EXPLOITING THE PROBABILITY RANKING PRINCIPLE TO INCREASE THE EFFECTIVENESS OF CONVENTIONAL BOOLEAN RETRIEVAL SYSTEMS

Tadeusz RADECKI

Department of Computer Science, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0115, U.S.A.

Abstract

This paper reports on research aimed at developing practical methods for improving the performance of conventional Boolean information retrieval systems. More specifically, the objective of this research is to incorporate into these systems a mechanism for ranking the documents of a collection in descending order of their probabilities of usefulness to the user. There are several reasons why a ranking mechanism of this type may be expected to provide retrieval results superior to those of traditional Boolean search techniques which normally do not furnish users with any indication of document relevance. In particular, the so-called output overload problem, which occurs when the size of the document set retrieved in response to a given query is unmanageable, could practically be eliminated since the ranking information would assist the searcher in deciding when to end an examination of the system's output with the confidence of identifying most of the useful items that have been retrieved. Moreover, since it would no longer be necessary to inspect all of the output documents, a user query could then be constructed broad enough to allow more relevant items to be retrieved. Accordingly, such a system enhancement may result in an improvement in retrieval effectiveness, especially in an increase in recall.

This paper presents and illustrates a theoretical framework, based on the Probability Ranking Principle, for implementing this enhancement. Research indicates that a relevance-based ranking scheme of this type could readily be incorporated into conventional retrieval systems without altering their underlying fundamental principles; some additional software in the form of a front-end is all that would be needed. As a result, the performance of these systems may be significantly improved at an acceptable cost.

INTRODUCTION

A great deal of research on information retrieval has been motivated by the recognition that clues for estimating the degrees of usefulness, or relevance, of documents in a collection are rarely complete. More specifically, it is widely recognized that predicting document relevance usually involves uncertainties which are inherently probabilistic, and as a result retrieval systems are generally unable to precisely determine the extent to which a given document will satisfy the user's needs. Thus, designing retrieval systems which are capable of identifying all and only relevant documents does not seem feasible under available theoretical models. However, it is often suggested that they should be able to present the retrieved documents in descending order of their estimated
probabilities of relevance. The desire to rank the output in this fashion underlies
the Probability Ranking Principle (PRP) (see, e.g., Ref. [1]), which is the
foundation for much of the research into probabilistic methods of information
retrieval since the earlier studies by Maron and Kuhns [2].

MOTIVATION
The ultimate objective of the PRP-based approach to information retrieval is to
rank the individual documents of a collection in descending order of their
estimated probabilities of usefulness to the user. Therefore, a ranking mechanism
of this type could be very helpful in operational retrieval environments. In
particular, it would practically eliminate the so-called output overload problem
which occurs when an unmanageable number of items are retrieved with respect
to a given query. More specifically, ranking information would assist the searcher
in deciding when to end an examination of the system's output with the
confidence of identifying most of the useful documents that have been retrieved.
Since it would no longer be necessary to inspect the entire output, user queries
could be constructed broad enough to allow the retrieval of more relevant items.
Accordingly, using a PRP-based ranking mechanism may result in an improvement
in retrieval effectiveness, especially in an increase in recall (*).

In view of the potential of the PRP-approach, it is of particular importance to
try to incorporate it into conventional Boolean retrieval systems. However, this
approach has been developed primarily for use in systems in which both
documents and queries are represented by sets of index terms. Thus, of the two
major components of any conventional retrieval system-(i) procedures for
document indexing, and (ii) procedures for query formulation-only the indexing
component is compatible with the PRP approach in conventional systems, queries
are represented by Boolean expressions referred to as Boolean search request
formulations (Boolean SRF)). Therefore, one may conclude that this ranking
mechanism cannot readily be incorporated into these systems. However,
preliminary studies indicate that a formal theory can be developed, on the basis
of which a front-end could be designed to merge the PRP-based ranking scheme
with conventional Boolean retrieval systems [3,4]. Since such an enhancement
would permit the fundamental principles behind these systems to be preserved,
significant improvement in their performance may be achieved at an acceptable
cost. The remainder of this paper presents a theoretical framework for extending
the PRP-based approach to conventional Boolean retrieval systems. First, the
original PRP approach is briefly described to help clarify this presentation.

OVERVIEW OF THE ORIGINAL PRP-APPROACH
As already mentioned, the PRP-based approach, as first introduced, requires both
documents and queries to be represented by sets of index terms. A major
observation underlying the development of this approach is that with respect to
this type of query representation, referred to as a binary SRF, in the document
collection a number of disjoint sets can be identified, each corresponding to a
different non-empty subset of the query's index terms. More specifically, each
set contains those documents whose representations include the terms that are
present in the corresponding subset, but do not include those terms which are
absent. For instance, if a query is represented by the three index terms, t_1, t_2,
and t_3 then seven disjoint document sets can be identified: one containing those
documents which are indexed by all three terms, a second containing those that
are indexed by t_1 and t_2, but not t_3 etc. The next step in the reasoning is to
note that individual query terms, present in or absent from a document
representation (DR), provide clues which can be used to estimate the probability
of the document being useful. From the perspective of a given binary SRF, all

(*) Recall is defined as the proportion of relevant documents which have been
retrieved (and presented to the user).
The Effectiveness of Conventional Boolean Retrieval Systems

The documents in a given set are indistinguishable. Thus, each document set may be characterized by a hypothetical document that is represented by the corresponding subset of the query's index terms. For example, a hypothetical document associated with the document set containing those items indexed by \( t_1 \) and \( t_2 \) but not \( t_3 \) is represented by \( \{ t_1, t_2 \} \). Therefore, it is sufficient to calculate the probability of relevance for each hypothetical document set. Accordingly, with respect to a given query, the number of different non-empty subsets of the query's index terms, or equivalently, the number of the corresponding document sets, there are several procedures available which can be used to calculate these probabilities (one of which is applied later in this paper). The details of these procedures may be found in the relevant literature on probabilistic design principles (see, e.g., Refs. [5,9]).

EXTENDING THE ORIGINAL PRP-APPROACH TO CONVENTIONAL BOOLEAN RETRIEVAL SYSTEMS

The rationale for extending the original PRP-based retrieval approach to conventional Boolean retrieval systems is based on the well-known result of Boolean algebra (see, e.g., [10]), which states that each element of such an algebra can be represented in a unique form referred to as its disjunctive normal form (DNF). From this result, it follows that any Boolean SRF can be uniquely transformed into the disjunction of a number of distinct expressions, each of which is a conjunction of all the query's index term-related Boolean variables, where each variable occurs once, either negated or unnegated.

Accordingly, the output retrieved by a conventional Boolean search strategy can be viewed as the union of a number of disjoint document sets, each of which corresponds to one of the query's DNF-conjuncts. This implies that a given document which satisfies the Boolean SRF matches exactly one of the constituent conjuncts, and consequently would be assigned to only one of the document sets. Thus, from the perspective of the Boolean SRF, all the documents matching a particular DNF-conjunct are indistinguishable and have the same representation. More specifically, all the documents in each set are seen as being represented by the index terms whose Boolean variables appear unnegated in the corresponding conjunct. Accordingly, each document set can again be characterized by a hypothetical document represented by these index terms. The PRP-based scheme can now be applied to rank the hypothetical documents in descending order of their probabilities of relevance. In estimating such a probability, the index terms which are associated with the negated Boolean variables in a given conjunct are to be regarded as absent from the representation of the corresponding hypothetical document. Since there is a one-to-one correspondence between the hypothetical documents and the respective document sets, these sets can be ranked according to the rank values of the hypothetical documents. Thus, by calculating the rank values of the hypothetical documents, we can determine a partial ranking of the output documents in descending order of their probabilities of relevance. Therefore, it can be concluded that the original PRP-based approach can be extended to conventional Boolean retrieval systems without changing any of their underlying fundamental principles.

To further clarify this theoretical framework, let us consider two query representations: (i) a Boolean SRF; and (ii) a binary SRF (a set of index terms), where each term in the binary SRF corresponds to a variable occurring in the Boolean SRF (there is a one-to-one mapping from the set of index terms onto the set of Boolean variables). Assuming that the Boolean and binary SRFs are processed against the same document collection by the extended and original PRP-based retrieval methods, respectively, each of the constituent output document sets for the Boolean query also occurs as one of the output's constituent document sets for the binary SR, and both of these sets have equal
Moreover, the respective rankings of the output document sets would be identical if the Boolean SRF was a disjunction of all the possible conjuncts formed from the Boolean SRF's variables, excluding the conjunct in which all the variables are negated, or equivalently, a disjunction of all the Boolean SRF's variables. Therefore, the extended ranking scheme can be regarded as more general than the original.

In order to illustrate this relationship between the extended and the original PRP-based retrieval models, let us consider the Boolean SRF, \( q_b = (t_1 \lor t_2) \) and not \( t_3 \), where \( t_1 \), \( t_2 \) and \( t_3 \) are the Boolean variables corresponding to the index terms \( t_1 \), \( t_2 \), and \( t_3 \), respectively, while \( \lor \), and \( \land \), and \( \neg \) denote the Boolean operations of disjunction, conjunction, and negation, respectively. Accordingly, the disjunctive normal form \( (q_b)_X \), of \( q_b \) for the Boolean set \( X = \{ t_1, t_2, t_3 \} \) is:

\[
(q_b)_X = \neg t_1 \land t_2 \land \neg t_3 \lor (t_1 \land \neg t_2 \land \neg t_3) \land (t_1 \land t_2 \land \neg t_3) \, .
\]

Thus, if \( q_b \) is processed against a collection of documents represented by sets of unweighted index terms, the resulting output can be viewed as the union of the three disjoint document sets: (i) the first containing those documents indexed by \( t_1 \) but not \( t_2 \) and \( t_3 \); (ii) the second containing those indexed by \( t_2 \) but not \( t_1 \) and \( t_3 \); and (iii) the third containing those indexed by \( t_1 \) and \( t_2 \) but not \( t_3 \). If the query is represented by the binary SRF, \( q_b = \{ t_1, t_2, t_3 \} \), then the ranked output generated by the original PRP-based procedure would include all of the document sets generated by the extended procedure with the same rank values assigned. Clearly, if \( q_b = t_1 \lor t_2 \lor t_3 \), then the resulting disjunctive normal form (for the same set \( X \)) would consist of seven conjuncts, and the ranked document sets, as well as their rank values, would be identical to those obtained for \( q_b \).

As can be seen from the above discussion, the extended Boolean retrieval scheme allows a wide range of user abilities in specifying an information need. One extreme is where the searcher can precisely represent such a need by providing a list of terms, as well as the logical relationships between them, and the other extreme is where the user is only able to furnish a list of index terms (without any of the logical relationships). Thus the extended scheme is considerably more flexible in terms of system-user communication than the conventional Boolean scheme. This is another significant advantageous feature of the suggested enhancement.

Any procedure that can be used to estimate the probability of relevance of a document represented by unweighted index terms could also be applied to a hypothetical document to determine its rank value. One such procedure, which is discussed below, is based on Bayesian decision theory and utilizes information concerning various frequency distributions of index terms within the document collection (see, e.g., Refs. [3-6]).

In the adopted procedure, a discriminant function, \( g(d_H) \), plays a central role in ranking the output documents. This function can be concisely expressed as:

\[
g(d_H) = \sum_{i=1}^{k} w_i x_i d_H + C,
\]

Provided, of course, that only those Boolean variables that occur in the Boolean SRF have been taken into account by the extended ranking scheme.
In this formula:

\[ w_i \] is the so-called relevance weight of the index term, \( t_i \), corresponding to the

\[ \text{ith} \] variable in the Boolean SRF (the number of relevance weights which are

to be calculated with respect to a given query is generally equal to the

number of variables in the Boolean SRF);

\[ x_d(i) \] is the value of the index term, \( t_i \), with regard to a hypothetical

\( d_H \) document, \( d_H \), which can take one of two values:

- 1 if the index term, \( t_i \), is present in the hypothetical document

representation \( d_H \), or equivalently, if the \( i \)th term-related Boolean

variable occurs unnegated in the corresponding conjunct;

- 0 if the index term, \( t_i \), absent from the hypothetical document

representation \( d_H \), or equivalently, if the \( i \)th term-related Boolean

variable occurs negated in the corresponding conjunct;

\( C \) is a constant, interpreted as the cutoff value \([6]\), which is the same for

all the hypothetical documents (and therefore all of the documents in the

collection) with respect to a given query.

In order to calculate the value of \( g(d_H) \) with regard to a particular hypothetical
document \( d_H \), the summation of the \( w_i \)'s for those index terms which are

present \( (x_d(i) = 1) \) in its representation, \( d_H \), is to be performed (or equivalently,

the summation of the \( w_i \)'s for those terms which correspond to the unnegated

Boolean variables in the associated conjunct). An implicit role of the discriminant

function is to classify each hypothetical document (and the corresponding
document set) into one of two classes: either the class of relevant documents or

the class of nonrelevant documents. Specifically, the decision rule is to classify a

given hypothetical document as relevant if \( g(d_H) > 0 \), and as non-relevant if

\( g(d_H) \leq 0 \). However, the principal role of the discriminant function is to rank

hypothetical documents, and simultaneously the retrieved documents, in decreasing

order of their probabilities of usefulness. More specifically, the greater the value

of \( g(d_H) \), the higher the probability of relevance of the hypothetical document,

\( d_H \), and thus all the hypothetical documents can be ranked accordingly. Since the

value of the constant \( C \) is the same for all of the documents in the collection,
it will not affect the resulting ranking (although it may be a factor in the

classification process). Accordingly, this constant can be ignored in generating the
ranked output, and will not be considered further in this paper. This will

considerably simplify further discussions without affecting the final conclusions.

Thus, the rank value of a hypothetical document, \( d_H \), can be determined by

\[ g(d_H) = \sum_{i=1}^{k} w_i x_d(i) + C. \]

The formula for \( w_i \) can be shown to be in the following

form \([5-6]\):

\[ w_i = \log \frac{\xi_i(1 - n_i)}{n_i(1 - \xi_i)}, \]

where \( \xi_i \) is the probability of occurrence of the index term, \( t_i \), in a DR given
that it represents a relevant document, and $p_i$ is the probability of occurrence of $t_i$ in a DR representing a non-relevant item. Thus, in order to calculate the $w_i$'s it is necessary to estimate the probabilities $\xi_i$ and $\eta_i$, for each index term, $t_i$, associated with the Boolean SRF. One possibility is to use the occurrence characteristics of $t_i$, for a sample document set that is representative of the entire collection, and estimate these probabilities in terms of simple proportions. Specifically, let $N$ be the total number of documents in the representative sample, in which $R$ items have been assessed as relevant by the user. Thus, it follows that there are $N - R$ non-relevant items in the sample. Furthermore, let $n_i$ denote the number of sample documents (relevant or not) which are indexed by the term $t_i$, and let $r_i$ stand for the number of relevant sample documents indexed by $t_i$. Using this notation, the occurrence characteristics for a particular index term, $t_i$, can be presented conveniently by means of a contingency table \[\begin{array}{c|cc}
Relevant & Non-relevant \\
\hline
r_i & n_i - r_i \\
N - r_i & N - n_i - R + r_i \\
\hline
R & N - R & N
\end{array}\] as follows:

Accordingly, $\xi_i$ and $\eta_i$ can be estimated by the following ratios:

$$\xi_i = \frac{r_i}{R},$$

and

$$\eta_i = \frac{n_i - r_i}{N - R}.$$

Thus, the formula for $w_i$ can now be expressed as:

$$w_i = \log \frac{r_i / (R - r_i)}{(n_i - r_i) / (N - n_i - R + r_i)}.$$

This is in fact the well-known index term weighting formula referred to as $F^4$ by Robertson and Sparck Jones \[5\]. The following simple example illustrates the use of the adopted $F^4$ formula in determining the relevance weights for index terms, and subsequently generating the ranked output from a given Boolean SRF.
Example. Suppose a query submitted to a conventional Boolean retrieval system is represented by the Boolean SRF, \( q_B \), as follows:

\[
q_B = (t_1 \text{ and not } t_2) \text{ or } (t_2 \text{ and } t_3).
\]

Thus, the disjunctive normal form, \((q_B)_X\), for the set \( X = \{t_1, t_2, t_3\} \) is:

\[
(q_B)_X = (\text{not } t_1 \text{ and } t_2 \text{ and } t_3) \text{ or } (t_1 \text{ and not } t_2 \text{ and not } t_3) \text{ or } (t_1 \text{ and not } t_2 \text{ and } t_3) \text{ or } (t_1 \text{ and } t_2 \text{ and } t_3).
\]

Accordingly, the retrieved output, \( \gamma(q_B) \), may equivalently be viewed as the union of the disjoint document sets, \( A_1, A_2, A_3, A_4 \), corresponding to the conjuncts: (i) not \( t_1 \) and \( t_2 \) and \( t_3 \); (ii) \( t_1 \) and not \( t_2 \) and not \( t_3 \); (iii) \( t_1 \) and not \( t_2 \) and \( t_3 \); and (iv) \( t_1 \) and \( t_2 \) and \( t_3 \), respectively. That is,

\[
\gamma(q_B) = A_1 \cup A_2 \cup A_3 \cup A_4.
\]

The first set, \( A_1 \), contains those documents in the collection which are indexed by the terms \( t_2 \) and \( t_3 \) but not \( t_1 \) (it is of no importance whether the items are indexed by any of the other terms); \( A_2 \) contains those documents which are indexed by \( t_1 \) but not \( t_2 \) and \( t_3 \); and so forth.

As already mentioned, from the perspective of a conventional Boolean search scheme, all the items in an output's constituent document set are indistinguishable. In other words, the documents contained in such a set are considered by the system to be represented by the same set of index terms (although, in fact, they may be indexed by other terms that are not referenced in the query). Accordingly, the documents in \( A_1, A_2, A_3, A_4 \) are viewed as being characterized by the index term sets \( \{t_2, t_3\} \), \( \{t_1\} \), \( \{t_1, t_3\} \), and \( \{t_1, t_2, t_3\} \), respectively. In compliance with the extended document ranking scheme, these sets of index terms can be regarded as the representations of the hypothetical documents, \( d_{H_1}, d_{H_2}, d_{H_3}, d_{H_4} \), respectively. Therefore, the system's output can be viewed as the set of four hypothetical documents. Formally,

\[
\gamma(q_B) = \{d_{H_1}, d_{H_2}, d_{H_3}, d_{H_4}\}.
\]

We can now apply the adopted \( F_4 \) formula to determine the rank values of these hypothetical documents in terms of their probabilities of relevance. Since there exists a one-to-one correspondence between the hypothetical documents and the output's constituent document sets, the rank values of the hypothetical documents are also the rank values of the associated document sets. Thus, having determined the rank values of the hypothetical documents, we can determine the ranking of the corresponding output's document sets in descending order of their
probabilities of relevance. In order to use the F4 formula, a representative
sample of the document collection is required. Assume that the sample document
set consists of 10 items, \( d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10} \), with the
following representations:

\[
d_1 = \{t_2, t_3, t_4\}, \\
d_2 = \{t_1, t_3\}, \\
d_3 = \{t_2, t_3, t_4\}, \\
d_4 = \{t_1, t_2\}, \\
d_5 = \{t_2, t_3\}, \\
d_6 = \{t_1, t_2, t_4\}, \\
d_7 = \{t_3\}, \\
d_8 = \{t_2\}, \\
d_9 = \{t_1, t_3\}, \\
d_{10} = \{t_1, t_2\}.
\]

Let us further assume that of the sample documents, \( d_1, d_4, d_6, d_8 \) and \( d_9 \) have
been assessed as relevant by the user.

This information is sufficient to determine the relevance weights for the
individual index terms associated with \( q_0 \). Specifically, for the term \( t_1 \), we have
\( r_1 = 3 \) and \( n_1 = 5 \). Accordingly, the related contingency table is \( (N = 10, R = 5, \\
r_1 = 3, n_1 = 5) \):

<table>
<thead>
<tr>
<th>Relevant</th>
<th>Non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Thus,

\[
w_1 = \log \frac{3}{2} \frac{2}{3} = \log \frac{9}{4} = 0.3522.
\]

The contingency table for \( t_2 \) is \( (N = 10, R = 5, r_2 = 4, n_2 = 7) \):

<table>
<thead>
<tr>
<th>Relevant</th>
<th>Non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Hence,

\[
w_2 = \log \frac{4/1}{3/2} = \log \frac{8}{3} = 0.4260.
\]
Finally, the contingency table for $t_3$ is ($N = 10$, $R = 5$, $r_3 = 2$, $n_3 = 6$):

<table>
<thead>
<tr>
<th>Relevant</th>
<th>Non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{d_H}(3) = 1$</td>
<td>2</td>
</tr>
<tr>
<td>$X_{d_H}(3) = 0$</td>
<td>3</td>
</tr>
</tbody>
</table>

Therefore,

$$w_3 = \log \frac{2/3}{6/7} = \log \frac{2}{12} = \log \frac{1}{6} = -0.7782.$$

These index term relevance weights will now allow us to calculate a rank value, $R_i$, for each hypothetical document, $d_{H_i}$, and simultaneously for the corresponding output document set, $A_i$.

These rank values are:

$$R_1 = -0.3522,$$

$$R_2 = 0.3522,$$

$$R_3 = -0.4260,$$

and

$$R_q = 0.0000.$$

This implies the following ranking, $\text{Ord } \gamma(q)$: $\text{Ord } \gamma(q) = < A_2, A_3, A_1, A_0 >$.

The $A_2$ document set is presumably the most likely to contain relevant documents. Thus, it should be presented to the user first, followed, if needed, by the next highest ranked document set, $A_3$, and so forth.

**CONCLUDING REMARKS**

This paper has presented a theoretical framework for incorporating a PRP-based document ranking mechanism into a conventional Boolean retrieval system. This extended system exhibits a number of advantages which include:

(i) accommodating traditional Boolean SRF's used in commercial retrieval systems without changing any of the fundamental principles upon which these systems are based;

(ii) providing for the retrieval of documents in descending order of their probabilities of usefulness to the user;

(iii) facilitating a system-user adaptation mechanism by determining relevance weights for a query's index terms;
allowing for a wide range of user abilities in specifying an information need.

To summarize, by demonstrating the suitability of Boolean logic for exploiting the Probability Ranking Principle, it has been shown that the implementation of an extended document ranking procedure should result in considerable improvement in the retrieval effectiveness of traditional search techniques at an acceptable cost.

BIBLIOGRAPHY