Optimum Database Selection in Networked IR

Norbert Fuhr
University of Dortmund

1 Introduction

Information retrieval (IR) deals with the problem of retrieving documents relevant to a query from a database. However, most IR models (e.g. Boolean, vector space or fuzzy retrieval models) do not refer explicitely to the concept of relevance. Instead, the only thing that can be shown is a statistical correlation between the output results of an application of the model and the relevance judgements given by the users.

Only probabilistic IR models refer directly to the concept of relevance: As described in [Robertson 77], the Probability Ranking Principle underlyig these models can be shown to give optimum retrieval performance. Here performance can be measured either in terms of precision and recall (which, in turn, refer to relevance), or by means of a decision-theoretic model which attributes different costs to the retrieval of relevant and nonrelevant documents.

In this paper, we make an attempt to apply these ideas to the problem of database selection in networked IR. The basic setting is as follows: A broker has access to a set of IR databases to which it may send a query. In response, each database produces a ranked list of documents, and the broker may request any number of documents from this list. Each database has its own performance curve in terms of recall and precision, and there are database-specific costs for the retrieval of documents. Given a specific query, we now want to retrieve a maximum number of relevant documents at minimum cost, i.e. one of the two parameters is specified by the user, and the broker aims at optimizing the other one.

More specifically, in order to select the databases to be used for processing a query, for each database D_i the broker estimates a function $C_i^r(n)$ giving the specific costs for retrieving n relevant documents from this database. Based on this information, a global function $C^r(n)$ can be derived which specifies the minimum costs for retrieving n relevant documents altogether from the databases.

As basic assumption underlying our approach, all IR systems running the different databases must be based on the probabilistic IR model, thus assigning an estimate of the probability of relevance to each document. If this assumption does not hold for a real system (e.g. for a Boolean system), we may be able to convert the output of such a system into the required form.

2 Cost model

As usual in decision-theoretic models, costs may stand for money as well as for computing time, response time or the time a user spends for doing her job.

On a coarse-grained level, we may assume that there is a cost function $C_i^s(k)$ for retrieving k documents from database D_i. However, in most cases the cost can be split up in fixed costs C_i^0 for processing a query and a factor C_i^d for each document delivered from the query result. So we have

$$C_i^s(k) = C_i^0 + kC_i^d. \tag{1}$$

symbol	meaning
q	query
d	document
D_i	database
R_i	# relevant documents in D_i
r_i	# relevant documents retrieved from D_i
s_i	# documents retrieved (selected) from D_i
$C_i^s(n)$	costs for selecting n documents from D_i
C_i^0	fixed costs for query processing in D_i
C_i^d	costs for retrieving a document from D_i
$C_i^r(n)$	costs for retrieving n relevant documents from D_i
C^R	user costs for viewing a relevant document
C^N	user costs for viewing a nonrelevant document
$C^r(n)$	global costs for n relevant documents
R	recall
P	precision
$P_i(R)$	recall-precision function for D_i

Table 1: Notations used throughout this paper

In order to make a statement about relevant documents, we must refer to the recall-precision curve $P_i(R)$ of each database D_i (for the specific query). Assume that we also know the total number of relevant documents R_i for each database. Then the number of documents s_i to be selected in order to retrieve r relevant documents from database D_i follows from $r/s_i = P_i(R) = P_i(r/R_i)$:

$$s_i(r) = \frac{r}{P_i(r/R_i)}. \tag{2}$$

As in traditional IR, we also assume user costs (or benefits) C^R and C^N for a user viewing a relevant document or a nonrelevant document, respectively.

Combining equations 1 and 2, we can compute the overall cost for retrieving r relevant documents from database D_i:

$$\begin{aligned} C_i^r(r) &= C_i^s(s_i(r)) + rC^R + (s_i(r) - r)C^N \\ &= C_i^0 + r(C^R - C^N) + \frac{r}{P(r/R_i)}\left(C_i^d + C^N\right). \end{aligned} \tag{3}$$

Given the database-specific cost functions, we can now compute the overall minimum costs $C^r(n)$ for retrieving n relevant documents. For that, let us assume that we have l databases and a binary vector $\vec{u} = (u_1, \ldots, u_l)$, where u_i denotes whether database D_i is used ($u_i = 1$) or not used ($u_i = 0$) for processing the current query at minimum cost.

Then the overall cost function $C^r(n)$ is defined as

$$C^r(n) = \min_{\vec{u}} \sum_{i=1}^{l} u_i C_i^r(r_i) \tag{4}$$

with the additional constraint

$$n = \sum_{i=1}^{l} u_i r_i.$$

3 Interpretation of results

Now we want to discuss the consequences of the cost functions 3 and 4. For that, we will ignore the fact that $C^r(n)$ is a discrete function and assume it to be continuous. Then we can make some observations that hold for the optimum solution.

First, we can make a general statement about those databases D_i that contribute to the query result, i.e. $u_i = 1$. By using Lagrange multipliers, we find out that

$$\frac{\partial C_i^r(r_i)}{\partial r_i}$$

is equal for all these databases. Roughly speaking, this means that the costs for the last relevant document retrieved from each of the databases involved are equal. Since the recall-precision curve usually is monotonously decreasing, we can conclude that the cost differential is monotonously increasing.

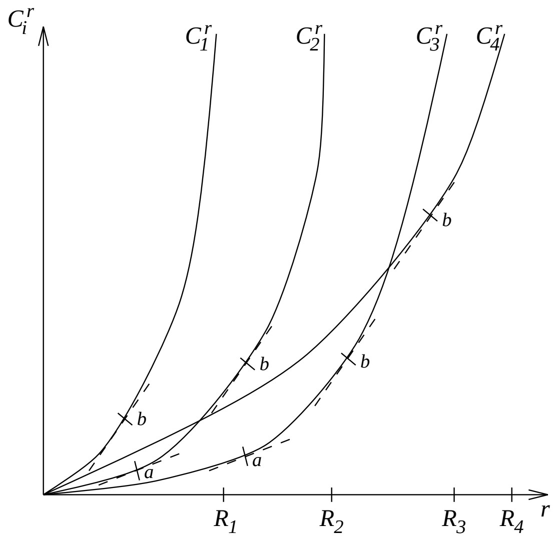

Figure 1: Sample cost functions for $C_i^0 = 0$ with optimum solutions a, b

In order to say more, we have to distinguish certain cases:

1. $C_i^0 = 0$ for $i = 1, \ldots, l$. Assuming that we only have to pay per document delivered, but not for processing the query, we get $C_i^r(0) = 0$ for all databases. Sample functions are depicted in figure 1. Since a specific number of total relevant documents implies an equal slope $\partial C_i^r(r_i)/\partial r_i$ for all curves, all databases for which there is a point with this slope on the curve will contribute to the optimum solution. In figure 1, the points corresponding to two solutions a and b are marked, showing that for the first solution, only two of the databases are involved. The set of databases involved grows as the total number of relevant documents increases; a database contributing to a small number always will stay involved for larger numbers, too. This feature is important for incremental retrieval where a user specifies neither the cost nor the number of relevant documents in advance.
   In addition, if we have equal costs per document C_i^d for all databases, then we can also make statements about recall and precision. In this case, the databases involved operate at the same precision level. Figure 2 shows the points for four different solutions $a, \ldots, d$), where e.g. for b, only databases 1 and 4 reach this precision level.

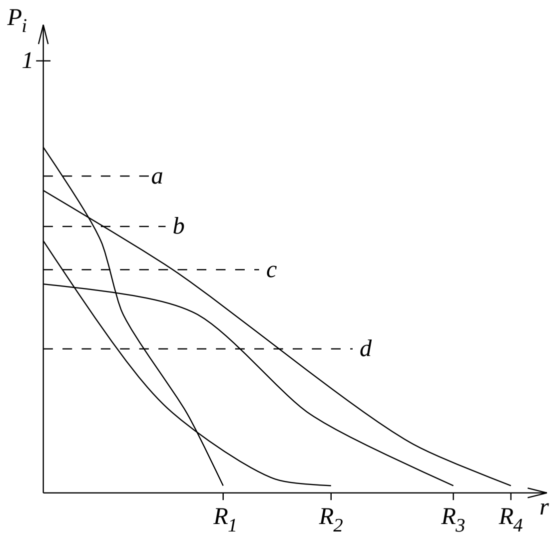

Figure 2: Sample recall-precision curves with optimum solutions $a, \ldots, d$

n	C1		C2		C	
	cost	gdl	cost	gdl	cost	gdl
1	4	$\langle(2,1)\rangle$	5	$\langle(1,2)\rangle$	4	$\langle(2,1)\rangle$
2	7	$\langle(4,1)\rangle$	6	$\langle(3,2)\rangle$	6	$\langle(3,2)\rangle$
3	10	$\langle(6,1)\rangle$	8	$\langle(7,2)\rangle$	8	$\langle(7,2)\rangle$
4	13	$\langle(8,1)\rangle$	11	$\langle(13,2)\rangle$	11	$\langle(13,2)\rangle$
5	16	$\langle(10,1)\rangle$	15	$\langle(21,2)\rangle$	15	$\langle(21,2)\rangle$
6	19	$\langle(12,1)\rangle$	20	$\langle(31,2)\rangle$	18	$\langle(4,1),(13,2)\rangle$
7	22	$\langle(14,1)\rangle$	26	$\langle(43,2)\rangle$	21	$\langle(6,1),(13,2)\rangle$

Table 2: Example for databases with nonzero query processing costs

2. $C_i^0 > 0$ for some $i \in [1, l]$. If there are databases with nonzero query processing costs, then the set of databases that actually contribute to the solution will depend on the total number n of relevant documents. Here databases involved for small values of n may not contribute to the optimum solution as n grows - as in the example in table 2 (here cost is the cost for retrieving n relevant documents and gdl gives pairs (number of documents to be selected, database number)). With regard to incremental retrieval, we have a conflict here: Given that the user first requested n_1 documents and then another n_2 documents, the minimum costs for this stepwise procedure may be higher than for retrieving $n_1 + n_2$ relevant documents at once.

In order to apply the formulas from above, we have to know the recall-precision (RP) curves of the databases plus the number of relevant documents in each database. The latter problem is discussed in [Fuhr 96]. The actual RP curves can hardly be known in advance, so we need some heuristics in order to get a good approximation. For example, a simple assumption would be a linear retrieval function, with $P(0) = P_0$ and $P(1) = 0$, thus leading to the equation $P = P_0(1 - R)$.

In the absence of any query-specific knowledge, one might assume that the RP function is equal for all queries. However, in some cases additional information may be available, e.g. if the query contains a condition which cannot be evaluated by the IR system running a specific database; then already P(0) will be very low. With respect to the different databases, one may start with the assumption that, in general, the RP function is the same for all databases. It also may be feasible to assume functions that are typical for certain kinds of IR systems, e.g. Boolean vs. probabilistic systems.

4 Towards application

In order to apply the cost estimation formulas from section 2, the following steps have to be performed:

1. For each database D_i, estimate the number of relevant documents R_i.
2. For each database D_i, determine (or assume) a recall-precision function $P_i(R)$.
3. Compute the database-specific cost functions $C_i^r(n)$.
4. Derive the global cost function $C^r(n)$ as combination of databases such that, for each value of n, the costs are minimum.

References

Fuhr, N. (1996). *A Decision-Theoretic Approach to Database Selection in Networked IR.* Technical report (http://ls6.informatik.uni-dortmund.de/ir/reports/96/Fuhr-96a.html), University of Dortmund, Computer Science Department.

Robertson, S. (1977). The Probability Ranking Principle in IR. *Journal of Documentation 33*, pages 294–304.