Information Retrieval, 3, 217–227, 2000


# The Probability Ranking Principle Revisited

MARTIN WECHSLER martin_wechsler@mckinsey.com
*McKinsey & Company, Switzerland*

PETER SCHÄUBLE schauble@eurospider.com
*Eurospider Information Technology AG, Zürich, Switzerland*

*Received August 30, 1999; Revised August 30, 1999; Accepted June 12, 2000*

**Abstract.** A theoretic framework for multimedia information retrieval is introduced which guarantees optimal retrieval effectiveness. In particular, a Ranking Principle for Distributed Multimedia-Documents (RPDM) is described together with an algorithm that satisfies this principle. Finally, the RPDM is shown to be a generalization of the Probability Ranking principle (PRP) which guarantees optimal retrieval effectiveness in the case of text document retrieval. The PRP justifies theoretically the relevance ranking adopted by modern search engines. In contrast to the classical PRP, the new RPDM takes into account transmission and inspection time, and most importantly, aspectual recall rather than simple recall.

**Keywords:** multimedia information retrieval, probability ranking principle, relevance ranking, optimal search performance, maximum retrieval effectiveness

## 1. Introduction

Multimedia Information Retrieval is becoming more and more feasible because both speech and image recognition methods have been improved significantly during the last years: A wealth of new information access techniques have been developed to find relevant information in large multimedia data collections (Schäuble 1997). While some of these new techniques—for example retrieval techniques for digitized speech documents (Wechsler 1998)—have been evaluated experimentally, they were hardly developed and studied within a theoretic framework that guarantees optimal retrieval effectiveness; in fact we lack such a framework that optimizes the probability that a user finds the desired information in a large multimedia document collection. In this paper, a Ranking Principle for Distributed Multimedia (RPDM) document collections is described that serves as a theoretic framework like the Probability Ranking Principle (PRP) by Robertson (1977) for centrally stored text documents. The PRP states that *a retrieval system performs optimally if the documents are ranked according to decreasing probabilities of relevance.* He showed that optimal performance can be expressed either in terms of precision and recall, or in terms of costs associated with the retrieval of non-relevant documents and the non-retrieval of relevant documents. We elaborate on the PRP in Section 5.

Nowadays, large *distributed multimedia* document collections are available. Local and global computer networks, for example the World-Wide Web, allow quick access and transfer of documents independent of their location. Further, new technologies enable digital

processing of non-text media, such as images, audio and video, which collectively comprises the notion of *multimedia*.

## 2. Criteria for distributed multimedia documents

For IR systems managing distributed multimedia documents, we believe that the optimal document ranking problem has to be revised. Such IR systems should allow for *additional criteria* other than solely the probability of relevance when suggesting an inspection order for documents with respect to a query. We identify the following two additional criteria:

1. *Transmission time* of a document, the time needed to transport a document from the source location across the network to the user.
2. *Inspection time* of a document, the time needed by the user to inspect a document.

We demonstrate the importance of these criteria by the following two examples that are also illustrated in figure 1: (1) Assume that two documents $d_j$ (text) and $d_k$ (text) have an equal probability of relevance to a given request. If $d_j$ is geographically closer to the user than $d_k$, then obviously $d_j$ should be transmitted and presented before $d_k$, such that the user's idle waiting time is minimized. (2) Assume that two documents $d_k$ and $d_l$ contain the same information but are from different media, say $d_k$ is text and $d_l$ is audio or video. In this case the IR system should favor the ranking of the text document first, since it is much faster both to transmit and to inspect.

The transmission time is affected (1) by the geographical distance, (2) by the network bandwidth, and (3) by the document's storage size, which depends on the document's medium and length. The media text, audio and video are orders of magnitude apart from each other with regard to their associated data rates. Table 1 shows the data rates for different (un) compressed formats of the three media (Lu 1996, p. 49, 108): The text data-rate is

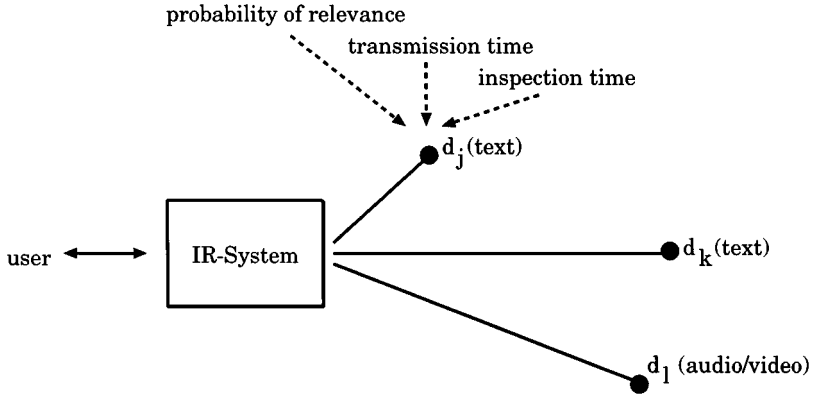


*Figure 1*. Retrieval of distributed multimedia documents and parameters affecting optimal ranking.

*Table 1*. Data rates for different compressed and uncompressed media.

| Document medium | Data format | Data rate (kByte/s) |
|---|---|---|
| Text | ASCII | 0.018 |
| Audio | Uncompressed CD-audio | 176.4 |
| Audio | Compressed CD-audio (MPEG-audio) | 29.4 |
| Video | Uncompressed VHS-video | 6750 |
| Video | Compressed VHS-video (MPEG-1) | 187.5 |

based on a speaking rate of 120 words per minute and an average word-length of 9 letters. For audio and video, sophisticated compression techniques such as MPEG (Pan 1995) make a data reduction possible, however the differences between the media are still considerable.

The inspection time is dependent on the medium and the length of the document as well. Due to the time-synchronous nature of audio and video, these media require much more time for inspection compared with for example an equivalent text document. Also, it has been found experimentally that users require more time to extract requested information from a video than from a text (Sutcliffe et al. 1997).

## 3. Ranking principle

Before we study the optimal ranking problem in an IR system, we make two general assumptions: First, we assume that the IR system is able to return (query-dependent) *passages* rather than documents in response to a query. Passages are motivated mainly by the fact that entire audio and video documents require very much time for transmission and inspection, and by the fact that a user would like to listen only to relevant parts of a (maybe long) recording. The second assumption is that a query $q$ consists of $k$ aspects $q_i$, $i = 1, \ldots, k$. For example, the query "W.A. Mozart" may ask for information about Mozart's compositions but also about his life in society.

To study the optimal ranking problem for distributed multimedia documents we assume that a user specifies a *total inspection time* when he submits a query. We further assume that, after query evaluation, the IR system has the following parameters available for each passage: (1) the probability of relevance, which is estimated by any retrieval method, (2) the transmission time, and (3) the inspection time.

**Ranking Principle for Distributed Multimedia Documents (RPDM)**

An IR system should present passages of distributed multimedia documents to a user query in such a way that

1. The passages can be inspected within the user-specified total inspection time,
2. There is no user waiting time between the inspection of passages due to their transmission,
3. The passages contain "a maximum amount of relevant information about various aspects of the query".

More formally, we formulate the RPDM as an optimization problem. Given are the

- total inspection time $T$ specified by the user,
- the set $X$ of passages $x_j, j = 1, \ldots, |X|$ found to query $q$,
- transmission time $t^t(x_j)$ for each passage $x_j$,
- inspection time $t^i(x_j)$ for each passage $x_j$,
- probabilities of relevance $P(R \mid q_i, x_j)$ for each passage $x_j$ with regard to a query aspect $q_i$.

The problem for the IR system is to find a sequence $Y = \langle y_1, \ldots, y_n \rangle$ of passages $y_j \in X$ such that the following conditions are satisfied:

$$t^t(y_1) + \sum_{j=1}^{n} t^i(y_j) \leq T, \tag{1}$$

$$t^t(y_j) \leq t^t(y_1) + \sum_{l=1}^{j-1} t^i(y_l) \quad \forall_j : 1 \leq j \leq n, \tag{2}$$

$$C(Y) := \sum_{i=1}^{k} \prod_{y \in Y} (1 - P(R \mid q_i, y)) \stackrel{!}{=} \min. \tag{3}$$

The first condition (1) requires that all selected passages can be transmitted and inspected within the total inspection time $T$.

The second condition (2) assures that the next passages to be inspected can be transmitted completely within the time used for the transmission of the first passages and the inspection of all preceding passages. Here we assume that the IR system is capable of requesting future passages in the background while the user is inspecting. This eliminates waiting time due to the transmission of passages. Figure 2 illustrates two time scenarios of the user's inspection process, where condition (2) is satisfied only in scenario (a).

Definition 3 is the main *cost function* which has to be *minimized*. We justify the choice for the cost function as follows: The term $(1 - P(R \mid q_i, y))$ denotes the probability that passage $y$ does *not* cover aspect $q_i$, and thus the product denotes the probability that aspect $q_i$ is not at all covered in the selected passages. If we associate constant costs for each aspect not covered, the cost function $C(Y)$ is proportional to the expected costs for missing relevant aspects of the query. In other words, the IR system should suggest passages that optimally cover the most aspects to the given query. Our cost function is inversely proportional to the aspectual recall used in TREC's interactive track (Voorhees and Harman 1999). In the next section we present an algorithm that solves this optimization problem.

## 4. An algorithm satisfying the RPDM

The simplest algorithm to the problem formulated in the previous section is a *backtracking procedure* (Nievergelt 1977, Kreher and Stinson 1998). The idea of backtracking applied to our situation is to *enumerate all* sequences of passages (i.e. possible solutions) while (1.)

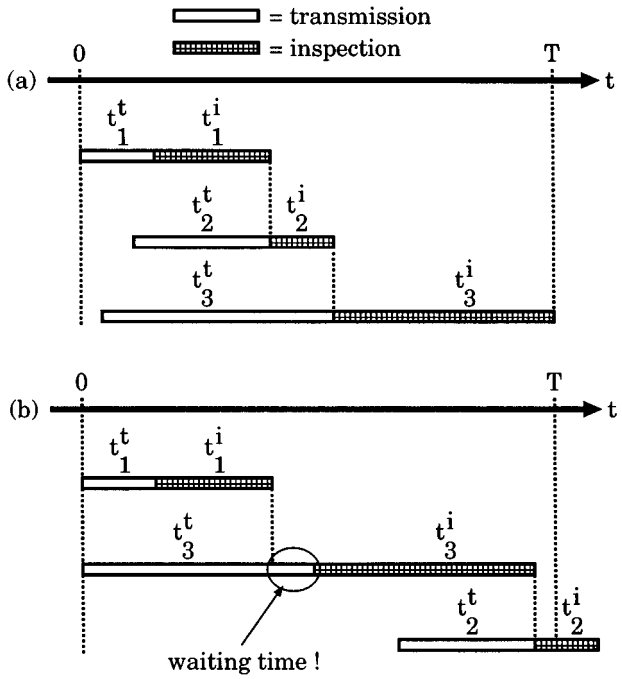


*Figure 2.* User inspection process for three passages. Future passages are transmitted in the background. (a) No user waiting time, (b) user waiting time between the first and second passage. Note also that in (b), the last passage cannot be inspected completely.

checking both conditions (1) and (2) for each sequence, (2.) computing its cost function $C(Y)$, and (3.) retaining the sequence that obtains minimal costs.

We show, however, that backtracking is not feasible for this problem since the number of solutions to be considered is far too high. Let $N = |X|$ be the total number of possible passages that have been found to a given query, and assume that the algorithm selects $n$ passages. The number of solutions with $n$ passages can be derived by combining $n$ out of $N$ passages and by permutating those sequences, which results in

$$\binom{N}{n} \cdot n! = \frac{N!}{(N-n)!}$$

solutions. Since $n$ may vary from 1 to $N$, the *total number of solutions* to be considered is

$$\sum_{n=1}^{N} \frac{N!}{(N-n)!}.$$

For $N = 10$ this value is $\approx 10^7$ and for $N = 100$ it is $\approx 2 \cdot 10^{158}$.

To reduce the number of solutions to consider, we propose the classic *branch-and-bound algorithm* (e.g. Domschke and Drexl (1991, p. 114–119)), which is a variant of

the backtracking algorithm. The idea of branch-and-bound is (1) to enumerate solutions with lowest *expected cost-values* first (branch), and (2) to discard entire sets of solutions where costs can certainly not be further reduced (bound). The pseudo-code of the algorithm is given in figure 3. The algorithm is shown as a recursive procedure *branch-and-bound* that calculates the cost function for a given sequence $Y$ and tries to extend this sequence with an additional passage at the end, before it is called recursively. The main call of this procedure is *branch-and-bound*($\epsilon$) where $\epsilon$ denotes the empty sequence. Figure 4 illustrates the extension of the sequence $Y$ graphically. The circles correspond to solutions and the arrows denote extensions with a single passage.

In line 7 of the algorithm, the costs $C_Y$ of the current solution are calculated according to the cost function (Definition 3). Then, the currently best solution $Y_{best}$ is updated if necessary. The rest of the procedure considers extensions of $Y$ with an additional passage $x$. We write $\langle Y, x\rangle$ for a new sequence consisting of $Y$ and an appended passage $x$. Also, we define $Y_x^+$ as the set of all sequences that consist of the sequence $Y$, an appended passage $x$ and any number of further passages. In lines 11–16, for each extension passage $x$ and each resulting set $Y_x^+$, a *lower bound of costs* LBC($Y_x^+$) is calculated (line 13, see next paragraph), and the extension passage is retained in an extension set $E_Y$ (line 14). However, this is only performed if two conditions are met (line 12): The procedure *within-user-time* $(Y, x)$ tests the condition (1), and *no-waiting* $(Y, x)$ tests the condition (2). Finally, in lines 17–23 the extension with the lowest LBC-value is pursued by a recursive call of *branch-and-bound* (branch). In line 18, the best passage is selected for extension, and simultaneously, extensions are only pursued if their costs (represented by the lower bound costs) may become smaller than the costs of the currently best solution $C_{best}$ (bound).

In the following, we derive a *lower bound of costs* LBC($Y_x^+$) for all solutions consisting of the solution $Y$ extended with a passage $x$ and any number of further passages: According to the cost function (Definition 3) we write for the costs of solution $Y$

$$C(Y) := \underbrace{\prod_{y\in Y}(1 - P(R \mid q_1, y))}_{P(q_1)} + \cdots + \underbrace{\prod_{y\in Y}(1 - P(R \mid q_k, y))}_{P(q_k)}$$

where $P(q_i)$ abbreviates a product term. If we extend the solution $Y$ with a passage $x$, it holds that

$$\begin{aligned} C(\langle Y, x\rangle) &:= P(q_1)\cdot(1 - P(R \mid q_1, x)) + \cdots + P(q_k)\cdot(1 - P(R \mid q_k, x)) \\ &\geq C(Y)\cdot \min_{i=1,\ldots,k}\{1 - P(R \mid q_1, x)\}. \end{aligned} \tag{4}$$

.

Now we consider any solution $Y' \in Y_x^+$, which starts with the sequence $Y$ followed by the passage $x$ and any number of further passages. From Eq. 4 follows:

$$\begin{aligned} &C(Y') \geq C(Y)\cdot \min_{i=1\ldots k}\{1 - P(R \mid q_i, x)\}. \\ &\underbrace{(\min\{1 - P(R \mid q_i, x') \mid i = 1, \ldots, k \wedge x' \in X\backslash \text{set}(\langle Y, x\rangle)\})^m}_{=: r(Y,x)}, \end{aligned} \tag{5}$$

```
    VAR
      Y_best; /* currently best solution */
      C_best; /* costs of best solution */

1   PROCEDURE branch_and_bound( Sequence Y )
2     VAR
3       C_Y; /* costs for solution Y */
4       LBC(Y_x^+); /* lower bound costs for extensions of ⟨Y, x⟩ */
5       E_Y; /* set of possible passage extensions of Y */
6
7     C_Y := C(Y);
8     IF C_Y < C_best
9       Y_best := Y; C_best := C_Y
10    END;
11    FOR ALL x ∈ X \ set(Y)
12      IF within_user_time(⟨Y, x⟩) ∧ no_waiting(⟨Y, x⟩)
13        calculate LBC(Y_x^+);
14        E_Y := E_Y ∪ {x};
15      END
16    END;
17    LOOP
18      x := argmin{LBC(Y_x^+) | x ∈ E_Y ∧ LBC(Y_x^+) < C_best}
19      IF x undefined
20        BREAK;
21      END
22      branch_and_bound(⟨Y, x⟩);
23    END;
24    RETURN
25  END PROCEDURE;

    /* INPUT: */
    /* X: set of passages found to a request */
    /* t^t(x): transmission times */
    /* t^i(x): inspection times */
    /* P(R|q_i, x): probabilities of relevance */
    /* T: total user inspection time */
    MAIN
      Y_best := ϵ; C_best := ∞;
      branch_and_bound(ϵ);        3
    END
    /* OUTPUT: Y_best */
```

*Figure 3.* Recursive branch-and-bound algorithm satisfying the RPDM.

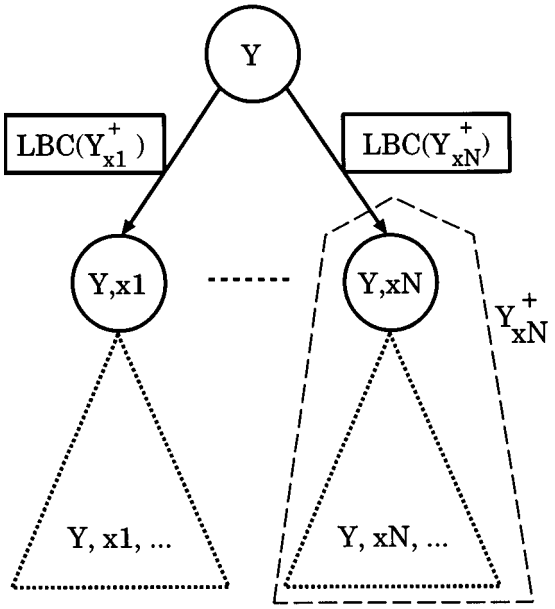


*Figure 4.* Search tree of the branch-and-bound algorithm. Each node represents a solution. Successor nodes are extensions of the solution $Y$. The LBC-values are used either to select an extension as the next solution, or to discard an entire subtree.

where $r(Y, x)$ is an abbreviation and $m$ denotes the maximum number of passages that may be added to the sequence $(Y, x)$. This number is limited by the remaining inspection time and by the minimum inspection time of the passages not yet selected. Let $t_Y$ be the time necessary to inspect all passages of the solution $Y$. Then we can formulate an upper bound for $m$ by

$$m \leq \left\lfloor \frac{T - t_Y - t^i(x)}{\min\{t^i(x') \mid x' \in X \backslash \text{set}(\langle Y, x \rangle)\}} \right\rfloor.$$

Thus, the right part of Eq. 5 contains a lower bound for the costs of a solution $Y' \in Y_x^+$ and therefore we can define

$$\text{LBC}(Y_x^+) := C(Y) \cdot \min_{i=1\cdots k} \{1 - P(R \mid q_i, x)\} \cdot r(Y, x).$$

The lower bound costs denote the minimum possible costs of any solution extended from a solution $Y$ and the passage $x$. If they are larger than the currently minimum costs, then the entire search subtree $\langle Y, x \rangle$ (figure 4) may be discarded.

The output of the branch-and-bound algorithm satisfies the RPDM as stated on page 2. A crucial element of the branch-and-bound algorithm is the derivation of a highest possible lower bound for the costs of extended solutions. The higher the lower bound is, the more

efficient the algorithm is because the number of solutions to consider is reduced. However, more accurate lower bounds are usually more complex to compute. This results in a trade-off between the effort to compute a lower bound and the effort to consider more solutions.

In a practical situation it may be that the IR system does not have enough time to run the branch-and-bound algorithm because the search tree is still too large. For such cases we propose the use of suboptimal stochastic optimization techniques (e.g. simulated annealing (Vidal 1993) or genetic algorithms (Mitchell 1996)). These approaches do not guarantee that the best solution will be found, but they may find a sufficiently good solution in a short time.

## 5. The RPDM and the PRP

In this section we show that our ranking principle for distributed multimedia collections is compatible to the *Probability Ranking Principle* (PRP) which was developed in the context of retrieval in rather short texts and bibliographic records (Robertson 1977).

We map our distributed multimedia model into the context of the PRP, which originally was to retrieve bibliographic references or (rather short) text documents. We make the following assumptions:

- We do not deal with passages but with documents because the (text) documents are very short (less than 1 page).
- The transmission time of each document is neglected ($t^t(d_j) = 0 \ \forall j$) because each document consists of only a few text words.
- The inspection time $t^i(d_j)$ for each document is constant, say $\Delta t$, because the short text documents are able to be inspected at a glance.
- There is only one aspect in the query.

Again, we assume that the user specifies a total inspection time $T$. With constant inspection times, it follows that the user inspects $n := \lfloor \frac{T}{\Delta t} \rfloor$ documents. Let $Y$ be the sequence of those $n$ documents. Since the transmission times are neglected, the documents of $Y$ may be presented in any order.

The cost function (Definition 3) is simplified due to the presence of only one aspect:

$$C(Y) := \prod_{d_j \in Y} (1 - P(R \mid q, d_j)). \tag{6}$$

**Proposition.** *If the solution $Y$ contains the top $n$ documents with respect to $P(R \mid q, d_j)$, $d_j \in Y$ (and thus conforms to the PRP), it follows that its costs $C(Y)$ are minimal.*

**Proof (by contradiction):** Assume that the solution $Y$ contains the top $n$ documents, and that its costs are *not* minimal. We can choose any document $d \in Y$ and replace it by any other document $d' \notin Y$ form the rest of the collection. According to Definition 6 the costs

of this new set $Y'$ are

$$C(Y') := \frac{1 - P(R \mid q, d')}{1 - P(R \mid q, d)} \cdot \prod_{d_j \in Y} (1 - P(R \mid q, d_j))$$
$$= \underbrace{\frac{1 - P(R \mid q, d')}{1 - P(R \mid q, d)}}_{=: f} \cdot C(Y).$$

Since $P(R \mid q, d) \geq P(R \mid q, d')$ it follows $f \geq 1$ and thus $C(Y)' \geq C(Y)$ or $C(Y)$ is minimal. □

The statement made in the proposition conforms to the Probability Ranking Principle under the assumption that the number of documents to inspect is known *a priori*. Thus we have shown that the context of the Probability Ranking Principle is a special case of our RPDM, and that the PRP is satisfied if the RPDM is satisfied.

## 6. Conclusions

We have introduced a new *ranking principle for distributed multimedia documents* (RPDM) and we have shown that the RPDM is a generalization of the probability ranking principle. In contrast to the classic principle, the RPDM takes into account the user's context in a more detailed way: First, the user's information need is structured into query aspects. The RPDM is aimed at covering all aspects as good as possible whereas the classic principle is aimed at retrieving relevant and only relevant documents even when the top ranked documents cover but a single aspect. Furthermore, the RPDM takes into account transmission and inspection time. This way, text is preferred to other media types because of its short transmission and inspection times. Finally, we would like to point out that the RPDM can easily be extended by further criteria such as cost for transmission, cost for content and QoS (Quality of Service) aspects.

## References

Domschke W and Drexl A (1991) Einführung in Operations Research. Springer, Berlin.
Kreher D and Stinson D (1998) Combinatorial Algorithms: Generation, Enumeration and Search. CRC Press, Boca Raton.
Lu G (1996) Communication and Computing for Distributed Multimedia Systems. Artech House, Boston.
Mitchell M (1996) An Introduction to Genetic Algorithms. The MIT Press, Cambridge.
Nievergelt J (1977) Combinatorial Algorithms. Prentice Hall, Englewood Cliffs.
Pan D (1995) A tutorial on MPEG/audio compression. IEEE Multimedia, 2(2):60–74.
Robertson SE (1977) The probability ranking principle in IR. Journal of Documentation, 33(4):294–304.
Schäuble P (1997) Multimedia Information Retrieval—Content-Based Information Retrieval from Large Text and Audio Databases. Kluwer Academic Publishers, Boston.
Sutcliffe A, Hare M, Doubleday A and Ryan M (1997). Empirical studies in multimedia information retrieval. Intelligent Multimedia Information Retrieval, AAAI Press, pp. 449–472.
Vidal R (1993) Applied Simulated Annealing. Springer, Berlin.

Voorhees E and Harman D (1999) Overview of the seventh text retrieval conference (TREC-7). In: TREC-7 Proceedings.

Wechsler M and Schäuble P (1999). A New ranking principle for multimedia information retrieval. In: Proceedings of the Fourth ACM Conference on Digital Libraries.

Wechsler M (1998) Spoken document retrieval based on phoneme recognition. PhD Thesis, ETH Zurich. Diss. No. 12879.