A theoretical study of recall and precision using a topological approach to information retrieval

A THEORETICAL STUDY OF RECALL AND PRECISION USING A TOPOLOGICAL APPROACH TO INFORMATION RETRIEVAL

by

LEO EGGHE* and RONALD ROUSSEAU**

* LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
E-mail : legghe@luc.ac.be

** UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
and
KHBO, Zeedijk 101, B-8400 Oostende, Belgium
E-mail : Ronald.Rousseau@kh.khbo.be

Abstract - Topologies for retrieval systems are generated by certain subsets, called retrievals. In this article we show how recall and precision can be expressed using only retrievals. Different types of retrieval systems are investigated : both threshold systems and "close match" systems, and both "optimal" and "non-optimal" retrieval. The relation with the hypergeometric and some "non-standard" distributions is highlighted.

1. INTRODUCTION

The late Jean Tague-Sutcliffe noted that (Tague-Sutcliffe, 1996)

> *Most information retrieval evaluation is experimental. However, in related sciences, such as computer science, results may be obtained by analysis as well as by experiment. This approach has been little used in information retrieval evaluation, perhaps because the problems are not so well-defined as in computer science.*

Taking up Tague-Sutcliffe's implicit challenge, we will work in this article within a well-defined theoretical framework. This will enable us to obtain precise, analytical results. Moreover, following Buckland and Gey, we stress the fact that in order to obtain a clear understanding of the notions of recall and precision, it is important and useful to study the theoretical behavior of these measures (Buckland & Gey, 1994). They constitute a basic building block for the understanding of any retrieval model.

A retrieval system is a triple (DS, QS, sim) consisting of a document space DS, a query space QS and a similarity function sim. In (Everett & Cater, 1992) the authors introduced the retrieval topology, denoted as $\mathcal{T}$, generated by the subbasis of retrievals :

$$\{R(Q,r) \| r \in \mathbb{R},\ Q \in QS\}$$

where a retrieval R(Q,r), $r \in \mathbb{R}$, is defined as :

$$R(Q,r) = \{D \in DS \| sim(D,Q) > r\} \quad .$$

Note that the set of retrievals of a query Q is the set of all possible answers to the query Q in the system (DS, QS, sim) with the retrieval topology. Different answers are obtained by changing the threshold. For definitions of topological notions used in this article we refer the reader to Appendix B, our earlier articles (Egghe & Rousseau, 1998) and the mathematical literature.

Further, the topology $\mathcal{T}''$, referred to as the similarity topology, is defined as the coarsest topology on DS that makes all similarity functions sim(.,Q) continuous. It is generated by the following subbasis of retrievals :

$$\{U(Q,r_1,r_2) \| Q \in QS,\ r_1 < r_2\}$$

where

$$U(Q,r_1,r_2) = \{D \in DS \| r_1 < sim(D,Q) < r_2\} = sim(.,Q)^{-1}(]r_1,r_2[)$$

with $r_1 < r_2$, $r_1,r_2 \in \mathbb{R}$.

We will assume that the document space DS consists of the documents $\{D_1,D_2,...,D_n\}$ ordered - in increasing order - according to the similarity values of one particular query (it does not matter which). However, to simplify the analysis we assume that all these similarity values are different. Consequently, a retrieval in the retrieval topology has the form $\{D_i,D_{i+1},...,D_n\}$ (i = 1,...,n) and a retrieval in the similarity topology has the form $\{D_i,D_{i+1},...,D_m\}$ (i,m = 1,...,n). More precisely we will assume that we can retrieve no other sets than retrievals.

We will show that it is possible to express the notions recall and precision using the topological approach introduced and studied in (Cater, 1986; Everett & Cater, 1992; Egghe, 1998; Egghe & Rousseau, 1997a, 1998; Rousseau,1998).

We will investigate (both for the retrieval topology and the similarity topology) non-optimal searches and optimal searches (in that order). By optimal searches we mean that the used retrieval is the best one with respect to the requested documents. More concretely, if we want k documents from DS = $\{D_1,...,D_n\}$ from which we will find p documents as a result ($p \leq k$) and if we retrieve the set $\{D_i,...,D_n\}$ (using the retrieval topology) then D_i is one of these p documents. In the same way, if we use the similarity topology, we will retrieve $\{D_i,...,D_m\}$ and we assume that D_i and D_m are amongst the p documents. If this is not necessarily the case we call the search non-optimal. We will begin with the latter case (although the methodology is the same for both cases, it turns out that the non-optimal case yields simple formulae whereas the other case can only be monitored using approximations).

Note that $p = 0,1,...,k \in \mathbb{N}$. In this connection we will show that we are dealing with the hypergeometric distribution (see e.g. Olkin, Gleser and Derman (1980) or Rothschild and Logothetis (1986)). If p = k we have perfect recall.

It turns out that the average recall-precision values (in short:(R,P) values) in the non-optimal case and in the optimal case are close to each other. This, in turn, yields accurate average (R,P) values as they are experienced in random sampling.

The paper closes with some open problems that arise from these models. We stress again that the study of the mathematical modelling of (R,P) is very important for a basic understanding of IR.

2. NON-OPTIMAL SEARCHING

We consider a document space $DS = \{D_1,D_2,...,D_n\}$ in which the documents have an increasing similarity w.r.t. one particular query Q, i.e. the finite sequence $(sim(D_i,Q))_{i=1,...,n}$ is increasing. We study the retrieval topology $\mathcal{T}$ as well as the similarity topology $\mathcal{T}''$ as explained in the introduction.

2.1. The case of retrieval via $\mathcal{T}$

We retrieve via the retrievals (Q fixed):

$$R(Q,r) = \{E \in DS \| sim(E,Q) > r\}.$$

Suppose we want to retrieve k documents from which we find p documents.

We take $k = 1,2,...,n$ and $p = 0,...,k$. Observe that, although the case $k = 0$ occurs in practice (e.g. when a would-be inventor does a search in a patent database), we will not consider this case. We notice that for every $r \in \mathbf{R}^+$ (the positive real numbers):

$$R(Q,r) = \{D_j \| j > i\} \subset DS$$

for a certain $i = 0,...,n$ (here D_0 denotes a symbolical document to allow for the case that $R(Q,r) = DS$). We hence have a situation as in figure 1.

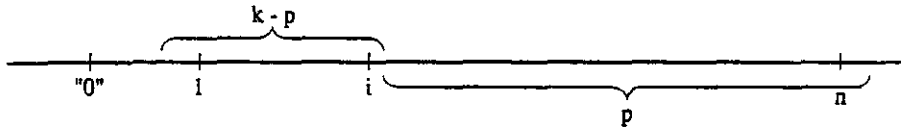

Figure 1 : Retrieval via $\mathcal{T}$, non-optimal searching: k documents are found in the set $\{D_{i+1},...,D_n\}$, k-p documents are found in the set $\{D_1,...,D_i\}$

In the retrieved set $\{D_{i+1},...,D_n\}$ there are p out of the k requested documents, leaving k-p requested documents in the non-retrieved set $\{D_1,...,D_i\}$. Obviously

$$R = \frac{p}{k} \tag{1}$$

$$P = \frac{p}{n - i} \quad , \tag{2}$$

are the recall and precision values. For fixed n,k,p and i these values occur $\binom{n-i}{p}\binom{i}{k-p}$ times. As $p = 0,1,...,k$ and $i = k-p,...,n-p$ we are dealing with a bivariate (in i and p) discrete distribution $(i = k-p,...,n-p;\ p = 0,...,k)$:

$$P(X=i, Y=p) = \frac{\binom{n-i}{p}\binom{i}{k-p}}{(k+1)\binom{n+1}{k+1}} \tag{3}$$

since it follows from Prudnikov, Brychkov and Marichev (1992), p.618(39)[1] that

$$\sum_{i=k-p}^{n-p} \binom{n-i}{p}\binom{i}{k-p} = \binom{n+1}{k+1} \quad . \tag{4}$$

Note that the binomial coefficient $\binom{0}{0} = 1$ and that $\binom{s}{t} = 0$ if $s < t$. This distribution gives the probability that among k "desired" documents in a database containing n items, there are p of the k documents in the set $\{D_{i+1},...,D_n\}$.

---

(1) Note : In Prudnikov, Brychkov and Marichev (1992) one must correct formula 39 on p.618 by replacing $n+\ell$ by $n+1$.

From this it follows that the marginal distributions are :

$$P(Y=p) = \sum_{i=k-p}^{n-p} \frac{\binom{n-i}{p}\binom{i}{k-p}}{(k+1)\binom{n+1}{k+1}} = \frac{1}{k+1} \tag{5}$$

$$P(X=i) = \sum_{p=0}^{k} \frac{\binom{n-i}{p}\binom{i}{k-p}}{(k+1)\binom{n+1}{k+1}} = \frac{1}{n+1}. \tag{6}$$

The latter result is obtained by using Vandermonde's convolution formula:

$$\sum_{p=0}^{k} \binom{n-i}{p}\binom{i}{k-p} = \binom{n}{k} \tag{7}$$

(see Gradshteyn and Ryzhik (1965), p.4, 0.156 (1) or Prudnikov, Brychkov and Marichev (1992), p.616 (13)).

From this, the conditional distributions follow (i = k-p,...,n-p; p = 0,...,k).

For p fixed :

$$P(X=i|Y=p) = \frac{P(X=i, Y=p)}{P(Y=p)}$$

$$= \frac{\binom{n-i}{p}\binom{i}{k-p}}{\binom{n+1}{k+1}}. \tag{8}$$

For i fixed :

$$P(Y=p|X=i) = \frac{P(X=i, Y=p)}{P(X=i)}$$

$$= \frac{\binom{n-i}{p}\binom{i}{k-p}}{\binom{n}{k}} \tag{9}$$

Hence for every fixed i = k-p,...,n-p, P(Y=p|X=i) is the classical hypergeometric distribution. See Olkin, Gleser and Derman (1980) and Rothschild and Logothetis (1986) for more information on this important distribution. There one can find that

$$\mu_i = k\,\frac{n-i}{n} \tag{10}$$

where μ_i denotes the average (over p: i is fixed) of the hypergeometric distribution (9). From (10) we obtain the following formulae for the average precision and recall values that are encountered in this system :

$$\overline{P}_i = \sum_{p=0}^{k} \frac{p}{n-i}\,P(Y=p|X=i) = \frac{k}{n} \tag{11}$$

$$\overline{R}_i = \sum_{p=0}^{k} \frac{p}{k}\,P(Y=p|X=i) = \frac{n-i}{n}. \tag{12}$$

Since P(X=i|Y=p) is not a "standard" distribution, it takes more work to calculate the average (over i) precision value $\overline{P}_p$. For recall, however, it is clear that

$$\overline{R}_p = R = \frac{p}{k} \tag{13}$$

since R is independent from i. For P, we have : $\overline{P}_0 = P_0 = 0$ and, for $p \neq 0$:

$$\overline{P}_p = \sum_{i=k-p}^{n-p} \frac{p}{n-i}\,P(X=i|Y=p)$$

$$= \frac{1}{\binom{n+1}{k+1}} \sum_{j=p}^{n-k+p} \frac{p}{j}\binom{j}{p}\binom{n-j}{k-p}$$

$$= \frac{1}{\binom{n+1}{k+1}} \sum_{j=p}^{n-k+p} \binom{j-1}{p-1} \binom{n-j}{k-p} , \qquad (14)$$

by using again formula (4) but with other symbols. Finally we can also calculate $\overline{P}$ and $\overline{R}$ over the bivariate distribution (3).

$$\overline{P} = \sum_{p=0}^{k} \sum_{i=k-p}^{n-p} \frac{p}{n-i} P(X=i, Y=p)$$

$$= \sum_{p=0}^{k} \frac{1}{k+1} \sum_{i=k-p}^{n-p} \frac{p}{n-i} \frac{\binom{n-i}{p}\binom{i}{k-p}}{\binom{n+1}{k+1}}$$

$$= \sum_{p=1}^{k} \frac{1}{n+1} = \frac{k}{n+1} \qquad (15)$$

$$\overline{R} = \sum_{p=0}^{k} \sum_{i=k-p}^{n-p} \frac{p}{k} P(X=i, Y=p)$$

$$= \frac{1}{k} \sum_{p=1}^{k} p \left[\sum_{i=k-p}^{n-p} \frac{\binom{n-i}{p}\binom{i}{k-p}}{(k+1)\binom{n+1}{k+1}} \right]$$

$$= \frac{1}{k} \sum_{p=1}^{k} p \frac{1}{k+1} \quad \text{(by (4))}$$

$$\overline{R} = \frac{1}{2} . \qquad (16)$$

This leads to the following theorem:

Theorem 2.1.1

In the case of retrieval via the retrieval topology $\mathcal{T}$, the probability to retrieve p documents from k relevant ones via the retrieval $\{D_{i+1},...,D_n\}$ is

$$P(X=i,Y=p) = \frac{\binom{n-i}{p}\binom{i}{k-p}}{(k+1)\binom{n+1}{k+1}} \tag{3}$$

(i=k-p,...,n-p;p=0,...,k). This gives rise to the conditional distributions

$$P(X=i|Y=p) = \frac{\binom{n-i}{p}\binom{i}{k-p}}{\binom{n+1}{k+1}} \tag{8}$$

and

$$P(Y=p|X=i) = \frac{\binom{n-i}{p}\binom{i}{k-p}}{\binom{n}{k}} . \tag{9}$$

The latter is the hypergeometric distribution. For (9) the average recall and precision values are:

$$\overline{R}_i = \frac{n-i}{n} \tag{12}$$

$$\overline{P}_i = \frac{k}{n} . \tag{11}$$

For (8) these are

$$\overline{R}_p = R = \frac{p}{k} \tag{13}$$

$$\overline{P}_p = \frac{k+1}{n+1} \quad (p \neq 0, \overline{P}_0 = P_0 = 0) \tag{14}$$

and calculated over (3) these are

$$\overline{R} = \frac{1}{2} \tag{16}$$

$$\overline{P} = \frac{k}{n+1} . \tag{15}$$

Discussion

1. We have found a bivariate probability distribution (3). The marginal distribution for fixed i is the hypergeometric distribution (in p). This was already remarked by Shaw, Burgin and Howell (1997). Their paper became available at the writing of the present paper so that the findings are independent of each other, but Shaw, Burgin and Howell deserve the credit for being the first to remark this. Surprisingly, their formula is somewhat different from ours; one formula can be recovered from the other one by interchanging the terms "retrieved" and "relevant". In any case it can be readily checked that the probabilities appearing in their formula are exactly the same as the one in (3). These findings in the context of duality have been studied in a separate note (Egghe and Rousseau (1997b)).

2. Formulae (11)-(14) are the most interesting results. It follows that

$$\overline{P}_i = \frac{k}{i} (1 - \overline{R}_i) \quad , \tag{17}$$

leading to a decreasing linear relationship between recall and precision. Also, for $p \neq 0$,

$$\overline{P}_p = \frac{1}{n+1} \left(1 + \frac{p}{R} \right) \tag{18}$$

($R = \overline{R}_p$), a hyperbolically decreasing relationship between recall and precision. Formula (18) is illustrated in figure 2 for $n = 10$ ($p \leq k$, $p,k=1,...,10$). We note that, although these recall-precision curves are decreasing, they are not concave as required by Egghe's model (Egghe, 1992), nor have they the form of tangent parabolic recall (Buckland & Gey, 1994). The reason for this difference is that there is no element of time or causality in our theoretic model : we do not require or even expect that 'first' relevant documents are found and 'later' the other ones. In fact we deal here with random sampling. Random sampling is unbiased as opposed to any result of an IR action. Random sampling is the basis of IR from which IR-results can be studied. Let us give an example of this. Random sampling is used (in DS) e.g. to determine the number of relevant documents to a certain query. Confidence intervals of this can be built based on the knowledge that the hypergeometric distribution is approximated by the normal (Gaussian) distribution.

Note also the remarkable fact that no $\overline{P}$, $\overline{P}_i$, $\overline{P}_p$ value depends on i and p! The case p=k corresponds to perfect retrieval (R=1).

Further we note that all $\bar{P}$, $\bar{P}_i$, $\bar{P}_p$ values are decreasing in n (if k stays fixed) and decreasing in n (for a fixed (k/n)-value). In the former case the limit is 0; in the latter case it is equal to k/n, as is readily seen. The latter case (for $\bar{P}_p$) is shown in figure 3.

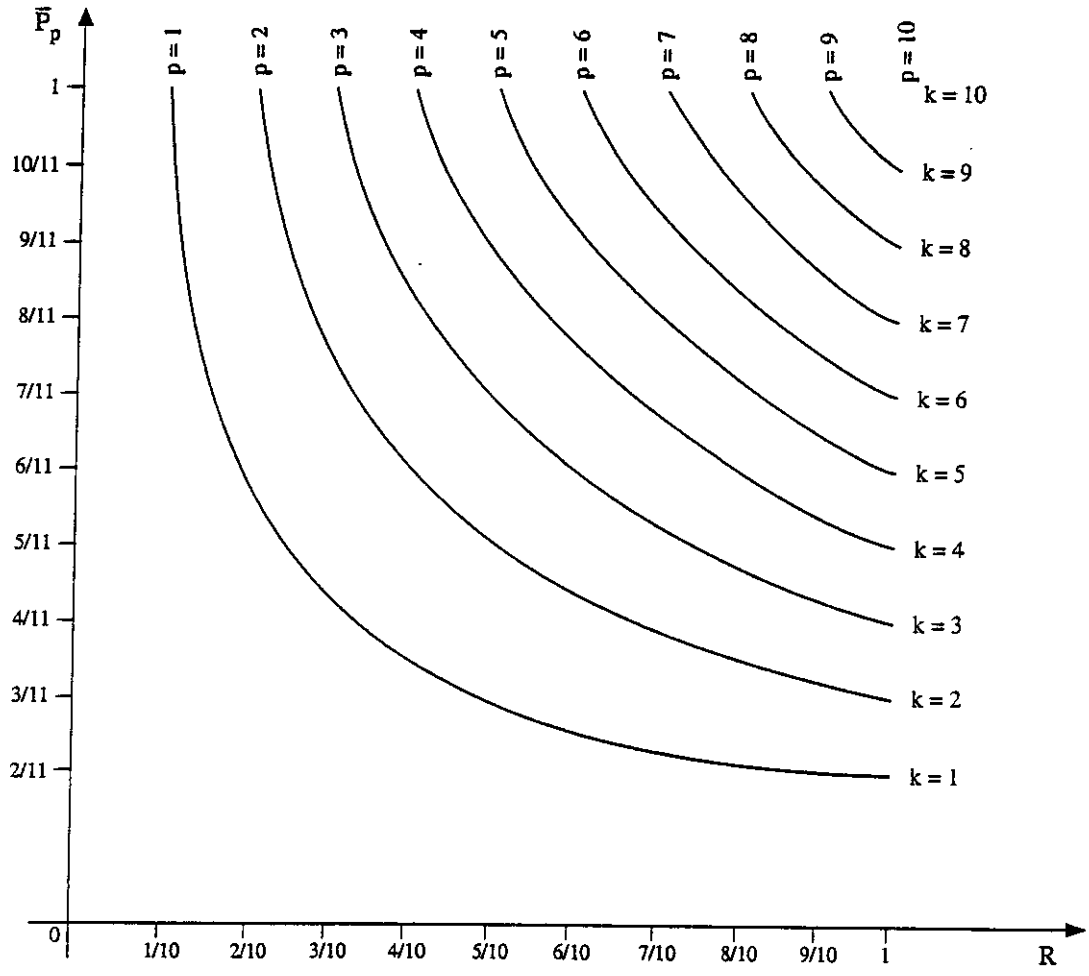

Figure 2

Recall-precision graph for non-optimal searching using the retrieval topology for a document space containing n = 10 documents

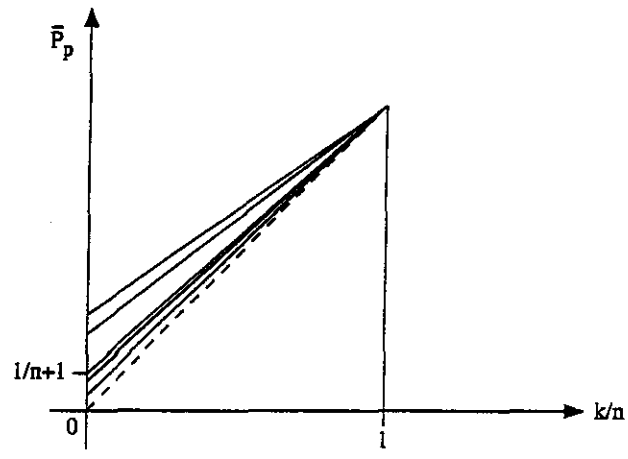

Figure 3 : $\bar{P}_p$ versus k/n

Note concerning the mode of P

Another way to study average behavior is to look at the mode or the median. (Information on the median will be given later.) Here we will show that, for the case of non-optimal searching, using the retrieval topology, i.e. formula (3) or (8), for fixed p, a mode is attained for

$$i_m = \left\lfloor \frac{(k-p)n}{k} - \frac{p}{k} + 1 \right\rfloor .$$

The symbol $\lfloor x \rfloor$ denotes the floor function, i.e. the largest integer smaller than or equal to x. If (k-p)n/k happens to be an integer then $i_m = (k-p)n/k$. This means that the precision value corresponding to a mode is $p/(n-i_m)$, by (2), and is approximately equal to k/n, i.e. is (almost) independent of p.

Theorem

$$\max_i \binom{n-i}{p}\binom{i}{k-p}$$

is attained for

$$i_m = \left\lfloor \frac{(k-p)n}{k} - \frac{p}{k} + 1 \right\rfloor .$$

If

$$\frac{(k-p)n}{k} \in \mathbb{N}$$

then

$$i_m = \frac{(k-p)n}{k} .$$

In this case,

$$\mathrm{Mod}(P) = \frac{k}{n} .$$

The proof is provided in Appendix C. Fig.4 illustrated the occurrence of the mode.

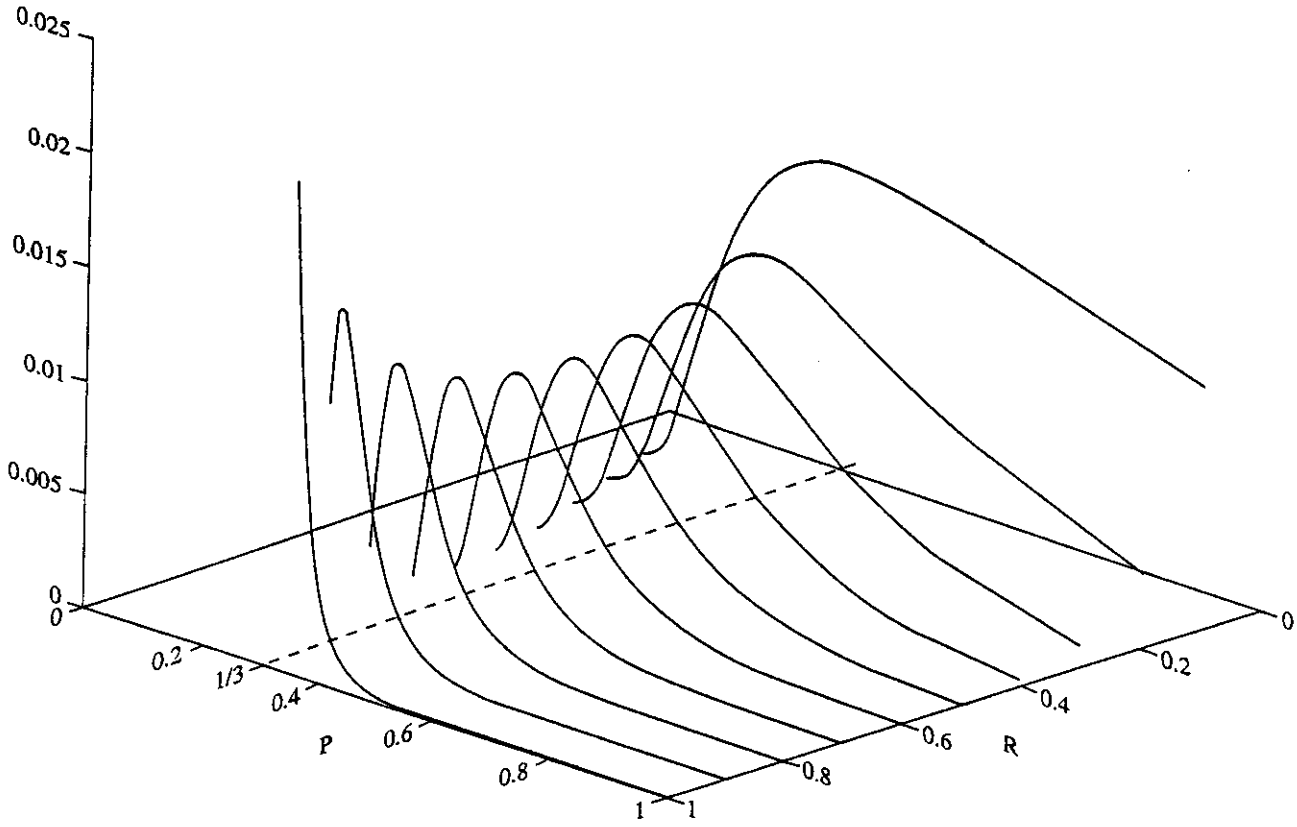

Figure 4

Bivariate distribution, with recall and precision values in the plane, illustrating that for different recall values, the mode occurs at - approximately - the same precision value (here n = 30, k = 10, hence k/n = 1/3)

Figure 4 shows ten curves for the case of non-optimal searching, using the retrieval topology, n = 30, k = 10. Each curve corresponds to a fixed recall value, equal to p/10, p = 1,...,10. The other axis corresponds to precision values p/(10-i); with p fixed, i takes values between 10-p and 30-p. The number of times each (R,P) value occurs (formula (3)) corresponds with the height of the curves. Figure 4 clearly shows that the modes of these curves all occur at the same P-value, namely for P = k/n = 1/3.

2.2. The case of retrieval via $\mathcal{T}''$

We now use the retrievals

$$U(Q, r_1, r_2) = \{E \in DS \| r_1 < sim(E,Q) < r_2\} \ .$$

This set looks like $\{D_i \| i < j < m\} \subset DS$, where $i = 0,...,n$ and $m = 1,...,n+1$ (as in 2.1 D_0 and D_{n+1} denote fictitious documents to allow for the case that $U(Q,r_1,r_2) = DS$). We hence have a situation as in figure 5.

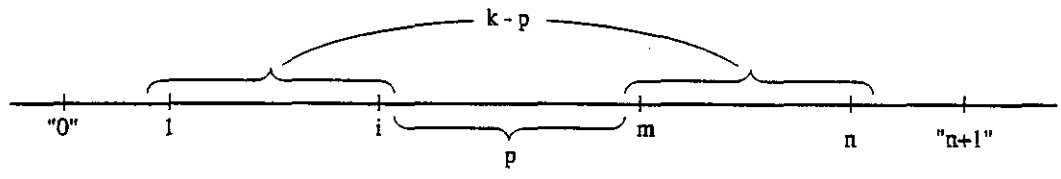

Figure 5 : Retrieval via $\mathcal{T}''$, non-optimal searching

In the retrieved set $\{D_{i+1},...,D_{m-1}\}$ there are p out of the k requested documents leaving k-p requested documents in the non-retrieved set $\{1,...,i\} \cup \{m,...,n\}$. Obviously

$$R = \frac{p}{k} \tag{19}$$

$$P = \frac{p}{m - i - 1} \ . \tag{20}$$

These values occur $\binom{m-i-1}{p}\binom{n-m+i+1}{k-p}$ times : p documents in $\{D_{i+1},...,D_{m-1}\}$ and k-p

in the set $\{D_1,...,D_i\} \cup \{D_m,...,D_n\}$.

We have now a trivariate (in i, m and p) discrete distribution :

$$P(X=i, Y=m, Z=p) = \frac{\binom{m-i-1}{p}\binom{n-m+i+1}{k-p}}{(k+1)\binom{\frac{k}{2}+1}{}\binom{n+2}{k+2}} \tag{21}$$

$(m = i+p+1,...,n+1; i = 0,...,n-p; p = 0,...,k)$.

Here we use the formula :

$$\sum_{i=0}^{n-p} \sum_{m=i+p+1}^{n+1} \binom{m-i-1}{p} \binom{n-m+i+1}{k-p} = (k-p+1) \binom{n+2}{k+2} . \qquad (22)$$

The proof is given in Appendix D since we were not able to trace this formula in the literature. Summation of (22) over p = 0,...,k yields the denominator of (21).

We will consider the marginal distributions of p and of i and m (together, by the very nature of close match retrieval via $\mathcal{T}''$).

$$P(Z=p) = \sum_{i=0}^{n-p} \sum_{m=i+p+1}^{n+1} \frac{\binom{m-i-1}{p} \binom{n-m+i+1}{k-p}}{(k+1)\left(\frac{k}{2}+1\right)\binom{n+2}{k+2}} = \frac{k-p+1}{(k+1)\left(\frac{k}{2}+1\right)} \qquad (23)$$

$$P(X=i, Y=m) = \sum_{p=0}^{k} \frac{\binom{m-i-1}{p} \binom{n-m+i+1}{k-p}}{(k+1)\left(\frac{k}{2}+1\right)\binom{n+2}{k+2}}$$

$$= \frac{k+2}{(n+1)(n+2)\left(\frac{k}{2}+1\right)} \qquad (24)$$

again using (7) but with n-i replaced by m+i+1.

From this, the conditional distributions follow (m = i+p+1,...,n+1; i = 0,...,n-p; p = 0,...,k):

$$P(X=i, Y=m \mid Z=p) = \frac{P(X=i, Y=m, Z=p)}{P(Z=p)} = \frac{\binom{m-i-1}{p} \binom{n-m+i+1}{k-p}}{(k-p+1)\binom{n+2}{k+2}} \qquad (25)$$

$$P(Z=p|X=i,Y=m) = \frac{P(X=i,Y=m,Z=p)}{P(X=i,Y=m)} = \frac{\binom{m-i-1}{p}\binom{n-m+i+1}{k-p}}{\binom{n}{k}} \qquad (26)$$

as is readily seen. It follows that, for every fixed m = i+p+1,...,n+1 and i = 0,...,n-p, P(Z=p|X=i,Y=m) is the classical hypergeometric distribution.

Expression (10) now becomes :

$$\mu_{i,m} = k\,\frac{m-i-1}{n}\ , \qquad (27)$$

where $\mu_{i,m}$ denotes the average (over p) of the hypergeometric distribution (26). This yields formulae for the average precision and recall values that are encountered in this system :

$$\overline{P}_{i,m} = \sum_{p=0}^{k} \frac{p}{m-i-1} P(Z=p|X=i,Y=m) = \frac{k}{n} \qquad (28)$$

$$\overline{R}_{i,m} = \sum_{p=0}^{k} \frac{p}{k} P(Z=p|X=i,Y=m) = \frac{m-i-1}{n}. \qquad (29)$$

What about the R and P-averages with respect to P(X=i,Y=m|Z=p)?

Clearly

$$\overline{R}_p = R = \frac{p}{k} \qquad (30)$$

since R is independent from i and m. For P we have

$$\overline{P}_p = \sum_{i=0}^{n-p} \sum_{m=i+p+1}^{n+1} \frac{p}{m-i-1} \frac{\binom{m-i-1}{p}\binom{n-m+i+1}{k-p}}{(k-p+1)\binom{n+2}{k+2}}$$

$$= \sum_{i=0}^{n-p} \sum_{\ell=p}^{n-i} \frac{p}{\ell} \frac{\binom{\ell}{p}\binom{n-\ell}{k-p}}{(k-p+1)\binom{n+2}{k+2}} \ .$$

It turns out that

$$\sum_{i=0}^{n-p}\sum_{\ell=p}^{n-i}\frac{1}{\ell}\binom{\ell}{p}\binom{n-\ell}{k-p}=\frac{1}{p}\sum_{i=0}^{n-p}\sum_{\ell=p}^{n-i}\binom{\ell-1}{p-1}\binom{n-\ell}{k-p}$$

$$=\frac{1}{p}\sum_{i'=0}^{n'-p'}\sum_{\ell'=p'}^{n'-i'}\binom{\ell'}{p'}\binom{n'-\ell'}{k'-p'}$$

if $p \neq 1$, with $\ell' = \ell-1$, $p' = p-1$, $k' = k-1$, $n' = n-1$, $i' = I$. Note that $\bar{P}_0 = P_0 = 0$.
We apply the formula proved in Appendix D, yielding

$$\bar{P}_p = \frac{(k'-p'+1)\binom{n'+2}{k'+2}}{(k-p+1)\binom{n+2}{k+2}}$$

$$\bar{P}_p = \frac{k+2}{n+2}\,. \tag{31}$$

Finally we now calculate the "overall" averages $\bar{P}$ and $\bar{R}$ for the trivariate distribution

$$\bar{P} = \sum_{p=0}^{k}\sum_{i=0}^{n-p}\sum_{m=i+p+1}^{n+1}\frac{p}{m-i-1}\,P(X=i, Y=m, Z=p)$$

$$=\sum_{p=1}^{k}\frac{k-p+1}{(k+1)\left(\frac{k}{2}+1\right)}\,\frac{k+2}{n+2}=\frac{k}{n+2} \tag{32}$$

and

$$\bar{R} = \sum_{p=0}^{k}\sum_{i=0}^{n-p}\sum_{m=i+p+1}^{n+1}\frac{p}{k}\,P(X=i, Y=m, Z=p)$$

$$=\sum_{p=0}^{k}\frac{p}{k}\,P(Z=p)=\frac{1}{3}, \tag{33}$$

using that

$$\sum_{p=0}^{k} p = \frac{k(k+1)}{2} \quad \text{and} \quad \sum_{p=0}^{k} p^2 = \frac{k(k+1)(2k+1)}{6} .$$

Concluding, we have proved the following theorem :

Theorem 2.2.1

In the case of retrieval via the similarity topology $\mathcal{T}$", the probability to retrieve p documents from k relevant ones via the retrieval $\{D_{i+1},..,D_{m-1}\}$ is

$$P(X=i, Y=m, Z=p) = \frac{\binom{m-i-1}{p}\binom{n-m+i+1}{k-p}}{(k+1)\binom{\frac{k}{2}+1}{}\binom{n+2}{k+2}} \tag{21}$$

$(m = i+p+1,...,n+1; i = 0,...,n-p; p = 0,...,k)$.

This gives rise to the conditional distributions:

$$P(X=i, Y=m \mid Z=p) = \frac{\binom{m-i-1}{p}\binom{n-m+i+1}{k-p}}{(k-p+1)\binom{n+2}{k+2}} \tag{25}$$

and

$$P(Z=p \mid X=i, Y=m) = \frac{\binom{m-i-1}{p}\binom{n-m+i+1}{k-p}}{\binom{n}{k}} \tag{26}$$

The latter is the hypergeometric distribution. Over (26) the average recall and precision values are

$$\overline{R}_{i,m} = \frac{m-i-1}{n} \tag{29}$$

$$\overline{P}_{i,m} = \frac{k}{n} . \tag{28}$$

Over (25) these are :

$$\bar{R}_p = R = \frac{p}{k} \tag{30}$$

$$\bar{P}_p = \frac{k+2}{n+2} \tag{31}$$

and over (21) these are

$$\bar{R} = \frac{1}{3} \tag{33}$$

$$\bar{P} = \frac{k}{n+2} \quad .$$

Discussion

It follows that

$$\bar{P}_{i,m} = \frac{k}{i}\left(\frac{m-1}{n} - \bar{R}_{i,m}\right), \tag{34}$$

which indicates a decreasing linear relationship between recall and precision. Also

$$\bar{P}_p = \frac{1}{n+2}\left(2 + \frac{p}{R}\right), \tag{35}$$

a hyperbolically decreasing relationship between recall and precision.

The curves are similar to the $\mathcal{T}$-case (figure 2). Note again that no $\bar{P}$, $\bar{P}_{i,m}$, $\bar{P}_p$ depends on i, m and p! All these values are decreasing in n (k fixed) and decreasing in n (k/n fixed). In the former case the limit is 0 and in the latter case it is equal to k/n. A similar figure as figure 5 can be drawn.

Note

The case $\mathcal{T}$ (section 2.1) follows from the case $\mathcal{T}''$ (section 2.2) by fixing m = n+1. However, in 2.2 we did not fix m because we dealt with $\mathcal{T}''$. Hence the average formulae for $\bar{P}_p$ and $\bar{P}$ in section 2.1 do not follow from those in section 2.2!

3. OPTIMAL SEARCHING

In optimal searching we assume that the first retrieved document (if we use $\mathcal{T}$) and the first and the last retrieved document (if we use $\mathcal{T}''$) belong to the k relevant ones. We thus assume that the retrieval engine is capable of making this type of search. We think that in a theoretical investigation such as the one performed here, such an assumption is allowed. However, as this case has less practical value we refer the calculations to Appendix A.

4. SUMMARY

In this article we have studied recall-precision values for random retrieval, important e.g. in statistical investigations. Results, however, are described using the underlying topological structure, i.e. using retrievals in the retrieval and the similarity topology. We have studied general, non-optimal, searches and more focused, optimal searches. Different distributions have been obtained, among which the hypergeometric one.

Our results yield the basic structure of random topological retrieval and consequently, the resulting recall-precision values constitute a lower level performance standard, cf. (Shaw, Burgin & Howell, 1997). Real IR results should be obtained by combining (convolving?) the distributions obtained in this article with other distributions, such as perhaps a "relevance" distribution, or a preference structure, or a distribution describing the precise behavior of the search mechanism. Moreover, it is clear that feedback should play a decisive role in real-world retrieval processes.

Acknowledgements

We thank several anonymous referees whose comments may not have led to an article that is easy to read, but that at least is better structured than the original. We also thank M. Dekeyser (KHBO, Oostende) for help during the rewriting process.

REFERENCES

Buckland, M. & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science, 45*, 12-19.

Cater, S.C. (1986). *The topological information retrieval system and the topological paradigm: a unification of the major models of information retrieval.* PhD. Dissertation, Louisiana State University, Baton Rouge, LA.

Egghe, L. (1992). Qualitative analysis of the recall-precision relationship in information retrieval. *Informetrics-91* (R. Rao, ed.). Sarada Ranganathan Endowment, Bangalore; 148-174.

Egghe, L. (1998). Properties of topologies of information retrieval systems. *Mathematical and Computer Modelling* (in press).

Egghe, L. & Rousseau, R. (1997a). Everett and Cater's retrieval topology (letter to the editor). *Journal of the American Society for Information Science, 48(5),* 479-480.

Egghe, L. & Rousseau, R. (1997b). Duality in information retrieval and the hypergeometric distribution. *Journal of Documentation* (in press).

Egghe, L. and Rousseau, R. (1998). Topological aspects of information retrieval. *Journal of the American Society for Information Science* (in press).

Everett, D.M. & Cater, S.C. (1992). Topology of document retrieval systems. *Journal of the American Society for Information Science, 43,* 658-673.

Gradshteyn, I.S., Ryzhik, I.M. & Jeffrey, A. (1980). *Table of integrals, series and products.* New York : Academic Press.

Olkin, I., Gleser, L.J. & Derman, C. (1980). *Probability. Models and applications.* New York: MacMillan.

Prudnikov, A.P., Brychkov, Yu.A. & Marichev, O.I. (1992). *Integrals and series. Vol.1 : Elementary functions.* New York: Gordon & Breach.

Rothschild, V. & Logothetis, N. (1986). *Probability distributions.* New York: Wiley.

Rousseau, R. (1998). Jaccard similarity leads to the Marczewski-Steinhaus topology for information retrieval. *Information Processing and Management, 34(1),* (in press).

Shaw, W.M. Jr., Burgin, R. & Howell, P. (1997). Performance standards and evaluations in IR test collections : vector-space and other retrieval models. *Information Processing & Management, 33(1),* 15-36.

Tague-Sutcliffe, J.M. (1996). Some perspectives on the evaluation of information retrieval systems. *Journal of the American Society for Information Science, 47(1),* 1-3.

APPENDIX A: Optimal searching

Part 1. The case of retrieval via $\mathcal{T}$"

The general case $p \leq k$

We still work with $DS = \{D_1,...,D_n\}$ and retrieve $\{D_i,...,D_n\}$ $(i = 1,...,n)$.
Suppose k documents are relevant and we retrieve p of them $(0 < p \leq k \leq n)$.Now, D_i is one of these p documents. We thus have a situation as depicted in figure 6.

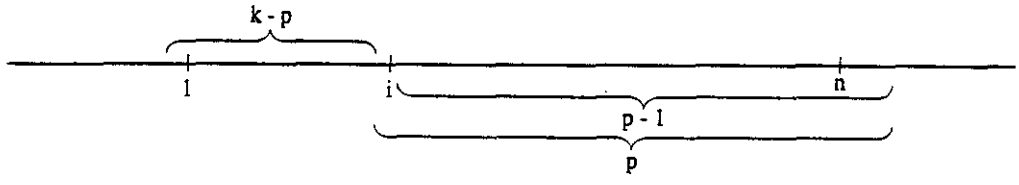

Figure 6 : Rettrieval via $\mathcal{T}$, optimal searching

The recall value is still $R = p/k$ while the precision is

$$P = \frac{p}{n - i + 1} . \tag{36}$$

These values occur $\binom{n-i}{p-1}\binom{i-1}{k-p}$ times $(i = k-p+1,...,n-p+1)$. The fact that n-i+1 is appearing in the denominator of P and that n-i occurs in the frequency of occurrence, gives rise to serious calculating difficulties : we will explain this in the sequel. ~~Let us proceed as in the previous section~~.
We have :

$$\sum_{i=k-p+1}^{n-p+1} \binom{n-i}{p-1}\binom{i-1}{k-p} = \binom{n}{k},$$

using formula (4) but with other symbols. Since this is p-independent and since p ranges in {1,...,k} (we exclude retrieving an empty set when p= 0; p ≥ 1 since D_i is relevant) we have now that optimal searching via T is governed via the bivariate distribution

$$P(X'=i,Y'=p) = \frac{\binom{n-i}{p-1}\binom{i-1}{k-p}}{k\binom{n}{k}} . \tag{37}$$

The marginal distributions are

$$P(Y'=p) = \sum_{i=k-p+1}^{n-p+1} \frac{\binom{n-i}{p-1}\binom{i-1}{k-p}}{k\binom{n}{k}} = \frac{1}{k} \tag{38}$$

$$P(X'=i) = \sum_{p=1}^{k} \frac{\binom{n-i}{p-1}\binom{i-1}{k-p}}{k\binom{n}{k}} = \frac{1}{n} \tag{39}$$

using (7) but with other symbols. Hence the conditional distributions are

$$P(X'=i|Y'=p) = \frac{P(X'=i,Y'=p)}{P(Y'=p)}$$

$$= \frac{\binom{n-i}{p-1}\binom{i-1}{k-p}}{\binom{n}{k}} \tag{40}$$

$$P(Y'=p|X'=i) = \frac{P(X'=i,Y'=p)}{P(X'=i)}$$

$$= \frac{\binom{n-i}{p-1}\binom{i-1}{k-p}}{\binom{n-1}{k-1}} . \tag{41}$$

Again as in the previous section we have here that (41) represents a hypergeometric distribution with parameters k-1, p-1 and with n replaced by n-1 and i by i-1 (i-1 = k-p,...,n-p). The average is now

$$\mu_i' = (k-1)\ \frac{n-i}{n-1}\ . \tag{42}$$

Hence, average R and P values can be calculated :

$$\overline{P}_i = \sum_{p=1}^{k} \frac{p}{n-i+1}\ P(Y'=p|X'=i)$$

$$= \sum_{p-1=0}^{k-1} \frac{p-1}{n-i+1}\ P(Y'=p|X'=i) + \frac{1}{n-i+1}$$

$$= \frac{k-1}{n-1}\ \frac{n-i}{n-i+1} + \frac{1}{n-i+1}$$

$$\overline{P}_i = \frac{k(n-i) + i - 1}{(n-1)\ (n-i+1)} \tag{43}$$

$$\overline{R}_i = \sum_{p=1}^{k} \frac{p}{k}\ P(Y'=p|X'=i)$$

$$= \sum_{p-1=0}^{k-1} \frac{p-1}{k}\ P(Y'=p|X'=i) + \frac{1}{k}$$

$$= \frac{k(n-i) + i - 1}{k(n-1)}\ . \tag{44}$$

Of course, according to the other conditional distribution we have that

$$\overline{R}_p = R = \frac{p}{k}\ . \tag{45}$$

For $\overline{P}_p$, we have

$$\overline{P}_p = \sum_{i=k-p+1}^{n-p+1} \frac{p}{n-i+1} P(X'=i \mid Y'=p) \tag{46}$$

The reader might so far find this appendix very similar to section two. However, as far as we know, formula (46) does not allow for analytical reductions. This is due to the occurrence of n-i+1 in the denominator of P and of n-i in one of the combinations. This slight difference is the basis of the difficulties. Of course (46) equals

$$\sum_{i'=k'-p'}^{n'-p'} \frac{p'}{n'-i'+1} \frac{\binom{n'-i'}{p'}\binom{i'}{k'-p'}}{\binom{n'+1}{k'+1}} + \frac{1}{n'-i'+1}$$

with $n' = n-1$, $k' = k-1$, $p' = p-1$, $i' = i-1$. One could then approximate this by deleting $1/(n'-i'+1)$ and by assuming that

$$\frac{1}{n'-i'+1} \approx \frac{1}{n'-i'}$$

in the summation over i'. Then applying formula (14) yields

$$\overline{P}_p \approx \frac{k}{n} . \tag{47}$$

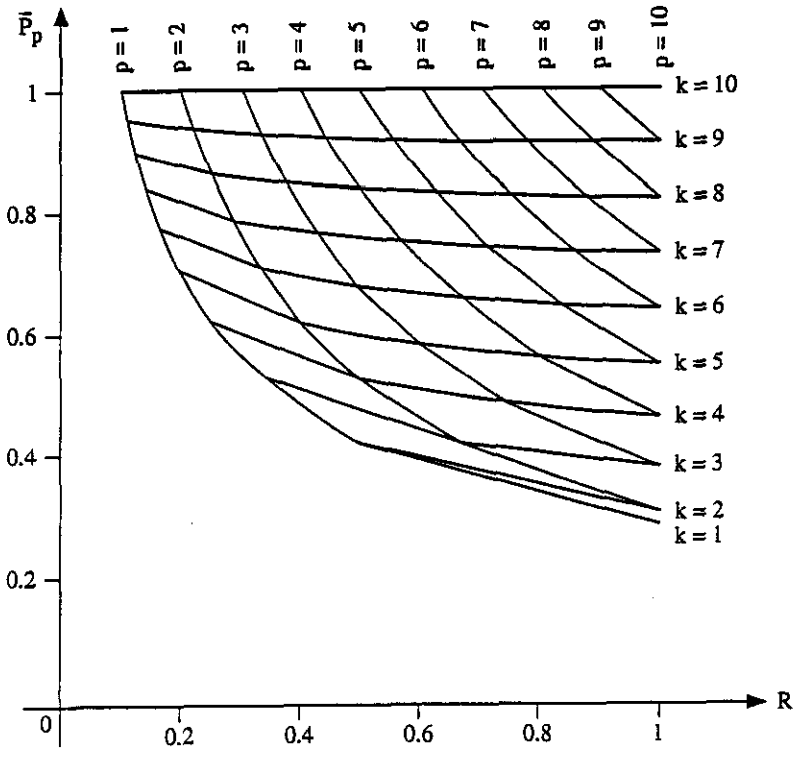

Figure 7

Recall-precision graph for optimal searching, using the retrieval topology, for a document space containing n = 10 documents

Even the case p = k (R = 1) will show that this approximation is rather rough. Still it is hardly possible to calculate the sum in (46) directly since i runs from k-p+1 to n-p+1. Now $p \leq k << n$ since k denotes the number of relevant (wanted) documents and n is the number of documents in the database. Hence n can easily be in the order of 10^5 or even 10^7 while k usually is of the order of 10 or 10^2. To get an idea what (46) looks like, figure 7 has been constructed for n = 10, which is the analogue of figure 2, but now for the case of optimal searching.

It is obvious that figure 7 ressembles figure 2 very much. The horizontal k = constant lines of figure 2 are slightly decreasing now. The p = constant lines are comparable with the hyperbolae of figure 2.

The remainder of this section is devoted to the case k = p where we will find more information on the intricate formula (46) and where approximations, better than (47) are given.

The case p = k (i.e. R = 1)

In this case, formula (46) reduces to

$$\overline{P}_p = \sum_{i=1}^{n-k+1} \frac{k}{n-i+1} \frac{\binom{n-i}{k-1}}{\binom{n}{k}}$$

$$= \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j} \,. \qquad (48)$$

This function is depicted in figure 8 for n = 5, 7, 10, 15 and 20, where $\overline{P}_p$ *is shown in function of* k/n.

Although a workable exact analytical expression of (48) is unknown to us we are able to explain figure 8 to a large extent. Indeed, we have the following theorem.

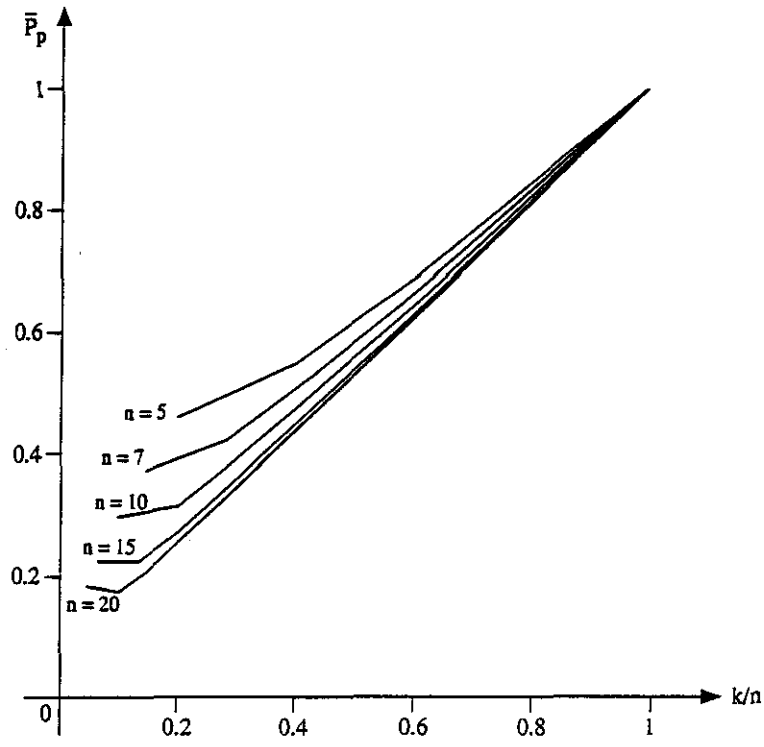

Figure 8

Average precision, using the retrieval topology and optimal searching,
for document spaces containing n = 5, 7, 10, 15 and 20 documents.
The x axis shows k/n values, the y axis shows the corresponding average precision

Theorem A.1

Consider

$$\overline{P}_p(n) = \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j} \tag{48}$$

as a function of n. Then we have

(i) $\overline{P}_p(n)$ decreases in n (fixed k)

(ii) $\lim_{\substack{n\to\infty \\ k=cst}} \overline{P}_p(n) = 0$

(iii) $\lim_{\substack{n\to\infty \\ k/n=cst}} \overline{P}_p(n) = \frac{k}{n}$ (here we look at vertical k/n = constant trajectories).

The proof is given in Appendix E.

The graph in figure 8 should also be compared to the non-optimal search case. There we found figure 3, which shows a similar behavior except for the initial decrease at the lower values of k/n. This part of the curve, however, is the most important one in practice since, in retrieval, $k << n$. Again we can say that, for the lower values of k/n, better approximations (better than k/n) are required in order to have workable and accurate approximations. This is done in the rest of this subsection.

We will approximate

$$\overline{P}_p(n) = \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j} \tag{48}$$

as follows.

Theorem A.2

$\forall n \in \mathbb{N}, \forall k \in \mathbb{N}, k \leq n$:

$$\overline{P}_p = \sum_{\ell=1}^{k-1} \psi_\ell + \varphi \ , \tag{49}$$

where

$$\psi_\ell = \frac{(-1)^{\ell-1} \ (\ell-1)! \ k^2}{(k-\ell) \ n(n-1)\ldots(n-\ell+1)} \tag{50}$$

$$\varphi = (-1)^{k-1} \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \frac{1}{j} \ . \tag{51}$$

If we denote

$$\varphi_i = \sum_{\ell=1}^{i} \psi_\ell$$

then, $\forall i = 1,\ldots,k-1$

$$\lim_{n\to\infty} |\bar{P}_p - \varphi_i| n^i = 0 \qquad (52)$$

if k is fixed.

The proof is given in Appendix F. Note that (49) is exact and only requires the calculation of k terms. In fact, we have reduced the summation to n in (48) to a summation of k terms and the calculation of $\sum_{j=k}^{n} \frac{1}{j}$ which is well-known (accurate tables exist and moreover: $\sum_{j=k}^{n} \frac{1}{j} \approx \ell n\ n + \gamma - \sum_{j=1}^{k-1} \frac{1}{j}$, for n large). Here γ denotes Euler's number. For a good approximation of $\bar{P}_p$ even less than k terms are required (by (52)). We go into this in more detail.

Suppose we did not wish to use the exact formula (49) but to approximate $\bar{P}_p$ by using just a few terms. Could this be done? Formula (52) shows that this must be possible at least for large n. To see what happens we calculated $\varphi_1,\varphi_2,\ldots,\varphi_7$ for n = 10 (not at all large but enough to see what happens) and k = 2,...,10. Note that

$$\psi_1 = \frac{k^2}{n(k-1)}, \quad \psi_2 = -\frac{k^2}{(k-2)\ n(n-1)}$$

and so on. We have the following table.

Table 1

Comparison between $\bar{P}_p$ and consecutive approximations

Case n = 10, k = 2,3,...,10

k	$\bar{P}_p$	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7
2	0.31427	0.4	-	-	-	-	-	-
3	0.38572	0.45	0.35	-	-	-	-	-
4	0.46802	0.5333	0.44444	0.48889	-	-	-	-
5	0.55415	0.625	0.53241	0.56713	0.53737	-	-	-
6	0.64203	0.72	0.62	0.65333	0.63190	0.66048	-	-
7	0.73086	0.81667	0.70778	0.74181	0.72236	0.75153	0.71264	-
8	0.82025	0.91429	0.79577	0.83133	0.81227	0.82921	0.80381	0.88000
9	0.91	1.0125	0.88393	0.92143	0.90214	0.91821	0.89679	0.94500
10	1	1.11111	0.97222	1.01191	0.99206	1.00794	0.98809	1.02778

One can see that the fit to the perfect $\bar{P}_p$-value is good and is reached quickly. It is now clear that the rough estimate k/n as mentioned in formula (47) is not particularly accurate but gets better the closer we are to k/n = 1, a fact that also follows from the graphs in figure 8. In table 1, the underlined values are the best approximations. This is surprising : at first it seems that the best values are on the diagonal but this stops at φ_4 : all φ_4-values below the diagonal are now best. How can this be explained?

Consecutive refinements can only be expected if the corrections get smaller and smaller. This gives the inequality :

$$|\psi_\ell| \geq |\psi_{\ell+1}| \ .$$

Using formula (50) this gives :

$$\frac{(\ell-1)!}{(k-\ell)\ n(n-1)\ldots(n-\ell+1)} \geq \frac{\ell!}{(k-\ell-1)\ n(n-1)\ldots(n-\ell)} \ .$$

Hence

$$\frac{\ell}{n-\ell} \leq \frac{k-\ell-1}{k-\ell} < 1$$

or

$$\ell < \frac{n}{2} \quad . \tag{53}$$

One can verify that $|\psi_\ell|$ increases from [n/2] + 1 on ([x] = the largest integer smaller than or equal to x). (53) implies that ℓ = [n/2] (n odd) or $\ell = n/2 - 1$ (n even) are the critical limits. Let us continue with the latter. Since $\ell \leq k - 1 < k$ we hence find the critical condition

$$\ell = \frac{n}{2} - 1 < k \ . \tag{54}$$

Interpretation

(i) In practice k is smaller than (n/2 - 1) since n is the size of the database and k is the number of relevant documents. In this case we use the diagonal elements (or even better : use (49) exactly, the diagonal elements are in fact $\bar{P}_p - \varphi$). If this is too much work then use any φ_ℓ , $\ell < k-1$ as a good approximation.

(ii) If $k > n/2 - 1$ then it is better <u>not</u> to use the diagonal elements : it is then optimal to use φ_ℓ for $\ell = n/2 - 1 < k$ (in the table : $\ell = 4$).

Part 2. The case of retrieval via T''

2.1 The general case $p \leq k$

We still work with DS = $\{D_1,...,D_n\}$ and retrieve now $\{D_i,...,D_m\}$ ($i \leq m$, i,m = 1,...,n) because of the use of T''. Suppose k documents are relevant and that we retrieve p of them ($p \leq k \leq n$). As said in this section on optimal searching, D_i and D_m belong to these p documents. Hence we have a situation as depicted in figure 9.

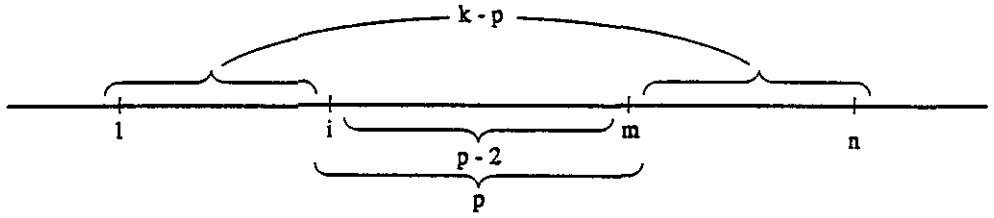

Figure 9 : Retrieval via $\mathcal{T}$", optimal searching

The recall value is still R = p/k while the precision is

$$P = \frac{p}{m - i + 1} . \tag{55}$$

These values occur

$$\binom{m-i-1}{p-2}\binom{n-m+i-1}{k-p}$$

times : p-2 documents in $\{D_{i+1},...,D_{m-1}\}$ and k-p documents in $\{D_1,...,D_{i-1}\} \cup \{D_{m+1},...,D_n\}$ (m = i+p-1,...,n; i = 1,...,n-p+1; p = 2,...,k). Hence here p = 0 or p = 1 is excluded since we have already the relevant documents D_i and D_m. We have now that optimal seaching via $\mathcal{T}$" is governed via the trivariate distribution

$$P(X'=i, Y'=m, Z'=p) = \frac{\binom{m-i-1}{p-2}\binom{n-m+i-1}{k-p}}{\frac{1}{2}k(k-1)\binom{n}{k}} . \tag{56}$$

Here we use that

$$\sum_{i=1}^{n-p+1} \sum_{m=i+p-1}^{n} \binom{m-i-1}{p-2}\binom{n-m+i-1}{k-p} \tag{57}$$

$$= \sum_{i'=0}^{n'-p'} \sum_{m'=i'+p'+1}^{n'+1} \binom{m'-i'-1}{p'}\binom{n'-m'+i'+1}{k'-p'} ,$$

where n' = n-2, k' = k-2, p' = p-2, m' = m-1, i' = i-1. We make again use of formula (22) but now with primes. In this way we obtain that (57) equals

$$(k'-p'+1)\binom{n'+2}{k'+2} = (k-p+1)\binom{n}{k} . \tag{58}$$

Finally

$$\sum_{p=2}^{k} (k-p+1)\binom{n}{k} = \frac{1}{k}\, k(k-1)\binom{n}{k}$$

yields the denominator of (56).

The marginal distributions are

$$P(Z'=p) \longleftarrow = \sum_{i=1}^{n-p+1} \sum_{m=i+p-1}^{n} \frac{\binom{m-i-1}{p-2}\binom{n-m+i-1}{k-p}}{\frac{1}{k}\, k(k-1)\binom{n}{k}}$$

$$= \frac{2(k-p+1)}{k(k-1)} \tag{59}$$

$$P(X'=i, Y'=m) = \sum_{p=2}^{k} \frac{\binom{m-i-1}{p-2}\binom{n-m+i-1}{k-p}}{\frac{1}{2}\, k(k-1)\binom{n}{k}} = \frac{2}{n(n-1)} , \tag{60}$$

using (7) again. From this the conditional distributions follow ($m = i+p-1,\ldots,n$; $i = 1,\ldots,n-p+1$; $p = 2,\ldots,k$)

$$P(X'=i, Y'=m \mid Z'=p) = \frac{P(X'=i, Y'=m, Z'=p)}{P(Z'=p)}$$

$$= \frac{\binom{m-i-1}{p-2}\binom{n-m+i-1}{k-p}}{(k-p+1)\binom{n}{k}} \tag{61}$$

$$P(Z'=p|X'=i,Y'=m) = \frac{P(X'=i,Y'=m,Z'=p)}{P(X'=i,Y'=m)}$$

$$= \frac{\binom{m-i-1}{p-2}\binom{n-m+i-1}{k-p}}{\binom{n-2}{k-2}}, \tag{62}$$

a hypergeometric distribution with parameters k-2, p-2 and with n replaced by n-2. The average is now

$$\mu'_{i,m} = (k-2)\,\frac{m-i-1}{n-2}\,. \tag{63}$$

We now find the average R and P-values

$$\overline{P}_{i,m} = \sum_{p=2}^{k} \frac{p}{m-i+1} P(Z'=p|X'=i,Y'=m)$$

$$\overline{P}_{i,m} = \sum_{p-2=0}^{k-2} \frac{p-2}{m-i+1} P(Z'=p|X'=i,Y'=m) + \frac{2}{m-i+1}$$

$$\overline{P}_{i,m} = \frac{k-2}{n-2}\,\frac{m-i-1}{m-i+1} + \frac{2}{m-i+1}$$

$$= \frac{k(m-i-1)+2(n-m+i-1)}{(n-2)(m-i+1)} \tag{64}$$

$$\overline{R}_{i,m} = \sum_{p=2}^{k} \frac{p}{k} P(Z'=p|X'=i,Y'=m)$$

$$= \sum_{p-2=0}^{k-2} \frac{p-2}{k} P(Z'=p|X'=i,Y'=m) + \frac{2}{k}$$

$$= \frac{k-2}{k}\,\frac{m-i-1}{n-2} + \frac{2}{k}$$

$$= \frac{k(m-i-1) + 2(n-m+i-1)}{k(n-2)}\,. \tag{65}$$

Again

$$\overline{R}_p = R = \frac{p}{k} \tag{66}$$

and

$$\overline{P}_p = \sum_{i=1}^{n-p+1} \sum_{m=i+p-1}^{n} \frac{p}{m-i+1}\, P(X'=i, Y'=m \mid Z'=p)$$

$$= \sum_{i=1}^{n-p+1} \sum_{n=i+p-1}^{n} \frac{p}{m-i+1}\, \frac{\binom{m-i-1}{p-2}\binom{n-m+i-1}{k-p}}{(k-p+1)\binom{n}{k}}\,. \tag{67}$$

As explained in the previous section this last formula cannot be transformed into elementary functions. Of course, the T''-analogue of figure 7 can be constructed. Its graph is depicted in figure 10.

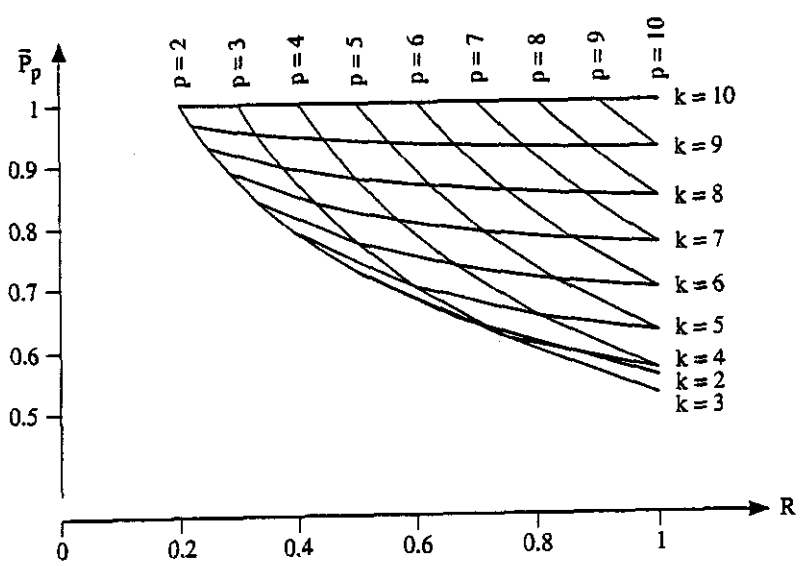

Figure 10

Recall-precision graph for optimal searching, using the similarity topology, for a document space containing n = 10 documents

For p = k, formula (67) reduces to

$$\overline{P}_p = \sum_{i=1}^{n-k+1} \sum_{m=i+k-1}^{n} \frac{k}{m - i + 1} \frac{\binom{m-i-1}{k-2}}{\binom{n}{k}} . \qquad (68)$$

Its graph can be depicted as in figure 11 : $\overline{P}_p$ versus k/n for n = 5, 7, 10, 15, 20.

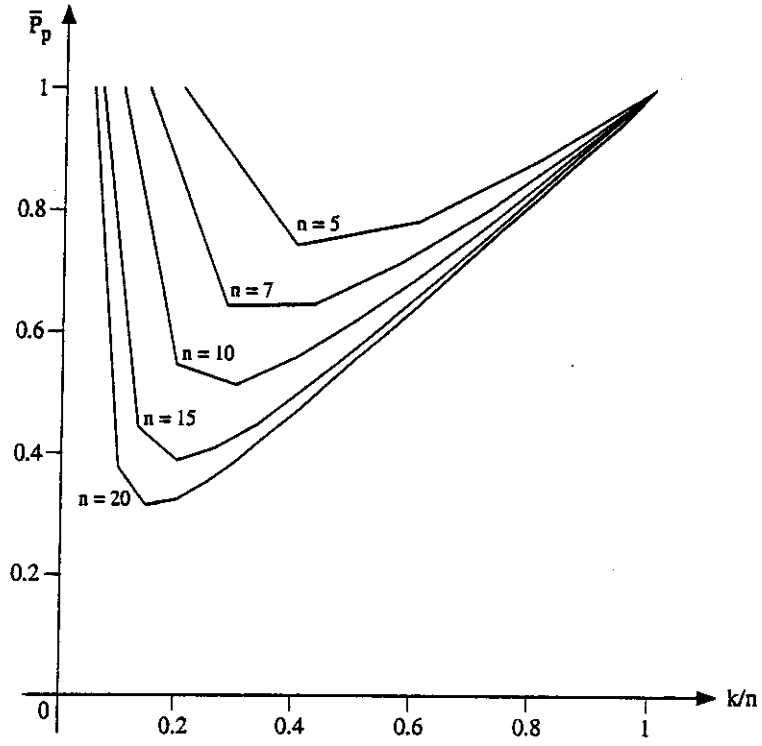

Figure 11

Average precision, using the similarity topology and optimal searching,
for document spaces containing n = 5, 7, 10, 15 and 20 documents.
The x axis shows k/n values, the y axis shows the corresponding average precision.

Based on this graph we make the following observations.

For a fixed k/n-value, the average precision decreases with n (except, of course, in 1, i.e. k = n). In addition, it seems that here too this average precision tends to k/n. The average precision in 1/n always begins at the value 1. For a fixed value of k, the average precision also decreases (probably to zero), except for k = 1. Finally, for a fixed value of n, the minimum value of the average precision decreases, but occurs at a larger value for k : it is first at k = 2, then at k = 3, and further calculations (not shown on the graph) show that it moves on to k = 4.

Note: if you allowed not only retrievals for the similarity topology, but also finite unions and intersections (i.e. the complete topology) this would always result in a precision equal to one. Yet, dropping the requirement that all similarity values are different, would reveal another aspect of the retrieval process. Indeed, in that case the precision in retrieving one document would be equal to the reciprocal of the number of documents with the same similarity value as the requested document. In general, retrieving k documents would yield a precision equal to k divided by the number of documents that have the same similarity values as - at least - one of the requested documents. So, in that case, the precision gives us information about the clustering of similarity values.

p 42 is empty

APPENDIX B: Topological spaces

In this appendix we recall, for the reader's convenience, some basic notions on topological spaces.

Let X be a set. Denote by $\wp(X)$ the set of all subsets of X. A topology τ on X is, by definition, a subset of $\wp(X)$ satisfying the following four axioms.

(O1) The empty set $\emptyset$ belongs to τ.

(O2) The set X itself belongs to τ.

(O3) Any union (hence, also infinite unions) of elements in τ belongs to τ.

(O4) Any finite intersection of elements of τ belongs to τ.

The elements of τ are called open sets. The couple (X, τ) is called a topological space. Note that the same set X can have many different topologies.

Given two topologies τ_1 and τ_2 on X, then τ_1 is said to be weaker or coarser than τ_2 if every element of τ_1 is also in τ_2. This relation between these topologies is also expressed by saying that τ_2 is stronger or finer than τ_1.

If (X,τ) is a topological space, then a subset $\mathcal{B}$ of $\wp(X)$ is a base for τ if every element of τ can be written as a union of elements in $\mathcal{B}$. A subbase for a topology τ is a subset $\mathcal{C}$ of $\wp(X)$ such that every element of τ is a union of finite intersections of elements of $\mathcal{C}$. This can also be expressed by saying that all finite unions of sets in $\mathcal{C}$ form a base for τ. Any collection of subsets of X is a subbase of some topology on X.

Consider a function f from the topological space (X,τ) to the topological space (Y,τ'). The function f is said to be continuous if the inverse image $f^{-1}(U) = \{ x \in X ; f(x) \in U \}$ of each open set U in (Y,τ') is open in (X,τ).

APPENDIX C

Theorem

$$\mathrm{MAX}_i \binom{n-i}{p}\binom{i}{k-p}$$

is attained for

$$i_m = \left\lfloor \frac{(k-p)n}{k} - \frac{p}{k} + 1 \right\rfloor . \tag{69}$$

If

$$\frac{(k-p)n}{k} \in \mathbb{N}$$

then

$$i_m = \frac{(k-p)n}{k} .$$

Proof:

We note first that for $p = k$, $i_m = 0$, hence (69) is correct for $k = p$. Let now, $p < k$ and let $i_m \in \mathbb{N}_0$ be such that for every $i \neq i_m$:

$$\binom{n-i_m}{p}\binom{i_m}{k-p} \geq \binom{n-i}{p}\binom{i}{k-p} . \tag{70}$$

Now, (70) is equivalent with :

$$\frac{(n-i_m)!}{(n-i)!}\,\frac{i_m!}{i!}\,\frac{(n-i-p)!}{(n-i_m-p)!}\,\frac{(i-k+p)!}{(i_m-k+p)!} \geq 1 . \tag{71}$$

a) Assume that $i < i_m$, then (71) is equivalent with :

$$\frac{i_m \ldots (i+1)\,(n-i-p) \ldots (n-i_m-p+1)}{(n-i) \ldots (n-i_m+1)\,(i_m-k+p) \ldots (i-k+p+1)} \geq 1$$

$$\Leftarrow \frac{n-i-p}{n-i} \cdots \frac{n-i_m+1-p}{n-i_m+1} \geq \frac{i_m-(k-p)}{i_m} \cdots \frac{i+1-(k-p)}{i+1}$$

or

$$\prod_{j=i}^{i_m-1} \left(1 - \frac{p}{n-j}\right) \geq \prod_{j=i}^{i_m-1} \left(1 - \frac{k-p}{j+1}\right) \quad .$$

This inequality is certainly satisfied if, for, every $j = i,\ldots,i_m-1$:

$$1 - \frac{p}{n-j} \geq 1 - \frac{k-p}{j+1}$$

or

$$j \leq \frac{(k-p)\,n}{k} - \frac{p}{k} \quad . \tag{72}$$

b) Assume now that $i > i_m$.

Interchanging the roles of i and i_m, and those of the numerator and the denominator of (71) leads to :

$$\prod_{j=i_m}^{i-1} \left(1 - \frac{p}{n-j}\right) \leq \prod_{j=i_m}^{i-1} \left(1 - \frac{k-p}{j+1}\right) \quad .$$

This inequality is certainly satisfied if, for every $j = i_m,\ldots,i-1$:

$$1 - \frac{p}{n-j} \leq 1 - \frac{k-p}{j+1}$$

or :

$$j \geq \frac{(k-p)n}{k} - \frac{p}{k} \quad . \tag{73}$$

Combining (72) and (73) proves this theorem.

APPENDIX D

Theorem

$\forall n \in \mathbb{N}, \forall p \le k \le n, p,k \in \mathbb{N}$ one has

$$\sum_{i=0}^{n-p} \sum_{m=i+p+1}^{n+1} \binom{m-i-1}{p} \binom{n-m+i+1}{k-p} = (k-p+1) \binom{n+2}{k+2} . \qquad (74)$$

Proof :

The formula can readily be checked for n = 1 (cases k = p = 0, p = 0 and k = 1, k = p = 1). We now prove the formula directly if k = n ($\forall n \in \mathbb{N}$). Then we have

$$\sum_{i=0}^{n-p} \sum_{m=i+p+1}^{n+1} \binom{m-i-1}{p} \binom{n-m+i+1}{n-p}$$

$$= \sum_{i=0}^{n-p} \sum_{\ell=p}^{n-i} \binom{\ell}{p} \binom{n-\ell}{n-p}$$

$$= \sum_{i=0}^{n-p} \binom{\ell}{p} \binom{n-p}{n-p} = n - p + 1 = (k-p+1) \binom{n+2}{k+2}$$

since $\ell \in \{p,...,n-i\} \Leftrightarrow n - \ell \in \{i,...,n-p\}$.

Next, we prove the formula directly if k = p ($\forall n \in \mathbb{N}$).

In this case (74) is

$$\sum_{i=0}^{n-k} \sum_{m=i+k+1}^{n+1} \binom{m-i-1}{k} = \sum_{i=0}^{n-k} \sum_{\ell=k}^{n-i} \binom{\ell}{k} = \sum_{i=0}^{n-k} \binom{n-i+1}{k+1} = \binom{n+2}{k+2} ,$$

where we used (4) twice (with different symbols) and k = p. This proves (74) in case k = p.

The rest of the proof is done by induction on n : by the above we can suppose (74) to be valid for n, $\forall p \le k \le n$ and have to prove it for n+1, $\forall p < k < n+1$ (since the cases k = p and k = n+1 have already been proved). We have

$$\sum_{i=0}^{n+1-p} \sum_{m=i+p+1}^{n+1} \binom{m-i-1}{p} \binom{n+1-m+i+1}{k-p}$$

$$= \sum_{i=0}^{n+1-p} \sum_{\ell=p}^{n+1-i} \binom{\ell}{p} \binom{n+1-\ell}{k-p}$$

$$= \sum_{i=0}^{n-p} \sum_{\ell=p}^{n-i} \binom{\ell}{p} \binom{n+1-\ell}{k-p} + \sum_{i=0}^{n+1-p} \binom{n+1-i}{p} \binom{i}{k-p}$$

(check the array $\sum_{i=0}^{n+1-p} \sum_{\ell=p}^{n+1-i}$).

Since

$$\binom{n+1-\ell}{k-p} = \binom{n-\ell}{k-p} + \binom{n-\ell}{k-1-p}$$

we have that the above equals

$$\sum_{i=0}^{n-p} \sum_{\ell=p}^{n-i} \binom{\ell}{p} \binom{n-\ell}{k-p} + \sum_{i=0}^{n-p} \sum_{\ell=p}^{n-i} \binom{\ell}{p} \binom{n-\ell}{k-1-p} + \sum_{i=0}^{n+1-p} \binom{n+1-i}{p} \binom{i}{k-p}.$$

We now apply the induction hypothesis on n and $p < k \leq n$ for the first term, the induction hypothesis on n and $p \leq k-1 < n$, for the second term and formula (4) on

$$\sum_{i=0}^{n+1-p} \binom{n+1-i}{p} \binom{i}{k-p} = \sum_{i=k-p}^{n+1-p} \binom{n+1-i}{p} \binom{i}{k-p} = \binom{n+2}{k+1} .$$

This yields

$$\sum_{i=0}^{n+1-p} \sum_{m=i+p+1}^{n+1} \binom{m-i-1}{p} \binom{n+1-m+i+1}{k-p}$$

$$= (k-p+1) \binom{n+2}{k+2} + (k-p) \binom{n+2}{k+1} + \binom{n+2}{k+1}$$

$$= (k-p+1) \left(\binom{n+2}{k+2} + \binom{n+2}{k+1} \right)$$

$$= (k-p+1)\binom{n+3}{k+2} \quad . \qquad \blacksquare$$

APPENDIX E

Theorem

The function

$$\bar{P}_p(n) = \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j} \tag{75}$$

satisfies

(i) $\bar{P}_p(n)$ decreases in n (k fixed)

(ii) $\lim_{\substack{n\to\infty \\ k=cst}} \bar{P}_p(n) = 0$

(iii) $\lim_{\substack{n\to\infty \\ k/n=cst}} \bar{P}_p(n) = \frac{k}{n}$.

Proof : (i)

$$\bar{P}_p(n+1) \leq \bar{P}_p(n)$$

$$\Leftarrow \sum_{j=k}^{n+1} \binom{j-1}{k-1} \frac{1}{j} \frac{n+1-k}{n+1} \leq \sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j}$$

$$\Leftarrow \binom{n}{k-1} \frac{1}{n+1} \leq \frac{k}{n+1} \sum_{j=k}^{n+1} \binom{j-1}{k-1} \frac{1}{j} \quad .$$

But

$$\frac{k}{n+1} \sum_{j=k}^{n+1} \binom{j-1}{k-1} \frac{1}{j} > \frac{k}{(n+1)^2} \sum_{j=k}^{n+1} \binom{j-1}{k-1} = \frac{k}{(n+1)^2} \binom{n+1}{k} = \frac{1}{n+1} \binom{k}{k+1}$$

(using (4) again).

(ii)

$$\sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j} \leq \sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j-1} = \frac{1}{k-1} \binom{n-1}{n-k}$$

The last equality follows from the lemma below this proof (this lemma will also be needed in Appendix F). Hence

$$0 \leq \bar{P}_p(n) \leq \frac{k^2}{n(k-1)} \quad .$$

Consequently,

$$\lim_{\substack{n\to\infty \\ k=cst}} \bar{P}_p(n) = 0 \quad .$$

(iii) From the above we have

$$\bar{P}_p(n) \leq \frac{k}{n} \frac{k}{k-1} \quad .$$

But $n \to \infty \Leftrightarrow k \to \infty$ since k/n = constant.
Hence

$$\lim_{\substack{n\to\infty \\ k/n=cst}} \bar{P}_p(n) \leq \frac{k}{n} \quad .$$

We will now prove that $\bar{P}_p(n) \geq k/n$, $\forall n \in \mathbb{N}$, $\forall k = 1,...,n$.

The proof goes by complete induction. For $n = 1$ is $k = 1$ and $\bar{P}_p(n) = 1 = k/n$. Let now this inequality be valid for n and all $k = 1,...,n$. Let now $k = 1,...,n+1$. For $k = n+1$ the inequality is trivial. Let now $k = 1,...,n$.

$$\frac{k}{\binom{n+1}{k}} \sum_{j=k}^{n+1} \binom{j-1}{k-1} \frac{1}{j}$$

$$= \frac{k}{\binom{n}{k}} \left(\sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j} \right) \frac{n+1-k}{n+1} + \frac{k}{\binom{n+1}{k}} \binom{n}{k-1} \frac{1}{n+1}$$

$$\geq \frac{k}{n} \frac{n+1-k}{n+1} + \frac{k^2}{(n+1)^2} ,$$

by the induction hypothesis since $k = 1,\ldots,n$.

Hence

$$\overline{P}_p(n) \geq \frac{k}{n+1} \left[1 + \frac{k}{n+1} - \frac{k-1}{n} \right] > \frac{k}{n+1}$$

since $k < n+1$. ■

Lemma : $\forall n \in \mathbb{N}, \forall m = 0,1,\ldots,n$

$$\sum_{j=k}^{n} \binom{j-m}{k-m} \frac{1}{j-m} = \frac{1}{k-m} \binom{n-m}{n-k} . \tag{76}$$

Proof :

For $m = 0$ we have to show that

$$\sum_{j=k}^{n} \binom{j}{k} \frac{1}{j} = \frac{1}{k} \binom{n}{k}$$

for all $n \in \mathbb{N}$. This is true for $n = 1$ and induction on n yields

$$\sum_{j=k}^{n+1} \binom{j}{k} \frac{1}{j} = \sum_{j=k}^{n} \binom{j}{k} \frac{1}{j} + \binom{n+1}{k} \frac{1}{n+1}$$

$$= \frac{1}{k}\binom{n}{k} + \binom{n+1}{k}\frac{1}{n+1}$$

$$= \frac{1}{k}\binom{n+1}{k} \quad .$$

Now follows the induction step on m : for m+1

$$\sum_{j=k}^{n}\binom{j-m-1}{k-m-1}\frac{1}{j-m-1} \qquad = \sum_{j'=k'}^{n'}\binom{j'-m}{k'-m}\frac{1}{j'-m}$$

(j' = j-1, k' = k-1, n' = n-1)

$$= \frac{1}{k'-m}\binom{n'-m}{n'-k'} = \frac{1}{k-m-1}\binom{n-m-1}{n-k}$$

by the induction hypothesis. ■

APPENDIX F

Lemma 1 :$\forall k,m,n \in \mathbb{N}$, $m \le k \le n$ one has

$$\sum_{j=k}^{n} \binom{j-m}{k-m} \frac{1}{j} = \frac{1}{k-m}\left[\binom{n-m}{n-k} - m \sum_{j=k}^{n} \binom{j-m-1}{k-m-1} \frac{1}{j}\right] . \tag{77}$$

Proof :

$$\binom{j-m}{k-m} \frac{1}{j} = \binom{j-m}{k-m} \frac{1}{j-m} - \binom{j-m-1}{k-m-1} \frac{m}{j(k-m)}$$

Hence

$$\sum_{j=k}^{n} \binom{j-m}{k-m} \frac{1}{j} = \sum_{j=k}^{n} \binom{j-m}{k-m} \frac{1}{j-m} - \frac{m}{k-m} \sum_{j=k}^{n} \binom{j-m-1}{k-m-1} \frac{1}{j} .$$

We now apply the lemma in Appendix E. This yields (77) directly. ■

Lemma 2 : $\forall n \in \mathbb{N}, \forall k \in \mathbb{N}, k \le n$:

$$\sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j} = \sum_{\ell=1}^{k-1} \binom{n-\ell}{n-k} \frac{(-1)^{\ell-1}\,(\ell-1)!}{(k-1)(k-2)\ldots(k-\ell)} + (-1)^{k-1} \sum_{j=k}^{n} \frac{1}{j} . \tag{78}$$

Proof :

We repeatedly apply the above lemma, yielding :

$$\sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j}$$

$$= \frac{1}{k-1}\left[\binom{n-1}{n-k} - \frac{1}{k-2}\left(\binom{n-2}{n-k} - 2 \sum_{j=k}^{n} \binom{j-3}{k-3} \frac{1}{j}\right)\right]$$

$$= \ldots\ldots$$

$$= \frac{1}{k-1}\binom{n-1}{n-k} - \frac{1}{(k-1)(k-2)}\binom{n-2}{n-k}$$

$$+ \frac{2}{(k-1)(k-2)(k-3)}\binom{n-3}{n-k} - \frac{2.3}{(k-1)(k-2)(k-3)(k-4)}\binom{n-4}{n-k}$$

$$+ \frac{2.3.4}{(k-1)(k-2)(k-3)(k-4)(k-5)}\binom{n-5}{n-k}$$

$$- \frac{2.3.4.5}{(k-1)(k-2)(k-3)(k-4)(k-5)} \sum_{j=k}^{n} \binom{j-6}{k-6} \frac{1}{j} .$$

From this it is clear that :

$$\sum_{j=k}^{n} \binom{j-1}{k-1} \frac{1}{j} = \sum_{\ell=1}^{k-1} \binom{n-\ell}{n-k} \frac{(-1)^{\ell-1}(\ell-1)!}{(k-1)(k-2)\ldots(k-\ell)} + (-1)^{k-1} \sum_{j=k}^{n} \frac{1}{j} .$$

■

Theorem 1 : $\forall k,n \in \mathbb{N}, k \leq n$:

$$\bar{P}_p = \sum_{\ell=1}^{k-1} \psi_\ell + \varphi , \tag{79}$$

where

$$\psi_\ell = \frac{(-1)^{\ell-1}(\ell-1)!\,k^2}{(k-\ell)\; n(n-1)\ldots(n-\ell+1)} \tag{80}$$

$$\varphi = (-1)^{k-1} \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \frac{1}{j} . \tag{81}$$

Proof :

This follows readily from lemma 2 and the formula (48) for $\bar{P}_p$. ■

Theorem 2

If we denote

$$\varphi_i = \sum_{\ell=1}^{i} \psi_\ell$$

then, $\forall i = 1,...,k-1$ and k fixed

$$\lim_{n\to\infty} \left| \overline{P}_p - \varphi_i \right| n^i = 0 \quad . \tag{82}$$

Proof : $\forall i = 1,...,k-1$

$$\left| \overline{P}_p - \varphi_i \right| \le \sum_{\ell=i+1}^{k-1} \left| \psi_\ell \right| + |\varphi| \quad .$$

But, for k fixed is

$$0 \le \lim_{n\to\infty} \left| \psi_\ell \right| n^{\ell-1} \le \lim_{n\to\infty} \frac{(\ell-1)!\ k^2\ n^{\ell-1}}{(k-\ell)\ n^\ell} = 0$$

and

$$0 \le \lim_{n\to\infty} |\varphi|\ n^{k-1} \le \lim_{n\to\infty} \frac{(k-1)!\ k^2\ n^{k-1}\ \ell n\ n}{n^k} = 0 \quad .$$

Here we used the fact that

$$\lim_{n\to\infty} \frac{\sum_{j=1}^{n} \frac{1}{j}}{\ell n\ n} = 1$$

and that

$$\lim_{n\to\infty} \frac{n^\ell}{n(n-1)\ldots(n-\ell+1)} = 1 \quad .$$

Hence we have

$$0 \le \lim_{n\to\infty} \left|\overline{P}_p - \varphi_i\right| n^i$$

$$\le \sum_{\ell=i+1}^{k-1} \lim_{n\to\infty} \left|\psi_\ell\right| n^i + \lim_{n\to\infty} |\varphi|\, n^i$$

$$\le \sum_{\ell=i+1}^{k-1} \lim_{n\to\infty} \left|\psi_\ell\right| n^{\ell-1} + \lim_{n\to\infty} |\varphi|\, n^{k-1} = 0 ,$$

since ℓ-1 $\ge$ i and i $\le$ k-1. ■

APPENDIX G

In this final apppendix we refer again to the problems of dealing with formula (48)

$$\overline{P}_p = \frac{k}{\binom{n}{k}} \sum_{i=1}^{n-k+1} \frac{k}{n - i + 1} \binom{n-i}{k-1} \tag{48}$$

and similar formulae, all in the context of optimal searching. Formula (48) above refers to the case $p = k$ and to retrieval via $\mathcal{T}$ (the simplest case). We have presented a non-trivial but proper approximation to (48), hereby reducing the sum up to n (large) to a sum up to k (small).

Another way to study the average behavior of P is to look at the median. We proceed as follows. Since there are (in total) $\binom{n}{k}$ cases we must find $s \in \{1,...,n\}$ such that

$$\frac{1}{2}\binom{n}{k} = \sum_{i=1}^{j} \binom{n-i}{k-1} . \tag{83}$$

But

$$\sum_{i=1}^{j} \binom{n-i}{k-1} = \sum_{i=1}^{n-k+1} \binom{n-i}{k-1} - \sum_{i=j+1}^{n-k+1} \binom{n-i}{k-1} = \binom{n}{k} - \binom{n-j}{k}$$

hence we have to solve

$$\frac{1}{2}\binom{n}{k} = \binom{n-j}{k} . \tag{84}$$

Hence

$$\prod_{i=0}^{j-1} \frac{n - i}{n - k - i} = 2$$

or

$$\frac{1}{1 - \frac{k}{n}} \; \frac{1}{1 - \frac{k}{n-1}} \; \cdots\cdots \; \frac{1}{1 - \frac{k}{n-j+1}} = 2 \; . \tag{85}$$

Since k << n we can delete second (and higher) order factors (we assume that also k << n-j, as j is a median). This yields:

$$\frac{1}{1 - \sum_{i=0}^{i-1} \frac{k}{n-i}} = 2 \; . \tag{86}$$

But

$$\sum_{i=0}^{i-1} \frac{k}{n - i} = \sum_{\ell=1}^{n} \frac{k}{\ell} - \sum_{\ell=1}^{n-j} \frac{k}{\ell}$$

$$\approx \ell n \left(\frac{n}{n - j} \right) \tag{87}$$

again since n-j is high. Together, formulae (86) and (87) yield

$$\frac{1}{1 - k \, \ell n \left(\frac{n}{n - j} \right)} \approx 2 \; .$$

Hence

$$j \approx n - \frac{n}{e^{1/2k}} \; . \tag{88}$$

For the median precision $M_d(P)$ we have

$$M_d(P) = \frac{k}{n - [j] + 1} \approx \frac{k}{2 + \left\lfloor \frac{n}{e^{1/2k}} \right\rfloor}$$

$$\left(\left\lfloor n - 1 - \frac{n}{e^{1/2k}} \right\rfloor = n - 1 - \left(\left\lfloor \frac{n}{e^{1/2k}} \right\rfloor + 1 \right) \right) \; .$$

So

$$M_d(P) \approx \frac{k}{2 + \frac{n}{e^{1/2k}}} \quad . \tag{89}$$

This is, roughly, of the order of k/n, a result that was already found for $\bar{P}_p$. Some values for n = 100.

k	$M_d(P)$
1	0.01613
2	0.02532
3	0.03488
50	0.49505
100	0.99010

Note. We can say that all "average" P-values (in all cases) are around k/n being the precision value obtained when we retrieve (or take) the complete database DS!