# Probabilistic search term weighting—some negative results

Norbert Fuhr, Peter Müller
TH Darmstadt, Fachbereich Informatik
6100 Darmstadt
West Germany

## Abstract

The effect of probabilistic search term weighting on the improvement of retrieval quality has been demonstrated in various experiments described in the literature. In this paper, we investigate the feasibility of this method for boolean retrieval with terms from a prescribed indexing vocabulary. This is a quite different test setting in comparison to other experiments where linear retrieval with free text terms was used. The experimental results show that in our case no improvement over a simple coordination match function can be achieved. On the other hand, models based on probabilistic indexing outperform the ranking procedures using search term weights.

## 1 Introduction

Probabilistic search term weighting is a well-known method to achieve a better retrieval quality. In various experiments, significant improvements could be shown (e.g. [Robertson & Sparck Jones 76], [Robertson et al. 81], [Croft & Harper 79]). These experiments were based on linear retrieval with free text terms, where the terms in general were reduced to their word stems. In this paper, probabilistic search term weighting is used to rank the documents from the output sets of boolean retrieval, where the query terms are descriptors from a prescribed indexing vocabulary. In a second series of experiments, models based on probabilistic indexing (that is, document term weighting) are tested and the results are compared with those of the search term weighting models.



## 2 Probabilistic search term weighting

In the binary independence retrieval (BIR) model ([Robertson & Sparck Jones 76]), documents are represented by a binary vector $\vec{x}$, where $x_i = 1(0)$ stands for the fact that in the actual document the index term $s_i \in S = \{s_1, \ldots, s_n\}$ is present (not present). The well-known ranking function $g(\vec{x})$ which gives the relevance value of a document $d_m$ described this way with respect to a request $f_k$ yields

$$g(\vec{x}) = \log \frac{P(\vec{x}|R)}{P(\vec{x}|\bar{R})} = \sum_{i=1}^{n} x_i \cdot \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_{i=1}^{n} \frac{1-p_i}{1-q_i} \quad (1)$$

with $p_i = P(x_i{=}1|R)$ and $q_i = P(x_i{=}1|\bar{R})$, the probabilities that $s_i$ occurs in a relevant/nonrelevant document.

In order to estimate the parameters $p_i$ and $q_i$, usually relevance feedback information has to be obtained from the user for a small set of documents. We will denote the corresponding ranking procedure BIR/RF. For comparison, we will perform upper bound experiments by using complete relevance information about the documents to be ranked (ranking procedure BIR/UB).

An alternative way to apply the BIR model which does not need any relevance information has been shown in [Croft & Harper 79], where the inverse document frequencies (IDF) are used as estimates for the parameters $q_i$, and constant values for the parameters $p_i$ are assumed (ranking procedure BIR/IDF).

## 3 Probabilistic document term weighting

In contrast to the model described above, probabilistic indexing models assign weights to the terms in

a document by regarding the document with respect to queries described by binary vectors. The indexing model of [Maron & Kuhns 60], here called binary independence indexing (BII) model ([Fuhr 86]), estimates the probability of relevance of a document $d_m$ with respect to a query $f_k$ described by vector $\vec{x_k}$:

$$P(R|\vec{x_k}, d_m) = P(R|d_m) \cdot \prod_{s_i \in f_k^S \cap d_m^S} \frac{P(R|s_i, d_m)}{P(R|d_m)} \quad (2)$$

Here and in the following, $d_m^S$ denotes the set of index terms occuring in the document $d_m$ and $f_k^S$ the set of index terms used in the query formulation of $f_k$. The index term weight $P(R|s_i, d_m)$ denotes the probability that the document is relevant to a request using $s_i$ in its query formulation. $P(R|d_m)$ is a probabilistic document weight; experiments described in [Fuhr 86] have shown that the assumption of a constant value for this weight yields better results than an estimation of this parameter.

The RPI model introduced in [Fuhr 86] can be regarded as an extension of the BIR model from binary to weighted (probabilistic) indexing.[1] Here the probability of relevance of a document $d_m$ with respect to a request $f_k$ is given by the formula:

$$P(R|f_k, d_m) = P(R|f_k) \cdot \prod_{s_i \in f_k^* \cap d_m^*} \left( \frac{p_{ik}}{q_i} u_{im} + \frac{1 - p_{ik}}{1 - q_i}(1 - u_{im}) \right) \cdot \prod_{s_i \in f_k^* \setminus d_m^*} \frac{1 - p_{ik}}{1 - q_i} \quad (3)$$

Here $u_{im} = P(R|s_i, d_m)$ is the index term weight of $s_i$ in $d_m$[2]. The parameters $p_{ik}$ and $q_i$ are the average index term weights of $s_i$ in the relevant documents of $f_k$ resp. in all documents. We will apply the RPI formula in three ways:[3]

1. with constant values for the $p_{ik}$'s and $q_i$'s, denoted as ranking procedure RPI/const
2. with the parameters $p_{ik}$ and $q_i$ estimated through relevance feedback (RPI/RF)
3. in an upper bound experiment by using complete relevance information (RPI/UB).

# 4 Boolean retrieval

As boolean retrieval is in widespread use in practice, there are attempts to find a combination with probabilistic ranking procedures. Here we will concentrate on a model aimed to rank the output set of a boolean query proposed by Radecki ([Radecki 82, 83])[4]. In this model, the BIR ranking procedure is applied to rank the output set. This approach, however, has one major disadvantage: caused by the boolean query structure, the independence assumption of the BIR model cannot hold on the output set (if it is assumed to hold on the whole document collection). This can be illustrated by an example: Suppose that we have a collection of 1000 documents with document frequencies 200 for $s_1$, 250 for $s_2$ and 100 for $s_3$, and the actual query is $s_1 \wedge (s_2 \vee s_3)$. If the three index terms are assumed to be distributed independently, we get an output set of 65 documents, namely

45 documents containing $s_1$ and $s_2$ only,
15 documents containing $s_1$ and $s_3$ only, and
5 documents containing all three index terms.

Therefore we get for a random document of the output set (OS): $P(s_1|OS) = 50/65$, but $P(s_1|s_2, OS) = 1$, which means that $s_1$ and $s_2$ are not distributed independently in the output set. These problems also remain in the case where the minterms $s_1 \wedge s_2$ and $s_1 \wedge s_3$ are taken as basic terms for the application of the BIR model instead.

From the above, it can be concluded that serious problems arise when the BIR or the RPI model is applied to rank the output set of a boolean query and the probabilistic parameters are estimated on (parts of) this output set.[5]

There are two ways to overcome these problems:

- The probabilistic parameters are estimated using a representative sample of the whole document collection. This is easy if only IDF weights are used, but for relevance feedback purposes, it seems to be rather difficult to select the 'right' documents which are to be judged by the user.
- The probabilistic models have to be modified. For example, a dependence model could be developed which pays attention to the fact that the

[1] A similar model has been described in [Croft 81], where documents are ranked according to the expected value $E(g(\vec{x}))$ of the BIR model. However, it can be shown theoretically that this model does not give a ranking according to the probability ranking principle.

[2] In the original RPI model, a distinction between the concepts 'relevance' (relating to the retrieval process) and 'correctness' (where the indexing weights relate to) is made, which is dropped here for simplicity.

[3] It is not possible to use IDF weights for the RPI model in the same way as with the BIR model, because the ranking formula (3) cannot be separated in two factors, one with the $p_{ik}$'s and the other with the $q_i$'s

[4] An alternative approach was suggested in [Bookstein 85], but has not been evaluated yet.

[5] In this case, the parameters $q_i$ of the RPI model are request-specific, because they have to be related to the output set of the actual request.

boolean query structure generates certain dependencies between the index terms in the output set. Another possible solution could be provided by the application of the maximum entropy principle (see e.g. [Kantor & Lee 86]).

In contrast to the BIR and the RPI model, the application of the BII model in combination with boolean retrieval bears no problem of term dependence in the documents, because only assumptions about the distribution of terms in the queries are made, and these distributions seem to be independent of the kind of query formulation. In the experiments described below, only for the BIR/IDF method the search term weights are estimated on the basis of the whole document collection. All experiments with relevance feedback only use relevance information from the output sets.

# 5 Controlled vocabulary

For many large retrieval data bases, a prescribed indexing vocabulary is used (e.g. MEDLARS, INSPEC). When query formulation uses terms from this vocabulary instead of free text terms, this has two serious consequences for the application of ranking procedures based on search term weighting

- The average number of terms per query is smaller (see also table 1 in the following chapter), so the number of documents having the same binary description with respect to the query and thus getting the same rank increases.

- The search terms from the controlled vocabulary will have a more similar significance than it would be the case for free text terms, so the scattering of the search term weights also will be smaller (and interfere with random differences in the weights which are intrinsic to the estimation of probabilistic parameters).

# 6 Test setting

For our experiments, we use the collection of the AIR retrieval test (see [Fuhr & Knorz 84]) from the physics data base PHYS of the Fachinformationszentrum Karlsruhe, Germany. This collection consists of a sample data base with 15,000 documents and 309 boolean search requests (without NOT operators) which were formulated by retrieval experts in dialogue sessions at the complete data base PHYS. Here we regard the query formulations using terms from the indexing vocabulary only. As weighted probabilistic indexing, the so-called indexing A1 was taken which was adopted by comparison with manual indexing (see also [Knorz 83]) using a learning sample of 1000 documents[6].[7]

For the ranking experiments, retrieval was made in two steps. In the first step, conventional boolean retrieval was performed: A very broad unweighted indexing was chosen by applying a cut-off value of 0.01 to the weighted indexing A1. The sets of output documents selected this way (only 244 of the 309 sets were non-empty) were ranked in the second retrieval step by applying the different ranking formulas.

The 244 queries with non-empty answer sets were randomly divided into three samples named A, B and C. For the ranking experiments, we use the samples A and C[8]. All global constants such as $P(R|d_m)$ of the BII model, $p_i$ of the BIR/IDF procedure and $p_{ik}$ and $q_i$ of the RPI/const function are estimated on sample C by choosing the best values from a series of retrieval runs; these global constants are used for sample A without modification. For the experiments using relevance feedback information, subsets of the samples are taken, because many answer sets are too small for this purpose. For the samples A10 and C10, 10 documents formed the feedback set of each query, and there had to be at least 4 documents left to be ranked (residual ranking). The samples A20 and C20 only contain queries with at least 25 answer documents from which 20 documents are used for relevance feedback. The sizes of the test samples are shown in table 1.

As evaluation measure, we take the $R_{norm}$ measure as defined by Bollmann [Bollmann 84] for multistage relevance scales (see Appendix), because such a relevance scale had been used in the AIR retrieval test. The macro average $R^M_{norm}$ is computed as well as the micro-macro average $R^m_{norm}$ (which is a weighted average according to the different sizes of the answer sets), because the latter measure seems to be more appropriate for the scattering of answer sizes in the test samples.

[6] None of these documents have been used for the retrieval test

[7] Although these weights refer rather to events of 'correctness' (judged by a human indexer) than of 'relevance' (which would theoretically be required for the BII model at least), we assume that the difference between these concepts is negligible, because there is no necessity to make this distinction in the application regarded here.

[8] Sample B is not regarded because of technical problems

| Sample | Queries | Documents | Terms/Query |
|---|---|---|---|
| A | 79 | 2835 | 5.8 |
| C | 83 | 2819 | 4.5 |
| A10 | 39 | 2242 | 5.9 |
| C10 | 41 | 2207 | 4.8 |
| A20 | 25 | 1884 | 5.5 |
| C20 | 25 | 1822 | 4.7 |

Table 1: Sizes of the test samples

# 7 Experimental results

Table 2 shows the results of the BIR model in comparison to the coordination match function. For these experiments with binary indexing, a cutoff value of 0.12 was applied to the indexing weights (at the cutoff value of 0.01, the ranking results were far worse). The BIR/IDF and the coordination match function yield nearly identical results. As the IDF weights are estimated on the whole document collection, they are independent of the boolean query structure. So the only reason for the lack of any improvement seems to be the kind of query terms employed here: in contrast to the experiments described in the literature, our queries only use descriptors from a prescribed vocabulary. Furthermore, the queries were formulated by retrieval experts. We assume that these descriptors have a more similar significance than free text terms, so a term weighting based on collection frequencies bears no useful information.

Replacing the IDF weights by weights based on relevance feedback (BIR/IDF), the results get even worse (however, only for the sample C10 the sign test shows a significant difference at a level of 95%). Obviously the estimation of the probabilistic parameters using only documents from the output set of the boolean queries is not suitable for the BIR model: the violation of the independence assumption cannot be ignored. The same problem arises with the application of the RPI model when relevance feedback is used, as can be seen from table 3.

Comparing the results of the BII and the RPI/const function with those of the coordination match function, a significant improvement (sign test: $> 97.5\%$) can be found. This difference is caused by the more detailed document representation (weighted instead of binary indexing) the BII and the RPI model are based upon. This fact is also illustrated by the results of the upper-bound experiments of the BIR and the RPI model.

| Ranking procedure | sample | $R^M_{norm}$ | $R^m_{norm}$ |
|---|---|---|---|
| COORD | A | 0.718 | 0.657 |
| BIR/IDF | | 0.711 | 0.653 |
| COORD | C | 0.672 | 0.623 |
| BIR/IDF | | 0.667 | 0.624 |
| COORD | A10 | 0.662 | 0.647 |
| BIR/IDF | | 0.659 | 0.642 |
| BIR/RF | | 0.611 | 0.626 |
| BIR/UB | | 0.711 | 0.683 |
| COORD | C10 | 0.614 | 0.619 |
| BIR/IDF | | 0.622 | 0.624 |
| BIR/RF | | 0.606 | 0.592 |
| BIR/UB | | 0.671 | 0.679 |
| COORD | A20 | 0.705 | 0.651 |
| BIR/IDF | | 0.705 | 0.647 |
| BIR/RF | | 0.634 | 0.640 |
| BIR/UB | | 0.743 | 0.684 |
| COORD | C20 | 0.640 | 0.613 |
| BIR/IDF | | 0.652 | 0.621 |
| BIR/RF | | 0.645 | 0.621 |
| BIR/UB | | 0.703 | 0.679 |

Table 2: Ranking results of the BIR model in comparison to the coordination match

| Ranking procedure | sample | $R^M_{norm}$ | $R^m_{norm}$ |
|---|---|---|---|
| BII | A | 0.740 | 0.693 |
| RPI/const | | 0.770 | 0.733 |
| BII | C | 0.734 | 0.702 |
| RPI/const | | 0.725 | 0.704 |
| BII | A10 | 0.706 | 0.685 |
| RPI/const | | 0.732 | 0.725 |
| RPI/RF | | 0.700 | 0.715 |
| RPI/UB | | 0.793 | 0.759 |
| BII | C10 | 0.732 | 0.704 |
| RPI/const | | 0.738 | 0.707 |
| RPI/RF | | 0.709 | 0.694 |
| RPI/UB | | 0.785 | 0.756 |
| BII | A20 | 0.773 | 0.712 |
| RPI/const | | 0.791 | 0.728 |
| RPI/RF | | 0.752 | 0.730 |
| RPI/UB | | 0.826 | 0.759 |
| BII | C20 | 0.722 | 0.697 |
| RPI/const | | 0.714 | 0.696 |
| RPI/RF | | 0.714 | 0.697 |
| RPI/UB | | 0.752 | 0.745 |

Table 3: Ranking results of the BII and the RPI model

# 8 Conclusions

Although boolean retrieval and controlled vocabulary are in widespread use in retrieval practice, there has not been an experimental evaluation of probabilistic IR models for this situation before. We have shown that in this case, the BIR model does not lead to improved retrieval quality. It seems that both parameters—boolean queries and controlled vocabulary—do have an influence on this result, but further experimental work is necessary where each of these parameters is considered separately. On the other hand, it is obvious that a more detailed document representation (than that of the BIR model) improves ranking results significantly (see also [Salton 86] and [Croft 83]).

From a theoretical point of view, none of the existing probabilistic retrieval models (except the maximum entropy principle, which has not been evaluated yet) seems to be appropriate for boolean retrieval. Similarly, there is a lack of probabilistic models which use a more detailed document representation (e.g. the number of occurrences of a term in a document). Here the 2 Poisson model yields rather unsatisfactory results ([Robertson et al. 81], [Losee et al. 86]). This shows that more elaborate probabilistic models are required: they should be adaptable to different requirements, but also simple enough to give us a better understanding of probabilistic retrieval.

# References


**Bollmann, P.** (1984). Private communication.

**Bookstein, A.** (1985). Representating boolean structure within a probabilistic framework. In: *Proceedings of the riao 85 (Recherche d' Information Assistée par Ordinateur)*, 18. - 20.3.85 in Grenoble (France), 373-386.

**Croft, W.B.; Harper, D.J.** (1979). Using Probabilistic Models of Document Retrieval without Relevance Information. *Journal of Documentation 35*, 285-295.

**Croft, W.B.** (1981). Document Representation in Probabilistic Models of Information Retrieval. *Journal of the American Society for Information Science* 32(6), 451-457.

**Croft, W.B.** (1983). Experiments with Representation in a Document Retrieval System. *Information Technology* 2(1), 1-22.

**Fuhr, N.; Knorz, G.** (1984). Retrieval test evaluation of a rule based automatic indexing (AIR/PHYS). In: Rijsbergen, C.J. van (Ed.): *Research and Development in Information Retrieval.* Cambridge University Press, Cambridge, 391-408.

**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*, 249-257.

**Kantor, P.B.; Lee, J.J.** (1986). The maximum entropy principle in Information Retrieval. In: Rabitti, F. (Ed.): *Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval*, 269-274.

**Losee,R.; Bookstein, A.; Yu, C.T.** (1986). Two Poisson and binary independence assumptions for probabilistic document retrieval. In: Rabitti, F. (Ed.): *Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval*, 258-264.

**Maron, M.E.; Kuhns, J.L.** (1960). On Relevance, Probabilistic Indexing, and Information Retrieval. *Journal of the Association for Computing Machinery 7*, 216-244.

**Radecki, T.** (1982). A Probabilistic Approach to Information Retrieval in Systems with Boolean Search Request Formulations. *Journal of the American Society for Information Science 33*, 365-370.

**Radecki, T.** (1983). Incorporation of relevance feedback into Boolean retrieval systems. In: Salton, G.; Schneider, H.-J. (Ed.): *Research and Development in Information Retrieval.* Springer, Berlin, Heidelberg, New York, 118-132.

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

**Robertson, S.E.; Rijsbergen, C.J. van; Porter, M.F.** (1981). Probabilistic models of indexing and searching. In: Oddy, R.N; Robertson, S.E.; van Rijsbergen, C.J.; Williams, P.W. (Ed.): *Information Retrieval Research.* Butterworth, London, 35-56.

**Salton, G.** (1986). Recent trends in automatic Information Retrieval. In: Rabitti, F. (Ed.): *Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval*, 1-10.


# Appendix

We give a short description of the $R_{norm}$ measure for multistage relevance scales which is an extension of the well-known $R_{norm}$ measure for binary scales. This measure only considers documents in different

ranks and with different relevance judgements. A pair of these documents is in right order if the document with the higher relevance judgement comes first, otherwise it is in wrong order. Let $S^+$ be the number of document pairs in right order, $S^-$ the number of those in wrong order, und $S^+_{max}$ the number of documents in right order for an optimum ranking. The normalized recall is defined then as follows:

$$R_{norm} = \frac{1}{2}\left(1 + \frac{S^+ - S^-}{S^+_{max}}\right)$$

For the cases with $S^+_{max} = 0$ we define $R_{norm} = 1$. In the other cases, a random ordering of documents will yield an $R_{norm}$ value of 0.5 in the average.
Because of the large scattering of the answer sizes, we use a second average method besides the macro average $R^M_{norm}$: the micro-macro average $R^m_{norm}$ is a weighted average with respect to the answer sizes. Let $n_i$ be the answer size of retrieval result $\Delta_i$, then the micro-macro average of $R_{norm}$ for a set of $t$ queries is defined as:

$$R^m_{norm}(\Delta_1, \ldots, \Delta_t) = \frac{\sum_{i=1}^{t} n_i \cdot R_{norm}(\Delta_i)}{\sum_{i=1}^{t} n_i}$$