Optimum Probability Estimation from Empirical Distributions

Norbert Fuhr, Hubert Hüther
Technische Hochschule Darmstadt, Fachbereich Informatik
Karolinenplatz 5, D-6100 Darmstadt / West Germany

Probability estimation is important for the application of probabilistic models as well as for any evaluation in IR. We discuss the interdependencies between parameter estimation and certain properties of probabilistic models: dependence assumptions, binary vs. non-binary features, estimation sample selection. Then we define an optimum estimate for binary features which can be applied to various typical estimation problems in IR. A method for computing this estimate using empirical data is described. Some experiments show the applicability of our method, whereas comparable approaches are partially based on false assumptions or yield biased estimates.

1 Parameter estimation in IR

In IR the development of theoretical models and their evaluation in experiments is of equal importance: A model which cannot be evaluated (applied) is of very little use, while an evaluation can show its weaknesses and strengths and give evidence for further developments. As will be discussed below, any evaluation in IR involves some kind of parameter estimation, even for non-probabilistic models. So it is interesting to note that the problem of parameter estimation has been discussed only by a few authors ([Rijsbergen 77], [Robertson & Bovey 82], [Bookstein 83], [?]). In this paper, an attempt is made to investigate the problem of parameter estimation in a more general way, showing that there are quite similar problems in different areas within IR. We define an optimum estimate and give also an appropriate estimation method which is compared with other methods discussed in the literature.

It is well understood that IR experiments are stochastic experiments. Any evaluation within IR has to consider this fact, and thus all parameters derived from retrieval results are to be regarded as estimates with a certain bias. Especially when recall or precision values for single queries are computed, very small numerators and denominators may occur. As an example, table 1 shows data from an evaluation of the AIR retrieval test [Fuhr & Knorz 84], where a representative sample of 15 000 documents (out of $\sim$ 400 000) was used. Here 84 from the total of 309 queries have precision values of 0/0 (and 148 queries retrieved only between 1 and 10 documents). This result is not very surprising, because most of the queries with empty answer sets on our test sample would retrieve some documents from the whole data base. A similar case is reported in [Womser-Hacker 87] for the PADOK retrieval test of the German patent documentation. Without probabilistic foundation, these cases cannot be handled appropriatly: Precision values are estimates of the probability $P(relevant|retrieved)$, that is the probability that an arbitrary retrieved document is judged relevant by the user. Similarly, most other retrieval measures can (and should) be regarded as probabilities. Following this view of retrieval evaluation, the methods for parameter estimation described in this paper can be applied.

# retrieved	# relevant retrieved										
	0	1	2	3	4	5	6	7	8	9	10
0	84										
1	13	33									
2	2	10	23								
3	0	3	2	8							
4	1	0	3	3	10						
5	0	0	1	3	1	4					
6	0	1	1	0	0	1	3				
7	0	1	1	0	0	1	1	1			
8	0	0	0	1	0	1	1	4	1		
9	0	0	0	0	0	1	2	1	1	1	
10	0	0	0	0	0	0	0	0	0	0	3

Table 1: Frequency distribution (numerators and denominators of precision values) of queries from the AIR retrieval test with $0 \dots 10$ documents retrieved (manual indexing, query formulations with descriptors only).

For probabilistic information retrieval models, it is obvious that an estimation of the different parameters of the model is required for an application. Several authors have discussed the problem of estimating probabilities for document retrieval on the basis of relevance feedback data (e.g. [Rijsbergen 77], [?]). This means that in document retrieval, after submitting a request to the IR system, the user is presented with some documents. The relevance judgements of the user concerning these documents form the basis for the estimation of the parameters of a probabilistic model which is applied to the ranking of the remaining documents. Here the required probabilities are mostly of the form $P(t_i|f_j, r)$, the probability that term t_i occurs in an arbitrary document which has relevance judgement r for request f_j.

Finally, we regard another situation where parameter estimation for a probabilistic IR model plays an important role: The Darmstadt Indexing Approach ([Fuhr 89], [Biebricher et al. 88]) is suited to automatic (probabilistic) indexing with descriptors from a prescribed indexing vocabulary. It uses parameters of the form $P(s_i|t_j)$, the probability that descriptor s_i should be assigned to a document, given that it contains term t_j. These values are derived from manually indexed documents, thus building up an indexing dictionary which stores the information necessary for the automatic indexing of new documents. Although large collections of manually indexed documents (up to 400 000) have been processed for the estimation of the required parameters, most of the estimates are based on document feature frequencies less than 10 and thus have a certain bias which cannot be ignored.

2 Parameter estimation for probabilistic models

In our view, parameter estimation is the crucial point of the application of probabilistic IR models. Besides the fact that the retrieval quality observed in experiments depends, in a yet unknown extent, on the quality of the estimation procedure for the required parameters of the model being tested, there are also several interdependencies between parameter estimation and certain aspects of the model itself. In the following, we will point out some of the resulting problems. For our discussion, we will relate to the estimation of the parameters for a probabilistic retrieval model through relevance feedback. However, many of the problems mentioned below also occur in other types of probabilistic models in IR, e.g. models of probabilistic indexing.

The form of *document representation* the model relates to determines the kind of parameters to be estimated. Mainly three forms of representation have been discussed in the literature:

- Binary features are most widespread in use: Here only the presence or absence of document features, e.g. terms, is regarded. The required parameters are the probabilities that a specific feature will occur in a relevant/non-relevant document. Well known examples for models of this kind are described in [Robertson & Sparck Jones 76] and [Yu & Salton 76], but also already in [Maron & Kuhns 60] for request features.
- The within-document-frequency of terms is the basis of the two-Poisson model ([Harter 75a], [Harter 75b], [Bookstein & Swanson 74]). Here the expected frequencies of terms in relevant/non-relevant documents are the parameters to be estimated.
- Probabilistic index term weights are used in the retrieval-with-probabilistic-indexing model ([Fuhr 86], [Fuhr 88]). For the application, the expected value of the index term weights in all documents and in relevant documents are to be estimated.

From the descriptions given above, it can be seen that different kinds of representation require quite different parameters to be estimated. In the following, we will only concentrate on parameters for binary features, for several reasons: The binary case has been discussed most in the literature, the basic principles of parameter estimation are easier to explain in this case, and finally we haven't yet evolved our approach for the non-binary cases.

The *independence assumptions* underlying the model determine which and how many parameters per feature regarded are to be estimated. Many models assume statistical independence ([Robertson & Sparck Jones 76], [Yu & Salton 76], [Robertson et al. 81], [Fuhr 86], [Fuhr 88]). In order to develop models which fit better to empirical data, several modifications of the basic models incorporating term dependence information have been proposed:

- The Bahadur-Lazarsfeld-expansion (BLE) ([Lam & Yu 82], [Yu et al. 83], [Salton et al. 83]) regards all possible dependencies of term pairs, triplets etc. For its application, the BLE has to be truncated in order to limit the number of parameters to be estimated. However, in the truncated expression, negative probabilities might occur.
- In the maximum spanning tree approach described in [Rijsbergen 77] only certain dependencies of term pairs are included in the model. The term pairs are selected in such a way that they form up a tree whose edge weights, representing the mutual information betweend the nodes (=terms), are maximized.
- In the Generalized Term Dependence Model described by [Yu et al. 83], the maximum spanning tree approach is extended to include also higher order dependencies (e.g. term triplets) in the ranking formula.
- The maximum entropy principle (MEP) has been suggested as an alternate approach to exploit available dependence information in probabilistic retrieval ([Cooper & Huizinga 82], [Kantor 84], [Kantor & Lee 86]). There is similar work in artificial intelligence research dealing with uncertainty reasoning [Cheeseman 83]. This principle is regarded as making as few assumptions as possible, but incorporating all known constraints and dependencies. However, because of the computational effort required for an application of the MEP there has been no experimental evaluation of this approach in IR to date.

The crucial point of the inclusion of dependency information within a probabilistic model is that there are more parameters to be estimated, and for each parameter the amount of data available for its estimation is smaller than in the independence case. With less data, the bias of each parameter increases, and the statistical errors cumulate for the whole ranking formula of the model under consideration. This way, the advantage of incorporating dependency information within a probabilistic model can be nullified. In fact, there are no convincing experimental results up to now where dependence models would outperform independence models.

The *parameter estimation procedure* used for the application of a model is in large parts independent of the tested model itself. So any experimental evaluation of IR models does not only compare the models itself, it compares the combination of IR models with specific parameter estimation procedures. Here we want to subdivide the whole estimation procedure into two parts: the estimation sample selection and the estimation method which uses the data from the sample in order to estimate the required parameters. The latter is the main topic of this paper and will be discussed in detail in the following sections.

The selection of the estimation sample can be done in principle in two different ways: Following the definitions and assumptions of the model to be applied, in most cases it would be necessary to have a random sample of documents from which the parameters are to be derived. But this approach would require too large numbers of documents to be selected in order to provide any valuable information for the estimation method. So the only way out is to use some kind of best-first selection method, that is to apply an initial ranking formula and collect feedback information from the top ranking documents. This means that instead of $P(t_i|f_j, r)$, the probability that a term t_i occurs in an arbitrary document judged relevant/non-relevant w.r.t. request f_j, we estimate of f $P(t_i|f_j, r, sel)$, the probability that t_i occurs in an arbitrary document with judgement r which is selected by the initial ranking function. As most of the ranking formulas used for the initial ranking (e.g. coordination level match or inverse document frequency weights) prefer documents which contain the query terms, it is obvious that these estimates in general will be higher than the probabilities f $P(t_i|f_j, r)$ for which they are used. As long as this problem has not been solved, all experimental results for more sophistiated models (like e.g. the dependence models described above) are of preliminary nature. Only few attempts have been made so far to overcome this dissatisfying situation:

- In [Harper & Rijsbergen 78] the parameters relating to non-relevant documents are derived from the documents known to be non-relevant as well as from all documents not known to be relevant. Although it is not mentioned in the original paper, it seems that this approach compensates in some way the error resulting from the best-first strategy.

- A quite different approach called “linear logistic model" has been investigated in [Robertson & Bovey 82]. It is claimed that this approach does not require a random sample for parameter estimation, but the experimental results showed lower performance levels than comparable methods.

- In [Losee 87], several models for best-first document retrieval models are discussed and compared with the older models based on the assumption of random document selection. It seems that the new models are more appropriate for the usual retrieval situation where users want to see only a few relevant documents. However, it is not yet clear how the estimation procedure for these models should look like.

Despite of the problems described above, we will follow the assumption of random estimation samples in the remainder of this paper. As mentioned already in section 1, there are quite different situations in IR where parameter estimation based on random samples plays an important role. On the other hand, the models for the best-first selection strategy need further development until appropriate estimation procedures can be proposed.

3 Optimum parameter estimation

Having described some of the peculiarities with parameter estimation for probabilistic IR models, we will now give a definition for an optimum parameter estimate and describe a method how this estimate can be achieved. For that, we will first give a more formal description of the three parameter estimation problems mentioned in section 1 which shows that these three situations can be treated in the same way.

The abstract situation is as follows: We have a collection of objects, where each object may have several features e_i. For a fixed set of n feature pairs (e_i, e_j), we are seeking for estimates of $P(e_i|e_j)$, the probability that a random object has feature e_i, given that it has feature e_j. In a random sample of g objects (called learning sample in the following), we observe f objects with feature e_i, of which h objects also have the feature e_j. In the three situations described in section 1, the objects are documents. All probabilities mentioned in the following relate to the occurrence of the different features in these documents:

- For the estimation of precision values for probability estimation, we regard the features $q_{j,retr}$ = "document retrieved for request q_j" and $q_{j,R}$ = "document relevant w.r.t. request q_j". The parameter we want to estimate is $P(q_{j,R}|q_{j,retr})$, the probability that a random document which is retrieved for request q_j will be judged relevant by the user. In a random sample of g documents from the collection, f documents are retrieved for request q_j of which h are relevant and n is the number of requests in the collection.

- In the relevance feedback situation we regard the features "relevance" and the occurrence of terms. Our aim is to estimate $P(t_i|q_{j,r})$, the probability that a random document with relevance judgement r for request q_j has term t_i. Having relevance feedback information about g documents, f documents are judged with r of which h have document feature t_i. The number of request-term pairs considered is n.

- In the probabilistic indexing approach, each document has a set of manually assigned descriptors and its set of document features associated with it. The probability we are asking for is $P(s_i|t_j)$, the probability that the document would be assigned manually descriptor s_i, given that it has document feature t_j. In a random sample of g documents, f documents have feature t_j of which h documents are indexed with descriptor s_i. The total number of descriptor-document feature pairs considered is n.

Our aim is to derive an optimum estimate[1] $p_{opt}(e_i, e_j)$ for a feature pair with the parameter triplet (h, f, g). We are regarding a total of n pairs, where $Z(h, f, g)$ is the set of feature pairs with (h, f, g) and $\sum_{h,f} |Z(h, f, g)| = n$. Based on this information we want to derive a point estimate for $P(e_i|e_j)$, which is required for all known probabilistic IR models.[2] We will justify our choice of $p_{opt}(e_i|e_j)$ in two different ways: First, we give a plausible definition for p_{opt}, and second, we justify p_{opt} by means of a loss function.

For the definition of p_{opt}, we assume that we have the empirical data mentioned before from a learning sample of size g. Now we regard a (not necessarily different) set of g' objects. Let $E_{g'}(e_i, e_j)$ denote the expected value of the occurrences of the pair (e_i, e_j) within this set and $E_{g'}(e_j)$ the expected number of occurrences of feature e_j. Then the optimum estimate for $P(e_i|e_j)$ for which we have observed the frequency pair (h, f) within the g objects is defined as follows:

$$p_{opt}(e_i|e_j, (h, f, g)) = \frac{\sum_{(k,l)} P((e_k, e_l) \in Z(h, f, g)) \cdot E_{g'}(e_k, e_l)}{\sum_{(k,l)} P((e_k, e_l) \in Z(h, f, g)) \cdot E_{g'}(e_l)} \tag{1}$$

[1] A preliminary version of this approach has been described in [Hüther & Knorz 83]. In [Hüther 87], a more detailed derivation for p_{opt} is given.

[2] Only for the case of evaluation measures, also interval estimates would be appropriate.

Here we sum up over all n feature pairs (e_k, e_l), taking the probability that we would observe (h, f, g) for each pair and multiplying it with the expected numbers of occurrences: In the numerator, we count the expected numbers of these pairs (e_k, e_l) within the g' objects, and in the denominator, the expected numbers of occurrences of the features e_l are summed up. Our definition can be regarded as a kind of micro average of the underlying probabilities (similar to the micro average of retrieval measures).

In order to express the expectations in the above formula (1) as functions of the underlying probabilities we introduce the following notations: Let $p_{kl}{=}P(e_k|e_l)$ and $p_l{=}P(e_l)$ denote the probability that a random object has feature e_l. Furthermore assume that Q is a random variable of the prior distribution of the p_{kl}'s. In contrast to the approaches described in the following section, no specific assumption about this prior distribution is made. Finally let Z_g be the random variable for the frequency pairs (h, f) we observed within the g objects, such that $P(Z_g{=}(h, f)|Q{=}p_{kl})$ gives us the probability that a feature pair (e_k, e_l) with underlying probability p_{kl} has the frequencies (h, f) within the g objects. With these notations, we can rewrite (1) as

$$p_{opt}(e_i|e_j, (h, f, g)) = \frac{\sum_{(k,l)} n \cdot g' \cdot p_{kl} \cdot p_l \cdot P(Z_g{=}(h, f)|Q{=}p_{kl})}{\sum_{(k,l)} n \cdot g' \cdot p_l \cdot P(Z_g{=}(h, f)|Q{=}p_{kl})} \tag{2}$$

It is obvious that the definition of p_{opt} is independent of the size g'. As we have only data about the g objects in most cases, we will set $g'{=}g$ in the following. Furthermore we will drop the constant factors n and g.

The second justification for our definition of p_{opt} is based on the well-known loss function

$$L_1(\hat{p}, p_{ij}) = (\hat{p} - p_{ij})^2$$

Now an estimate p_{min} is chosen such that the expected value of L_1 is minimized:

$$\begin{aligned} p_{min}(e_i|e_j, (h, f, g)) &= \min_{0\le\hat{p}\le 1} (E(L_1(\hat{p}, p_{ij})) \\ &= \min_{0\le\hat{p}\le 1} \left(\sum_{(k,l)} (\hat{p} - p_{kl})^2 p_l P(Z_g{=}(h, f)|Q{=}p_{kl}) \right) \end{aligned}$$

$$\begin{aligned} \frac{dE}{d\hat{p}} &= \sum_{(k,l)} 2\hat{p}p_l P(Z_g{=}(h, f)|Q{=}p_{kl}) - \sum_{(k,l)} 2p_{kl}p_l P(Z_g{=}(h, f)|Q{=}p_{kl}) \\ \frac{dE}{d\hat{p}} &\overset{!}{=} 0 \Longrightarrow \\ p_{min} &= \frac{\sum_{(k,l)} p_{kl}p_l P(Z_g{=}(h, f)|Q{=}p_{kl})}{\sum_{(k,l)} p_l P(Z_g{=}(h, f)|Q{=}p_{kl})} \\ &= p_{opt} \end{aligned}$$

Having justified our choice of p_{opt} this way, we now want to show how p_{opt} can be estimated on the basis of data from our learning sample of size g.

Therefore we define $E^+(h, f, g)$ as the numerator of (2), i.e.

$$E^+(h, f, g) = \sum_{(k,l)} p_{kl} \cdot p_l \cdot P(Z_g{=}(h, f)|Q{=}p_{kl})$$

(the expected number of occurrences of (e_k, e_l)) and

$$E^-(h, f, g) = \sum_{(k,l)} (1 - p_{kl}) \cdot p_l \cdot P(Z_g{=}(h, f)|Q{=}p_{kl})$$

(the expected number of occurrences of e_l without e_k), so that we get

$$p_{opt}(e_i|e_j,(h,f,g)) = \frac{E^+(h,f,g)}{E^+(h,f,g)+E^-(h,f,g)} \tag{3}$$

The expectations $E^+(h,f,g)$ and $E^-(h,f,g)$ can be approximated by the expected values $E(h,f,g)$ of the frequency distribution $|Z(h,f,g)|$:

$$\begin{aligned}
E^+(h,f,g) &= \sum_{(k,l)} \binom{g}{f} p_l^f (1-p_l)^{g-f} \binom{f}{h} p_{kl}^h (1-p_{kl})^{f-h} \cdot p_l \cdot p_{kl} \\
&= \frac{h+1}{g+1} \sum_{(k,l)} \binom{g+1}{f+1} p_l^{f+1} (1-p_l)^{g+1-(f+1)} \cdot \\
&\quad \binom{f+1}{h+1} p_{kl}^{h+1} \cdot (1-p_{kl})^{f+1-(h+1)} \qquad (4)\\
&= \frac{h+1}{g+1} E(h+1,f+1,g+1) \\
&\approx \frac{h+1}{g} E(h+1,f+1,g) \qquad (5)
\end{aligned}$$

The approximation used above is not critical, in comparison to the probability estimation the error is of second order.

$$\begin{aligned}
E^-(h,f,g) &= \sum_{k,l} \binom{g}{f} p_l^f (1-p_l)^{g-f} \binom{f}{h} p_{kl}^h (1-p_{kl})^{f-h} p_l \cdot (1-p_{kl}) \\
&= \frac{f+1-h}{g+1} \sum_{k,l} \binom{g+1}{f+1} p_l^{f+1} (1-p_l)^{g+1-(f+1)} \cdot \\
&\quad \binom{f+1}{h} p_{kl}^h (1-p_{kl})^{f+1-h} \qquad (6)\\
&= \frac{f+1-h}{g+1} E(h,f+1,g+1) \\
&\approx \frac{f+1-h}{g} E(h,f+1,g) \qquad (7)
\end{aligned}$$

With these approximations for $E^+(h,f,g)$ and $E^-(h,f,g)$, we can estimate p_{opt} according to formula (3) as

$$p_{opt}(e_i|e_j,(h,f,g)) \approx \frac{(h{+}1)\, E(h{+}1,f{+}1,g)}{(h{+}1)\, E(h{+}1,f{+}1,g) + (f{+}1-h) \cdot E(h,f{+}1,g)} \tag{8}$$

To apply this formula, we need a sufficient amount of data about our learning sample of size g (that is, we have to observe a large number of feature pairs (e_k,e_l)). Then we can use the numbers $|Z(h,f,g)|$ of the frequency distribution as approximation of the expected values $E(h,f,g)$. We will return to this problem in section 5.

4 Other methods of parameter estimation

In this section, we discuss some other estimation methods which have been applied in IR research.

The most simple method is the method of moments which suggests that the population moments (e.g. mean and variance) can be estimated from sample moments. This method has been used in the experiments described in ([Harter 75a], [Harter 75b]) for the estimation of the parameters of the two-Poisson model. The problem of this approach is that in many applications, most of the parameter values lie outside the proper range, for which ad hoc estimates have to be defined ([Raghavan et al. 83]).

The maximum likelihood method estimates the value of a parameter as that value which maximizes the likelihood function. For our estimation problem, the maximum likelihood estimate of $P(e_i|e_j)$ is simply h/f. Besides the problem with the quotient 0/0 for which an ad hoc estimate has to be defined, the maximum likelihood estimate may also bear a bias, as will be shown in the following section.

Bayesian estimates are preferred most in IR research. This method assumes that we have previous knowledge about the parameter Q to be estimated. Based on the knowledge of the prior distribution of this parameter, we use the sample data in order to derive the posterior distribution of the parameter: Assume that X is a random variable which can have discrete values $x_1, x_2, \ldots$ depending on parameter Q which is a continuous random variable (it is also possible to assume Q as a discrete variable, but all applications described in the following assume a continuous one). Then $P(X{=}x_k|Q{=}q)$ is the probability that X will take the value x_k given that parameter Q has the value q. With $f(q)$ describing the prior distribution of parameter Q we get the posterior distribution

$$g(q|x_k) = \frac{f(q) \cdot P(X{=}x_k|Q{=}q)}{\int_{-\infty}^{\infty} f(q) \cdot P(X{=}x_k|Q{=}q) dq} \tag{9}$$

Further methods have to be applied in order to derive an estimate for q from this formula.

In the following discussion, we will restrict to a specific application of the Bayesian method to our problem where a beta distribution is assumed as prior distribution. The density of the beta distribution is given by the formula

$$f(p) = \frac{1}{B(a,b)}\, p^{a-1}(1-p)^{b-1}$$

with $B(a,b){=}\Gamma(a) \cdot \Gamma(b)/\Gamma(a+b)$ and $a, b > 0$ are parameters to be chosen. The assumption of a beta distribution is made (explicitly or implicitly) for most applications of the Bayesian method in IR ([Robertson & Sparck Jones 76], [Rijsbergen 77], [Bookstein 83], [?]). In contrast to our approach, these authors assume a single (beta) distribution for the parameters p_{ij}, independently of the probabilities p_j. This can be regarded as a kind of macro estimate, so the comparison with our micro estimate may not be quite appropriate. Furthermore, our optimum estimate also depends on the collection size g, while the sequential learning model described in [Bookstein 83] assumes that the prior distribution is independent of g. With the beta distribution as prior and the fact that we have observed the frequencies (h, f), we get:

$$g(p|\tfrac{h}{f}) = \frac{p^{a-1}(1-p)^{b-1}\binom{f}{h}p^h(1-p)^{f-h}}{\int_0^1 p^{a-1}(1-p)^{(b-1)}\binom{f}{h}p^h(1-p)^{f-h}dp}$$

Using the relationship $B(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}dp$, we get as posterior distribution:

$$g(p|\tfrac{h}{f}) = \frac{p^{h+a-1}(1-p)^{f-h+b-1}}{B(h+a, f-h+b)} \tag{10}$$

From this distribution, different estimates can be derived. One possibility is to choose that value p_{max} for which $g(p|\frac{h}{f})$ takes its maximum value. This approach is quite similar to the maximum likelihood method. With

$$\frac{dg(p|\frac{h}{f})}{dp} \overset{!}{=} 0$$

we get

$$p_{max} = \frac{h+a-1}{f+a+b-2}$$

A second approach is based on the definition of a loss function. Besides our function L_1, we also regard the loss function

$$L_2(p,\hat{p}) = \frac{(p-\hat{p})^2}{p(1-p)}$$

discussed in [Rijsbergen 77].

Now we seek for estimates p_L minimizing the expectation of the loss function, that is

$$\frac{d}{dp_L}\int_0^1 L(p,p_L)g(p)dp \overset{!}{=} 0$$

This yields the estimates[3]

$$p_{L_1} = \frac{h+a}{f+a+b}$$

and

$$p_{L_2} = \frac{h+a-1}{f+a+b-2} = p_{max}$$

Finally, in [Bookstein 83] a proposal is made to use the expectation of p as estimate:

$$p_E(e_i|e_j,\tfrac{h}{f}) = \int_0^1 g(p|\tfrac{h}{f})\ p\ dp$$

For the beta prior, we get

$$p_E(e_i|e_j,\tfrac{h}{f}) = \frac{h+a}{f+a+b} = p_{L_2}$$

It is interesting to notice that the four different methods for deriving an estimate from be posterior distribution yield only two different results when a beta distribution is assumed as prior. In any case there is still the problem of the estimation of the parameters a and b (see also the following section). In [?] several heuristic strategies for the choice of these parameters are evaluated. [Robertson & Sparck Jones 76] assumed a=b=$\frac{1}{2}$ in their experiments (following a proposal in [Cox 70]) and in [Robertson 86] parameter combinations with $a+b$=1 are discussed (according to our definition of p_{L_1}). For a=b=1 the beta distribution is identical with the uniform distribution. In this case p_{L_2}=p_{max} yields the same estimate as the maximum likelihood method.

We will now show how the parameters of the beta distribution can be interpreted in our approach (although we think that the assumption of a beta distribution is not suitable). Using the approximation

$$p_{opt}(h,f) \approx \frac{(h+1)\cdot E(h+1,f+1)}{(h+1)\cdot E(h+1,f+1)+(f+1-h)\cdot E(h,f+1)}$$

and setting this equal to

$$p_{L_1} = \frac{h+a}{f+a+b}$$

[3] In Rijsbergen's paper, a false value (that of p_{L_1}) is given as result for p_{L_2}

we get the relationship

$$\frac{E(h+1,f+1)}{E(h,f+1)} = \frac{(h+a)(f+1-h)}{(h+1)(f+b-h)} \quad (f \geq h \geq 0)$$

This gives us a relationship between frequencies of our distribution $Z(h,f,g)$ for the same f value: The above formula can be transformed into

$$E(h,f) = \frac{E(0,f)}{h \cdot B(a,b)} \cdot \frac{(f+h-1) \cdot (f+h-2) \cdot \ldots \cdot f}{(f+b-h) \cdot (f+b-h+1) \cdot \ldots \cdot (f+b-1)} \quad (f \geq h \geq 1)$$

5 Some experimental evidence

Having presented different approaches to the problem of parameter estimation in the last two sections, we now want to discuss the problem of their application and illustrate the resulting problems. For our further discussion, we will only consider the estimate p_{L_1} besides p_{opt}, because the other approaches can be regarded as variants of p_{L_1}.

f\h	4	5	6	7	8	9	10	11
4	2388							
5	2487	1015						
6	2289	1144	539					
7	2289	1081	623	317				
8	2129	1105	583	382	235			
9	2064	999	614	358	201	125		
10		950	550	348	227	161	89	
11		884	546	308	229	187	105	62

Table 2: Frequency distribution $|Z(h,f,g)|$ for the estimation of $P(s_i|t_j)$ (t_j = occurrence of noun phrases) from 78 000 manually indexed PHYS documents

f\h	4	5	6	7	8	9	10
4	0.671						
5	0.555	0.739					
6	0.440	0.634	0.781				
7	0.393	0.513	0.696	0.831			
8	0.326	0.480	0.576	0.692	0.848		
9		0.410	0.525	0.635	0.761	0.847	
10		0.382	0.441	0.598	0.710	0.737	0.867

Table 3: Estimates p_{opt} for the frequency distribution of Table 2

In order to overcome the lack of experimental data which prevented most of the authors cited in this paper from a deeper investigation in probability estimation, we use some data from an application of the Darmstadt Indexing Approach: Tables 2-3 are derived from the computation of the probabilities $P(s_i|t_j)$ from 78 000 documents of the PHYS database from the Fachinformationszentrum Energie Physik Mathematik, Karlsruhe, Germany. Here the occurrences of noun phrases in the documents are regarded as document features. The frequency distribution given in table 2 indicates that this collection is sufficiently large for an application of our approach, so that we can assume $|Z(h,f,g)|=E(h,f,g)$. Table 3 shows the corresponding values of p_{opt} according to formula (6). From these results two main conclusions can be drawn (see also the tables given below):

- Since all values of p_{opt} are smaller than the corresponding maximum likelihood estimates (MLE) $p_{ML}=\frac{h}{f}$, it can be said that the MLE are biased (in this case).
- The difference between p_{opt} and p_{ML} is substantial for small h and f values, it decreases for higher frequencies. So the estimate p_{opt} seems to yield a substantial improvement for low f frequencies.

For the approach based on the assumption of a beta distribution values for the parameters a and b were computed in two different ways:

- Optimum values for a and b were estimated on a test sample of 1 000 documents in such a way that the quadratic error $(k - p_{L_1}(a,b))^2$ on this sample was minimized ($k = 0/1$ is the intellectual descriptor assignment decision).
- The values a' and b' are derived from the p_{opt} values given in table 3 by minimizing the quadratic difference $(p_{opt} - p_{L_1}(a,b))^2$.

f	$Z_{g'}(h,f)$	a	b	a'	b'
4	139	-3.42	0.47	-1.01	1.47
5	341	2.64	5.85	-0.97	1.43
6	349	-1.54	1.48	-1.37	1.24
7	546	-1.61	1.33	-1.43	1.12
8	526	-1.54	1.76	-1.35	1.31
9	816	-1.29	1.84	-1.28	1.28
10	374	-1.46	1.31	-1.13	1.35

Table 4: Estimates for the parameters of the beta distribution. The values a and b are derived from a sample of 1 000 PHYS documents (sample X of table 5), where $Z_{g'}$ gives the number of pairs $(e_i|e_j)$ in this collection which had the term frequencies in the learning sample ($g = 78\,000$). The values a' and b' are derived from the p_{opt} values of table 3.

These computations were performed for different term frequencies which is more appropriate for the macro-oriented p_{L_1} estimate. From the results given in table 4, two interesting conclusions can be drawn:[4] [5]

- The assumption of a beta distribution is not appropriate in this case, because all a values are smaller than 0, which is not allowed for the beta distribution.
- Despite of a large scattering of these parameter estimates, it seems that the assumption of a single (a,b)-pair (a single beta-like distribution) for different f frequencies is not appropriate.

sample	$Z_{g'}$	estimate	s^2
X	2852	p_{ML}	0.271
		p_{L_1}	0.231
		p_{opt}	0.231
Z	2709	p_{ML}	0.265
		p_{L_1}	0.226
		p_{opt}	0.224

Table 5: Comparison of different kinds of estimates for f=4...10, $h \geq 4$ and $\frac{h}{f} > 0.4$ on two samples of 1 000 documents from the PHYS database. $Z_{g'}$ gives the number of feature pairs in these documents fulfilling the above conditions, s^2 is the average value of the error function $(k-p)^2$, where k=0/1 is the intellectual descriptor assignment decision.

In order to show the effect of improved estimation methods, experiments were made where the average quadratic errors $s^2=(k-p)^2$ for the different estimates were computed on the basis of two test samples with 1 000 documents each. Here only estimates for feature pairs with $f \leq 10$ (and $h \geq 4, h/f > 0.4$) were regarded, and for the p_{L_1} estimates, constant values $a = -1.13$ and $b = 2.13$ for all term frequencies were used. The results are given in table 5. It is obvious that the MLE are far worse than the improved estimates. In fact, the quadratic error for p_{ML} is even worse than in the case where all feature pairs considered would be given the same estimate ($s^2 = 0.249$ for both samples). This is a clear justification for the improved estimates. On the other hand, the difference in the results for p_{L_1} and p_{opt} is negligible, so we can conclude that we get almost optimum estimates with p_{L_1} Although the p_{L_1} estimate is easier to apply, its theoretical justification is doubtful. The p_{opt} approach requires more experimental data for its application, but it has a clear theoretical justification.

f\h	4	5	6	7	8	9	10
4	1489						
5	1078	833					
6	1005	633	510				
7	875	540	368	319			
8	896	469	326	238	250		
9	866	455	271	201	158	186	
10	751	368	245	154	139	115	106

Table 6: Frequency distribution $Z(h, f, g)$ for the estimation of $P(s_i|t_j)$ (t_i = occurence of single words or noun phrases) for the DIA from 22 000 manually indexed documents of the Food Science and Technology Abstract Service

[4] Because of some problems with the numeric algorithms used for the optimum search, the values a and b given here are of limited precision.

[5] The values pairs (a, b) and (a', b') cannot be compared directly, because two different methods of term identification were used in order to derive these values. This has the effect that the p_{L_1} estimates are smaller than the corresponding p_{opt} values.

f\h	4	5	6	7	8	9
4	0.79					
5	0.61	0.83				
6	0.51	0.67	0.86			
7	0.39	0.59	0.72	0.89		
8	0.34	0.47	0.63	0.76	0.91	
9	0.29	0.44	0.52	0.71	0.79	0.90

Table 7: Estimates p_{opt} for the frequency distribution of Table 6

f\h	4	5	6	7	8	9	10
4	1016						
5	636	466					
6	562	366	273				
7	418	252	232	172			
8	412	206	124	125	127		
9	342	162	119	97	70	85	
10	310	137	78	74	46	56	46

Table 8: Frequency distribution $Z(h, f, g)$ for the estimation of $P(s_i|t_j)$ ($t_i =$ occurence of single words or noun phrases) for the DIA from 11 000 manually indexed documents of the Food Science and Technology Abstract Service

f\h	4	5	6	7	8	9
4	0.79					
5	0.62	0.82				
6	0.50	0.73	0.84			
7	0.38	0.55	0.78	0.89		
8	0.32	0.52	0.66	0.74	0.92	
9	0.27	0.41	0.62	0.62	0.85	0.89

Table 9: Estimates p_{opt} for the frequency distribution of Table 8

Tables 6-9 show the results of an experiment regarding the influence of the collection size g on the estimates p_{opt}. The frequency distribution given in table 8 and the corresponding estimates (table 9) are derived from a half of the collection used for tables 6-7. Here it can be seen that the estimates p_{opt} for the same (h, f)-pair depend heavily on the collection size g: except for the cases $4/f$ and f/f, we get significant differences in the estimates for the two collections. Although we are not able to draw any specific conclusions for very small collections (e.g. in feedback experiments), we should say that we have only contrary evidence for the assumption of the estimate being independent of g (see e.g. [Bookstein 83]).

f\h	0	1	2	3	4	5	6	7	8	9	10
0	43										
1	44	17									
2	40	24	15								
3	5	1	4	4							
4	18	18	3	6	13						
5	35	12	10	7	5	10					
6	48	14	15	7	7	9	10				
7	12	5	5	5	6	6	4	10			
8	9	0	0	3	0	2	3	3	5		
9	6	6	1	1	0	1	0	5	3	4	
10	4	6	5	3	1	4	4	0	4	4	3

Table 10: Frequency distribution $|Z(h, f, g)|$ for relevance feedback from 122 queries of the AIR retrieval test (g=10)

Finally we want to mention the problem of the insufficient data for the application of the p_{opt} estimate. Table 10 shows the frequency distribution for relevance feedback from 10 documents per query for a sample of 122 queries of the AIR retrieval test ([Fuhr & Knorz 84]). Because of the large scattering of the frequencies shown here, it is not suitable to apply the p_{opt} estimate directly. In our view, the only possibility is the development of an appropriate bibliometric distribution describing this data. With such a distribution, a smoothing of the experimental data could be performed so that we would get better estimates $|Z(h, f, g)|'$ for the expectations $E(h, f, g)$, and then our formula for p_{opt} could be applied.

6 Conclusions

In this paper, we have tried to show the importance of probability estimation in IR research, and we have developed a new kind of estimation method which seems to be superior to former approaches.

We think that the problem of probability estimation is central to most of the work in IR, and that more attention should be paid to this problem: nearly any result of experimental work is only valid in combination with the parameter estimation method used. We have shown that there are situations where the simple maximum likelihood method yields biased estimates. This fact may have serious consequences in IR for the results of evaluations as well as for the application of probabilistic models. So it is necessary to investigate the influence of the choice of the estimation method on experimental results.

The definition of the optimum estimate p_{opt} is suited to the given constraints in a variety of typical IR situations. Furthermore, an estimation method for p_{opt} is described. In contrast to other approaches, our method is not based on (possibly unrealistic) assumptions which are hard to verify. The experimental results given here show that the assumption of a beta distribution may not be appropriate in certain situations. The problem with insufficient data for the application of our method occurs only with small experimental collections: For example, in a practical application for relevance feedback, the IR system could gather enough data from new queries while running, so that the parameter estimates could be improved step by step.

Our work should be regarded as a starting point for further investigations, because there is a number of related problems:

- As a variant to our approach, one could also define an optimum estimate according to the macro method (arithmetic mean), that is $p'_{opt}(e_i|e_j,(h,f,g)) = \sum_{(k,l)} P(Z_g{=}(h,f)|Q{=}p_{kl}) \cdot p_{kl}$. However, we do not yet have a method for the computation of this estimate.
- For non-binary features like e.g. the within-document frequency of terms or indexing weights, our approach has to be extended.
- In order to cope with insufficient data for the application for p_{opt}, appropriate two-dimensional bibliometric distributions for $|Z(h,f,g)|$ have to be investigated.
- The estimation of probabilistic parameters for retrieval with relevance feedback poses special problems: Either specific estimation methods have to be developed, or we need other retrieval models.

Finally it should be noted that the usage of the optimum estimate in the application of IR models does not necessarily improve retrieval quality: There may be several systematic errors in the application of a model which might compensate each other partially, such that the correction of one of these errors also could increase the overall error. Only further research work with careful analysis of the test design used will give us new insights in these problems.

References

Biebricher, P.; Fuhr, N.; Knorz, G.; Lustig, G.; Schwantner, M. (1988). The Automatic Indexing System AIR/PHYS - from Research to Application. In: Chiaramella, Y. (ed.): *11th International Conference on Research and Development in Information Retrieval*, pages 333–342. Presses Universitaires de Grenoble, Grenoble, France.

Bookstein, A.; Swanson, D. (1974). Probabilistic Models for Automatic Indexing. *Journal of the American Society for Information Science 25*, pages 312–318.

Bookstein, A. (1983). Information Retrieval: A Sequential Learning Process. *Journal of the American Society for Information Science 34*, pages 331–342.

Cheeseman, P. (1983). A Method of Computing Generalized Bayesian Probability Values for Expert Systems. In: *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, volume 1, pages 198–202.

Cooper, W.; Huizinga, P. (1982). The Maximum Entropy Principle and its Application to the Design of Probabilistic Information Retrieval Systems. *Information Technology: Research and Development 1*, pages 99–112.

Cox, D. (1970). *Analysis of Binary Data.* Methuen, London.

Fuhr, N.; Knorz, G. (1984). Retrieval Test Evaluation of a Rule Based Automatic Indexing (AIR/PHYS). In: Van Rijsbergen, C. J. (ed.): *Research and Development in Information Retrieval*, pages 391–408. Cambridge University Press, Cambridge.

Fuhr, N. (1986). Two Models of Retrieval with Probabilistic Indexing. In: Rabitti, F. (ed.): *Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval*, pages 249–257. ACM, New York.

Fuhr, N. (1988). *Probabilistisches Indexing und Retrieval.* Dissertation, TH Darmstadt, Fachbereich Informatik. Available from: Fachinformationszentrum Karlsruhe, Eggenstein-Leopoldshafen, West Germany.

Fuhr, N. (1989). Models for Retrieval with Probabilistic Indexing. *Information Processing and Management 25(1)*, pages 55–72.

Harper, D. J.; van Rijsbergen, C. J. (1978). An Evaluation of Feedback in Document Retrieval using Cooccurrence Data. *Journal of Documentation 34*, pages 189–216.

Harter, S. (1975a). A Probabilistic Approach to Automatic Keyword Indexing. Part I: On the Distribution of Speciality Words in a Technical Literature. *Journal of the American Society for Information Science 26*, pages 197–206.

Harter, S. (1975b). A Probabilistic Approach to Automatic Keyword Indexing. Part II: An Algorithm for Probabilistic Indexing. *Journal of the American Society for Information Science 26*, pages 280–289.

Hüther, H.; Knorz, G. (1983). Schätzung von Zuteilungswahrscheinlichkeiten für Deskriptoren als Eintrag im Indexierungswörterbuch. In: Deutsche Gesellschaft für Dokumentation (ed.): *Deutscher Dokumentartag 1982*, pages 139–161. K.G. Saur, München, New York, London, Paris.

Hüther, H. (1987). *Schätzung von Wahrscheinlichkeiten aufgrund kleiner Vorkommenshäufigkeiten in großen Kollektionen.* Report DV II-87-1, TH Darmstadt, FB Informatik, Datenverwaltungssysteme II.

Kantor, P.; Lee, J. (1986). The Maximum Entropiy Principle in Information Retrieval. In: Rabitti, F. (ed.): *Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval*, pages 265–268. ACM, New York.

Kantor, S. (1984). Maximum Entropy and the Optimum Design of Automated Information Retrieval Systems. *Information Technology: Research and Development 3*, pages 88–94.

Lam, K.; Yu, C. (1982). A Clustered Search Algorithm Incorporating Arbitrary Term Dependencies. *ACM Transactions on Database Systems 7*.

Losee, R. (1987). The Effect of Database Size on Document Retrieval:Random and Best-First Retrieval Models. In: van Rijsbergen, C. J.; Yu, C. T. (eds.): *Proceedings of the Tenth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, pages 164–169. ACM, New York.

Maron, M.; Kuhns, J. (1960). On Relevance, Probabilistic Indexing, and Information Retrieval. *Journal of the ACM 7*, pages 216–244.

Raghavan, V.; Shi, H.; Yu, C. (1983). Evaluation of the 2 Poisson Model as a Basis for Using Term Frequency Data in Searching. In: *Proceedings of the 1983 ACM Conference on Research and Development in Information Retrieval*, pages 88–100.

van Rijsbergen, C. J. (1977). A Theoretical Basis for the Use of Co-Occurrence Data in Information Retrieval. *Journal of Documentation 33*, pages 106–119.

Robertson, S.; Bovey, J. (1982). *Statistical Problems in the application of probabilistic models to information retrieval.* Report 5739, British Library, London.

Robertson, S.; Sparck Jones, K. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science 27*, pages 129–146.

Robertson, S. (1986). On Relevance Weight Estimation and Query Expansion. *Journal of Documentation 42*, pages 182–188.

Robertson, S. E.; van Rijsbergen, C. J.; Porter, M. F. (1981). Probabilistic Models of Indexing and Searching. In: Oddy, R. N.; Robertson, S. E.; van Rijsbergen, C. J.; Williams, P. W. (eds.): *Information Retrieval Research*, pages 35–56. Butterworths, London.

Salton, G.; Buckley, C.; Yu, C. (1983). An Evaluation of Term Dependence Models in Information Retrieval. In: Salton, G.; Schneider, H.-J. (eds.): *Research and Development in Information Retrieval*, pages 151–173. Springer, Berlin et al.

Womser-Hacker, C. (1987). *Der PADOK-Retrievaltest.* Dissertation, Universität Regensburg.

Yu, C.; Salton, G. (1976). Precision Weighting. An Effective Automatic Indexing Method. *Journal of the ACM 23*, pages 76–88.

Yu, C.; Buckley, C.; Lam, K.; Salton, G. (1983). A Generalized Term Dependence Model in Information Retrieval. *Information Technology: Research and Development 2*, pages 129–154.