4. Probabilistic information retrieval

Early attempt to apply probabilistic theory to information retrieval dates back to the work by Maron and Kuhns [Maron60] on *probabilistic indexing*. Their work made substantial contribution to the formulation of probabilistic approaches to such principal concepts as relevance and ranking. However the idea of probabilistic indexing is essentially different from *probabilistic retrieval*, the models of which were developed by Robertson and Sparck Jones [Robertson76], Van Rijsbergen [Van Rijsbergen79] and Croft and Harper [Croft79]. Beginning from the eighties there has also been attempted integration of probabilistic indexing and retrieval approaches into a *unified theory* of information retrieval [Robertson82].

Before moving on to the discussion of the probabilistic model used in this project – Robertson & Sparck Jones model – it is worthwhile to mention the main principles characteristic of all probabilistic retrieval theories. The aim of probabilistic theories is to rank documents in descending order of the probability of their relevance to the query. Since each document is unique and it is impossible to estimate probabilities of unique events, the models derive these probabilities from the characteristics of non-unique events constituting the documents – terms. The estimation of the term's usefulness in discriminating relevant from non-relevant documents is done through *term weighting*. The probability of relevance of each document to the query is estimated through the corresponding weighting matching function. The main differences between probabilistic models are in the weighting functions they use, and the ways these functions are derived.

4.1 Robertson & Sparck Jones model of probabilistic IR

The probabilistic model developed by Robertson and Sparck Jones [Robertson76] approaches the general problem of information retrieval by asking for each document and query the *Basic Question* [Sparck Jones98, p. 5]:

> "What is the probability that *this* document is relevant to *this* query?"

There are two assumptions about relevance implied by the Basic Question:

1. Relevance is a binary attribute;
2. Relevance can be attributed to a document irrespectively of other documents in the collection.

These assumptions are of course simplifications, but they are essential for the model. The main idea of the model is to estimate the probability of relevance of each document in order to use this estimation in ranking the documents for presentation to the user. This idea of the model is known as a *Probability Ranking Principle* [Robertson77], quoted from [Sparck Jones98, p.7]:

> **P1** : If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best to be gotten for the data.

The probability of relevance of each document is estimated from the relevance-predicting characteristics of attributes constituting a document – terms, specifically

query terms. Relevance-based term weighting developed by Robertson & Sparck Jones, and which will be covered in more detail in the next subsection, has been derived from a formal *theory of relevance weights* [Robertson76]. The theory's aim is to achieve document ranking through weighting of terms given some known information on their behaviour in relevant and non-relevant documents. There are two principal assumptions underlying the theory [Robertson76, p.132]:

Independence assumption: The distribution of terms in relevant documents is independent and their distribution in non-relevant documents is independent.

Ordering principle: That probable relevance is based on both the presence of search terms in documents and the absence from documents.

For more detailed discussion of the independence assumptions see section 4.2.

4.1.1 Term relevance weighting

The theory of relevance weights based on the above assumptions, also known as the binary independence retrieval model, is represented on the general level by the formula:

$$w_i = \frac{p_i(1-\overline{p}_i)}{\overline{p}_i(1-p_i)} \qquad (4.1)$$

where p_i is the probability of occurrence of term i in a relevant document;
$\overline{p}_i$ is the probability of term i occurrence in a non-relevant document.

The estimation of the above probabilities requires presence of information about occurrence of terms in relevant and non-relevant documents. The relevance information for a query need not be complete, and in practice it is usually available only to some extent. This information is obtained in the form of user relevance judgements. The following contingency table of term incidence has been proposed as a basis for a more specific representation of the model [Robertson76]:

	Relevant	Non-relevant	
Containing the term	r	$n-r$	n
Not containing the term	$R-r$	$N-n-R+r$	$N-n$
	R	$N-R$	N

where R is the number of known relevant documents
r is the number of known relevant documents with term i
N is the number of documents in the collection
n is the number of known documents containing term i

Given the term incidence information in the contingency table, general term weighting function (4.1) can be re-written as following:

$$w = \log \frac{r(N-n-R+r)}{(R-r)(n-r)} \qquad (4.2)$$

Before any relevance information is obtained, the estimation of p_i and $\overline{p}_i$ is performed on the assumption that the number of relevant documents R in the collection is much smaller than the number of non-relevant, therefore $\overline{p}_i$ is estimated as the number of documents in the collection with term i, i.e. n/N, while p_i is assumed to be a constant. In these conditions the term weighting function (4.2) gives the same result as a collection frequency weight (*CFW*) function, introduced first in [Sparck Jones71(b)]:

$$CFW = \log \frac{N}{n_i} \tag{4.3}$$

The problem with function (4.2) is that when no or very little relevance information is known, i.e. when the value of any of the central cells in the contingency table is zero, it results in the infinite weights. To avoid this problem 0.5 was added to each of the four parameters of the function. The resulting Relevance Weight (*RW*) formula is:

$$RW = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \tag{4.4}$$

The matching score (MS) for a document is simply the sum of the weights of the present query terms:

$$MS - RW = \sum_i \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)}$$

Further elaboration of the term weighting function was targeted at exploiting the information on term frequencies within documents [Robertson94] and document length normalisation. Robertson and Walker suggested a formula which modulates the weight of term i by its within-document frequency - TF_i:

$$W(TF_i) = \frac{TF_i (k_1 + 1)}{k_1 + TF_i} w_i \tag{4.5}$$

The formula has the following behaviour: it returns zero for term absence, i.e. $TF_i = 0$, and it increases monotonically with term frequency, but only to a certain point. The weight increase is adjustable by the tuning constant k_1: if $k_1 = 0$, no account of TF_i is taken, whereas a high value for k_1 results in the linear increase of weight with TF_i. The most effective value was experimentally determined to be from 1.2 to 2 [Sparck Jones98]. From this evidence the authors infer that TF_i has a non-linear effect on performance, i.e. after a certain point, taking account of more term occurrences contributes very little.

Inclusion of term frequency into weight estimation in turn requires a mechanism to equally handle documents of different lengths. To develop document normalisation

function Robertson and Sparck Jones made a simple assumption that "where there are two documents about the same topic but of different lengths, this is just because the longer is more wordy" [Sparck Jones98, p.25]. The normalisation factor they suggested can be tuned using the constant *b* to vary the strength of the above assumption:

$$NF = ((1-b) + b\frac{DL}{AVDL}) \tag{4.6}$$

where *DL* is document length
AVDL is average document length
b is a tuning constant

If *b* is set to 1, the formula will normalise document length on the above assumption, i.e. that longer documents are simply more verbose. Whereas smaller values will reward longer documents on the premise that their length is not entirely due to higher verbosity, but to a greater topic development. Experimental evidence suggests an optimum value of 0.75 for *b* [Sparck Jones98].

Inclusion of the document normalisation factor (4.6) into function (4.5) leads to the following formula:

$$W(TF_i) = \frac{TF_i(k_1+1)}{k_1 * ((1-b) + b\frac{DL}{AVDL}) + TF_i} w_i \tag{4.7}$$

The function (4.7) can be instantiated either with, or without relevance information. The function where w_i is *CFW* formula (4.3), i.e. without relevance information, is known as Combined Weight (*CW*). The function with *RW* (4.4) as w_i is known as Combined Iterative Weight (*CIW*).

There have also been developed variations of function (4.7) to take account of query term frequency for long queries: query adjusted combined weight (*QACW*):

$$QACW = CW * QTF \tag{4.8}$$

and query adjusted combined iterative weight (*QACIW*):

$$QACIW = CIW * QTF \tag{4.9}$$

where in both formulas *QTF* is query term frequency.

4.1.2 Relevance feedback

The mechanism for reweighting/reformulating the original query by means of the terms in the retrieved relevant documents is known as *relevance feedback*. Since first suggested by [Rocchio65], the idea of reformulating the query by adding selected terms from the relevant documents is widely considered to be more effective than reweighting the original terms alone.

In the Robertson & Sparck Jones probabilistic model, relevance feedback is implemented through selection of terms from the relevant documents as candidates for the new expanded query. Robertson and Sparck Jones emphasise the *selective* approach to query expansion, as opposed to an alternative strategy of *massive* query expansion, advocated by some research groups, e.g. SMART [Buckley95], where any term is added to the query, for which there is some positive/negative evidence of usefulness. Robertson and Sparck Jones argue that selection is useful, since it makes possible adapting query expansion to the environment conditions of the search, i.e. varying the number of added terms according to the size of the original query. The size of the expanded query is, therefore, predictable, and not affected by the size of the relevant documents, which, if massive expansion is used, can yield a very large number of terms.

They suggest ranking terms in the relevant documents and applying a threshold for term selection, which varies with the size of the original query. The ranking function used for term selection should be different from the function used for term weighting. Robertson and Sparck Jones argue that the question answered by relevance weight:

> **Q1:** "How much evidence does the presence of this term provide for the relevance of this document?"

is different from the question that should be answered by the term selection value (TSV):

> **Q2:** "How much will adding this term to the request benefit the overall performance of the search formulation?" [Sparck Jones98, p.30]

The term selection value should, therefore, reward those terms which are likely to have a strong impact on bringing relevant documents to the top of the ranked list, and not those, which though being good relevance indicators for one document are unable to contribute more because they are too infrequent. To achieve this kind of term selection Robertson [Robertson90] derived the following term selection formula, known as offer weight (OW):

$$OW = rRW \qquad (4.10)$$

Terms selected by OW are then weighted using the term weighting function.

A set of Okapi experiments on TREC data [Sparck Jones98], aimed at identifying the size of the optimum query expansion set for different length types of original queries, established that the following sizes work reliably well: 32 terms for long queries (title, description and narrative fields of TREC topics), 24 for medium (title and description) and 16 for very short queries (title field only). The experiments also suggested that small changes in the sizes of expansion sets did not affect performance.

Another question put forward by Robertson et al. is whether original query terms should be rewarded by the term selection value. They conducted experiments, which emphasised original query terms by assuming the existence of 20 hypothetical documents, with each query term occurring in 19 of them. The results showed no performance improvement from rewarding original query terms.

A query expansion strategy related to relevance feedback, and which was extensively tested with Okapi, is *blind* query expansion. Instead of known relevant documents, it uses documents *assumed* to be relevant through the information implied by their ranking position in the initial ranked set. The documents for query expansion are taken from the top of the ranked set and terms are selected from them on the same principle as for relevance feedback described above. Okapi experiments showed that blind feedback, though not always as good as relevance feedback, can be better than no query expansion.

4.2 Term independence assumptions

There are three fundamental assumptions commonly underlying probabilistic retrieval theories:

I1: Terms occur independently within the whole collection.

$$P(A,B) = P(A)P(B)$$

I2: Terms occur independently within the set of relevant documents.

$$P(A,B \mid R) = P(A \mid R)P(B \mid R)$$

I3: Terms occur independently within the set of nonrelevant documents.

$$P(A,B \mid \overline{R}) = P(A \mid \overline{R})P(B \mid \overline{R})$$

Robertson and Sparck Jones [Robertson76] preferred assumptions I2 and I3 over I1 and I2, since the latter contradict each other, as was recognised by the authors, and later demonstrated by [Cooper90]. Because of I2 and I3 assumptions, the model developed by Robertson and Sparck Jones became known as *binary independence model*.

Though independence within relevant and non-relevant documents is a less stringent assumption than independence within the whole collection, it is still an oversimplification of the data known about the real word behaviour in texts. Independence assumptions are needed however for the following benefits they bring [Sparck Jones98]:

- They simplify the task of model development;
- Model instantiation as an operable system is made possible;
- They result in performance improvements over simple term matching.

If no term independence was assumed, the implementation of such model as a working system would be unrealistic, since the probabilities would need to be estimated for every possible combination of terms in the documents.

However it has been argued [Sparck Jones98] that estimation of all term dependencies may not be needed under the model, since the assumptions it is built on – I2 and I3 – imply term dependencies within the whole collection. Specifically, if it is known that two terms occur more often in relevant documents than in non-relevant, then they can be considered to have some degree of dependency in the whole collection.

Cooper [Cooper90] suggested a weaker assumption, which he called *linked dependence* assumption (I4):

I4: $$\frac{P(A,B \mid R)}{P(A,B \mid \overline{R})} = \frac{P(A \mid R)}{P(A \mid \overline{R})} \frac{P(B \mid R)}{P(B \mid \overline{R})}$$

This assumption associates the relevance-implied dependency between A and B with the nonrelevance-implied dependency. As Cooper argued this is not an independence assumption, but assumption of linked dependencies. What important is that it leads to exactly the same equation as that of binary independence model.

All this leads to the conclusion that the binary independence model already accounts to some extent for relevance-implied document-level term dependencies. In the early days after the appearance of Robertson & Sparck Jones model, there have been many formal/less formal attempts to compensate for independence assumptions in probabilistic retrieval [e.g., Van Rijsbergen77, Harper78, Smeaton83] (see section *3.3.1*), which hardly led to any significant or consistent improvements. Cooper argues that present cooling of interest to such modifications is due to the wider recognition of the above-described implied dependencies. Moreover, he hypothesises that document-level co-occurrence data may not in itself be an important factor in retrieval performance [Cooper90].