. Some inconsistentencies and misnomers in probabilistic information retrieval. In A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan, editors, *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 57–61, Chicago, Illinois, USA, October 1991.

[Daw89] A. P. Dawid. Probability forecasting. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley, New York, 1989.

[DF82] M. DeGroot and S. Feinberg. The comparison and evaluation of forecasters. *The Statistician*, 32:12–22, 1982.

[ES88] Gary J. Erickson and C. Ray Smith. *Maximum Entropy and Bayesian Methods in Science and Engineering.* Kluwer Academic Publishers, Norwell, MA, 1988.

[Fin73] Terrence L. Fine. *Theories of Probability: An Examination of Foundations.* Academic Press, New York, 1973.

[GD78] S. F. Gull and G. J. Daniell. Image reconstruction from incomplete and noisy data. *Nature*, 272:686–690, 1978.

[GJM96] Amos Golan, George G. Judge, and Doublas Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* John Wiley & Sons, New York, 1996.

[Goo50] I. J. Good. *Probability and the Weighing of Evidence.* Charles Griffin, London, 1950.

[Goo60] I. J. Good. Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society – Series B.*, 22:319–331, 1960.

[Hac65] Ian Hacking. *Logic of Statistical Inference.* Cambridge University Press, Cambridge, 1965.

[HvR78] D. J. Harper and C. J. van Rijsbergen. An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3):189–216, September 1978.

[Jay57a] E. T. Jaynes. Information theory and statistical mechanics: Part I. *Physical Review*, 106:620–630, 1957.

[Jay57b] E. T. Jaynes. Information theory and statistical mechanics: Part II. *Physical Review*, 108:171, 1957.

[Jay63] E. T. Jaynes. Information theory and statistical mechanics. In G. E. Uhlenbeck, editor, *Statistical Physics: Brandeis Summer Institute Lectures in Theoretical Physics*, volume 3 of *Brandeis Summer Institute Lectures in Theoretical Physics*, pages 182–218. W. A. Benjamin, New York, 1963.

[Jay79] E. T. Jaynes. Where do we stand on maximum entropy. In Raphael D. Levine and Myron Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118, Cambridge, Massachusetts, May 1979. MIT Press.

[Jay94] E. T. Jaynes. Probability theory: The logic of science. available via ftp://bayes.wustl.edu/pub/Jaynes/book.probability.theory/, 1994.

[Kan84] Paul B. Kantor. Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology, Research & Development*, 3(2):88–94, April 1984.

[KL86] Paul B. Kantor and Jung Jin Lee. The maximum entropy principle in information retrieval. In Fausto Rabitti, editor, *Proceedings of the 9th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, pages 269–274, Pisa, Italy, September 1986.

[KL98] Paul B. Kantor and Jung Jin Lee. Testing the maximum entropy principle for information retrieval. *Journal of the American Society for Information Science*, 49(6):557–566, 1998.

[LK91] Jung Jin Lee and Paul B. Kantor. A study of probabilistic information retrieval in the case of inconsistent expert judgments. *Journal of the American Society for Information Science*, 42(3):166–172, 1991.

[MO95] Kneale T. Marshall and Robert M. Oliver. *Decision Making and Forecasting: with Emphasis on Model Building and Policy Anal.* McGraw-Hill, New York, 1995.

[Rob77] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33:294–304, 1977.

[RS77] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1977.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423 & 623–656, 1948.

[Spa72] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[SvR83] A. F. Smeaton and C. J. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 25(3):239–246, 1983.

[SWY76] G. Salton, A. Wong, and C. T. Yu. Automatic indexing using term discrimination and term precision measurements. *Information Processing & Management*, 12:43–51, 1976.

[Tri69] Myron Tribus. *Rational Descriptions, Decisions, and Designs.* Pergamon-Hall, New York, 1969.

[Tri79] Myron Tribus. Thirty years of information theory. In Raphael D. Levine and Myron Tribus, editors, *The Maximum Entropy Formalism*, pages 1–14, Cambridge, Massachusetts, May 1979. MIT Press.

[van79] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 2 edition, 1979.

[vR77] C. J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33:106–119, 1977.