On Retrieval Properties of Samples of Large Collections

David Madigan[*]
Avaya Labs Research
233 Mount Airy Road
Basking Ridge, NJ 07920
madigan@stat.rutgers.edu

Yehuda Vardi[†]
Department of Statistics
Rutgers University
Piscataway, NJ 08855
vardi@stat.rutgers.edu

Ishay Weissman
Technion - Israel Institute of Technology
Technion City, Haifa 32000
Israel
ieriw01@ie.technion.ac.il

Abstract

We consider text retrieval applications that assign query-specific relevance scores to documents drawn from particular collections. Such applications represent a primary focus of the annual Text Retrieval Conference (TREC) where the participants compare the empirical performance of different approaches. "P@K," the proportion of the top K documents that are relevant, is a popular measure of retrieval effectiveness.

Participants in the TREC Very Large Corpus track have observed that P@K increases substantially when moving from a sample to the full collection. Hawking *et al.* (1999) posed as an open research question the cause of this phenomenon and proposed five possible explanatory hypotheses. In this paper we present a mathematical analysis of the phenomenon. We will also introduce "contamination at K," the number of irrelevant documents amongst the top K relevant documents, and describe its properties.

Our analysis shows that while P@K typically will increase with collection size, the phenomenon is not universal. That is, there exist score distributions for which P@K (and C@K) approach a constant limit as collection size increases.

1 Introduction

For TREC-like ad-hoc retrieval tasks, P@K, the proportion of the top K documents that are relevant, is a popular measure of retrieval effectiveness. Participants in the TREC Very Large Corpus (VLC) track have observed that P@K increases substantially when moving from a 10% sample to the the full collection of over seven million documents (Hawking *et al.*, 1999). This paper presents an analysis of this phenomenon.

In what follows we will:

- Provide a general mathematical expression for P@K, evaluate the expression in closed-form under particular distributional assumptions for document scores, and present an asymptotic analysis.
- Introduce "contamination at K", C@K, the expected number of irrelevant documents amongst the top K relevant documents and describe some of its properties.

Our analysis shows that while P@K typically will increase in expectation with collection size, the phenomenon is not universal. That is, there exist score distributions for which the expected values of P@K (and C@K) approach a constant limit as collection size increases.

Section 2 below describes the mathematical framework and presents the key results. Section 3 presents some examples. Section 4 considers asymptotic behavior and Section 5 suggests some future directions.

2 Mathematical Framework and Key Results

We consider a text retrieval application that, in response to a given query, assigns relevance scores to documents drawn from a collection. Denote by G the probability distribution of the scores of the relevant documents in the collection, and denote by F the probability distribution of the scores of the irrelevant documents. Let g and f denote the corresponding probability density functions. We assume that the application assigns scores $\boldsymbol{X} = \{X_1, \ldots, X_M\}$ to M documents drawn randomly from F, and scores $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\}$

[*] also with Rutgers University
[†] also with Avaya Labs Research

to N documents drawn randomly from G. We assume independence of scores across all documents.

We make the restrictive assumption that the score assigned to a particular document is independent of M and N. While this will be true for certain text retrieval applications (e.g., Boolean retrieval), it will generally not be true for applications that use collection-wide statistics such as *idf* (inverse document frequency). For very large collections, however, the distinction is likely to be unimportant.

Let $Z_1 > \cdots > Z_{M+N}$ be the ordered pooled scores $\boldsymbol{X} \cup \boldsymbol{Y}$ and let δ_i, $i = 1, \ldots, M+N$, be an indicator variable:

$$\delta_i = \begin{cases} 0 & \text{if } Z_i \in \boldsymbol{X} \\ 1 & \text{if } Z_i \in \boldsymbol{Y}. \end{cases}$$

Let T_K be the number of relevant documents among the K highest scores, in the pooled sample $\boldsymbol{Z} = \{Z_1, \ldots, Z_{M+N}\}$. (We assume $K < \min(M, N)$.) Then, $T_K = \sum_{i=1}^{K} \delta_i$.

Define *precision at* K as the proportion of relevant documents in the top K scores, i.e., T_K/K. We denote the expected value of T_K/K as P@K:

$$P@K = ET_K/K = K^{-1} \sum_{k=0}^{K} k \times P\{T_K = k\}.$$

Define *contamination at* K as the number of irrelevant documents scoring higher than the Kth highest relevant score, i.e., the number of X's exceeding the Kth highest Y score. As in the case of P@K, we denote its expected value as C@K.

In what follows, we use the following order statistics notation:

$X_{(1)} > \cdots > X_{(M)}$ (ordered X's),

$Y_{(1)} > \cdots > Y_{(N)}$ (ordered Y's),

and

$Y_{(0)} = X_{(0)} \equiv +\infty$, $X_{(M+1)} = Y_{(N+1)} \equiv -\infty$.

Standard order statistics theory uses the probability transform to simplify analyses; in fact, any increasing transformation applied to the scores leaves P@K and C@K invariant. Thus, there is no loss of generality in assuming:

$X_1, \ldots, X_M \sim H \equiv FG^{-1}$, and $Y_1, \ldots, Y_N \sim U[0,1]$,

or, alternatively,

$X_1, \ldots, X_M \sim U[0,1]$, and $Y_1, \ldots, Y_N \sim \tilde{H} \equiv H^{-1} \equiv GF^{-1}$.

2.1 A General Expression for P@K

Proposition 2.1 below, provides an explicit expression for $\mathrm{P}\{T_K = k\}$.

PROPOSITION 2.1. *For $k = 0, \ldots, K$, $\mathrm{P}\{T_K = k\}$ is given by:*

$$\mathrm{P}\{T_K = k\} = \binom{M}{l}\binom{N}{k}\left\{\int_0^1 t^{N-k}(1-t)^k \times H(t)^{M-l}(1-H(t))^l \left[l\frac{h(t)}{H(t)} + \frac{k}{1-t}\right] dt\right\},$$

where $H(t) = F(G^{-1}(t))$, $h(t) = H'(t)$ and $\ell = K - k$.

PROOF. For $1 \le K \le \min(M, N)$, an integer, and $k = 0, 1, \ldots, K$, we have:

$$\begin{aligned} P\{T_K = k\} &= P\Big(\sum_{i=1}^{K} \delta_i = k\Big) \\ &= P\Big\{\underbrace{[Y_{(k+1)} < X_{(K-k)} < Y_{(k)}]}_{A_k} \cup \underbrace{[X_{(K-k+1)} < Y_{(k)} < X_{(K-k)}]}_{B_k}\Big\} \\ &= P\{A_k\} + P\{B_k\}. \end{aligned}$$

Note that $A_k \cap B_k = \emptyset$ and $A_K = B_0 = \emptyset$. Based on the multinomial distribution, for $k \ge 1$, we have,

$$P\{A_k\} = C(N, M, k) \int\int\int_{0 \le y < x < y' \le 1} y^{N-k-1}(1-y')^{k-1} H(x)^{M-K+k} \bar{H}(x)^{K-k-1} h(x)\, dx\, dy\, dy', \quad (1)$$

where $\bar{H} = 1 - H$ and

$$\begin{aligned} C(N, M, k) &= \binom{N}{k-1, N-k-1, 1}\binom{M}{\ell - 1, M - \ell} \\ &= k(N-k)\ell \binom{N}{k}\binom{M}{\ell}. \end{aligned}$$

Now,

$$\int_x^1 (1-y')^{k-1} dy' = \int_0^{1-x} t^{k-1} dt = \frac{1}{k}(1-x)^k, \quad (2)$$

and

$$\int_0^x y^{N-k-1} dy = \frac{1}{N-k} x^{N-k}. \quad (3)$$

Substituting (1-3) into (1), we get

$$P\{A_k\} = \ell \binom{N}{k}\binom{M}{\ell} \int_0^1 H(x)^{M-\ell} \bar{H}(x)^{\ell-1} h(x) x^{N-k} (1-x)^k dx. \quad (4)$$

Turning now to the $P\{B_k\}$, we have:

$$
\begin{aligned}
&P\{B_k\}\\
&= P\{X_{(\ell+1)} < Y_{(k)} < X_{(\ell)}\}\\
&= P\{F(X_{(\ell+1)}) < F(Y_{(k)}) < F(X_{(\ell)})\}\\
&= C(M,N,l)\int\int\int_{0\le x<y<x'\le 1}\\
&x^{M-\ell-1}(1-x')^{\ell-1}\tilde{\bar{H}}(y)^{N-k}\tilde{\bar{H}}(y)^{k-1}\tilde{h}(y)dydxdx'\\
&= k\binom{M}{\ell}\binom{N}{k}\\
&\int_0^1 \tilde{H}(y)^{N-k}\tilde{\bar{H}}(y)^{k-1}\tilde{h}(y)y^{M-\ell}(1-y)^{\ell}dy.
\end{aligned}
$$

Let $t = \tilde{H}(y)$ so that $H(t) = y$ and $dt = \tilde{h}(y)dy$. Therefore,

$$P\{B_k\} = k\binom{M}{\ell}\binom{N}{k}\int_0^1 t^{N-k}(1-t)^{k-1}H(t)^{M-\ell}\bar{H}(t)^{\ell}dt$$

and the result follows.

2.2 A General Expression for C@K

Contamination at $K \equiv$ C@$K \equiv$ Expected number of irrelevant docs exceeding (in score) the K-th largest relevant score

$= E[\#X'\text{s exceeding } Y_{(K)}]$.

We have

$$
\begin{aligned}
&\text{C@}K = EE[\#X'\text{s exceeding } Y_{(K)}|\boldsymbol{Y}]\\
&= E[\text{Binomial}(M, \bar{H}(Y_{(K)}))]\\
&= ME\bar{H}(Y_{(K)})\\
&= M\int_0^1 \bar{H}(t)a_K(t)dt\\
&= M\int_0^1 A_K(t)h(t)dt, \qquad (5)
\end{aligned}
$$

where for $0 \le t \le 1$,

$$a_K(t) = K\binom{N}{K}t^{N-K}(1-t)^{K-1}$$

and

$$A_K(t) = \sum_{j=N-K+1}^{N}\binom{N}{j}t^j(1-t)^{N-j}$$

are respectively, the density function and the cumulative distribution function of the Kth largest order statistic of a sample of size N from the uniform distribution on $[0,1]$.

Therefore

$$
\begin{aligned}
&\text{C@}K = MK\binom{N}{K}\int_0^1 \bar{H}(t)t^{N-K}(1-t)^{K-1}dt\\
&= M\sum_{j=N-K+1}^{N}\binom{N}{j}\int_0^1 t^j(1-t)^{N-j}h(t)dt \qquad (6)
\end{aligned}
$$

and for calculations, we can use either one of these expressions, whichever is more convenient.

3 Specific Examples

For certain probability distributions, the integral in the expression for P$\{T_K = k\}$ is available in closed form, yielding in turn, explicit expressions for P@K and C@K.

PROPOSITION 3.1. *If F is the exponential distribution with parameter λ_x and G is the exponential distribution with parameter λ_y, then, for $\beta = \lambda_x/\lambda_y$:*

$$
\begin{aligned}
&\text{P@K} = \tfrac{1}{K}\sum_{k=1}^{K} k\left(\tfrac{(K-k)\beta}{\Gamma(k+1)} + \tfrac{1}{\Gamma(k)}\right)\cdot\sum_{j=0}^{M-K+k}\\
&(-1)^j\frac{\Gamma(N+1)\Gamma(M+1)\Gamma\big(\beta(K-k+j)+k\big)}{\Gamma(K-k+1)\Gamma(j+1)\Gamma(M-K+k-j+1)\Gamma\big(\beta(K-k+j)+N+1\big)},
\end{aligned}
$$

and

$$\text{C@K} = M\frac{\Gamma(N+1)\Gamma(K+\beta)}{\Gamma(N+\beta+1)\Gamma(K)}.$$

PROOF. First note that $H(t) = F\big(G^{-1}(t)\big) = 1-(1-t)^{\beta}$, $0 \le t \le 1$. The distribution H is in fact the Beta(1,β).

From (4) we have

$$
\begin{aligned}
&P\{A_k\} = \ell\binom{N}{k}\binom{M}{\ell}\int_0^1\big(1-(1-t)^{\beta}\big)^{M-\ell}\big((1-t)^{\beta}\big)^{\ell-1}\\
&\cdot\beta(1-t)^{\beta-1}\cdot t^{N-k}(1-t)^k\,dt\\
&= \ell\beta\binom{N}{k}\binom{M}{\ell}\int_0^1\big(1-(1-t)^{\beta}\big)^{M-\ell}(1-t)^{\beta\ell-1+k}t^{N-k}\,dt\\
&= \ell\beta\binom{N}{k}\binom{M}{\ell}\sum_{j=0}^{M-\ell}\binom{M-\ell}{j}(-1)^j\int_0^1(1-t)^{\beta(\ell+j)-1+k}t^{N-k}\,dt\\
&= \ell\beta\frac{\Gamma(N+1)\Gamma(M+1)}{\Gamma(k+1)\Gamma(\ell+1)}\sum_{j=0}^{M-\ell}\\
&\quad(-1)^j\frac{\Gamma\big(\beta(\ell+j)+k\big)}{\Gamma(j+1)\Gamma(M-\ell-j+1)\Gamma\big(\beta(\ell+j)+N+1\big)}.
\end{aligned}
$$

Similarly

$$
\begin{aligned}
&P\{B_k\} =\\
&\frac{\Gamma(N+1)\Gamma(M+1)}{\Gamma(k)\Gamma(\ell+1)}\sum_{j=0}^{M-\ell}(-1)^j\frac{\Gamma\big(\beta(\ell+j)+k\big)}{\Gamma(j+1)\Gamma(M-\ell-j+1)\Gamma\big(\beta(\ell+j)+N+1\big)}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\text{P@K} = \tfrac{1}{K}\sum_{k=1}^{K} k\left(\tfrac{\ell\beta}{\Gamma(k+1)} + \tfrac{1}{\Gamma(k)}\right)\cdot\sum_{j=0}^{M-K+k}\\
&(-1)^j\frac{\Gamma(N+1)\Gamma(M+1)\Gamma\big(\beta(\ell+j)+k\big)}{\Gamma(\ell+1)\Gamma(j+1)\Gamma(M-K+k-j+1)\Gamma\big(\beta(\ell+j)+N+1\big)}.
\end{aligned}
$$

To compute C@K, we use the first expression of (6):

$$\text{C@K} = MK\binom{N}{K}\int_0^1(1-t)^{\beta}t^{N-K}(1-t)^{K-1}dt$$

and the desired result follows.

PROPOSITION 3.2. *If $F(t) = t^{\theta}$ and $G(t) = t^{\eta}, 0 \le t \le 1$, then, for $\alpha = \theta/\eta$ we have:*

$$
\begin{aligned}
&\text{P@K} = \tfrac{1}{K}\sum_{k=1}^{K} k\frac{\Gamma(M+1)\Gamma(N+1)}{\Gamma(M-l+1)\Gamma(N-k+1)}\times\\
&\Big(\sum_{j=0}^{l-1}\alpha(-1)^j\frac{\Gamma(\alpha(M-l)+N-k+\alpha(j+1))}{\Gamma(j+1)\Gamma(l-j)\Gamma(\alpha(M-l)+N+\alpha(j+1)+1)}\\
&+\sum_{j=0}^{l}(-1)^j\frac{\Gamma(N-k+\alpha(M-l)+\alpha j+1)}{\Gamma(j+1)\Gamma(l-j+1)\Gamma(N+\alpha(M-l)+\alpha j+1)}\Big),
\end{aligned}
$$

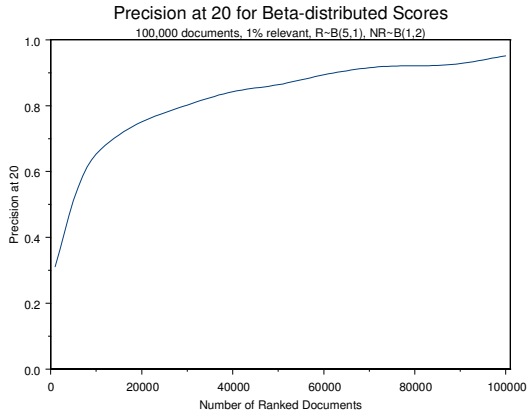

Figure 1: *P@20 for the Beta-Beta case.*

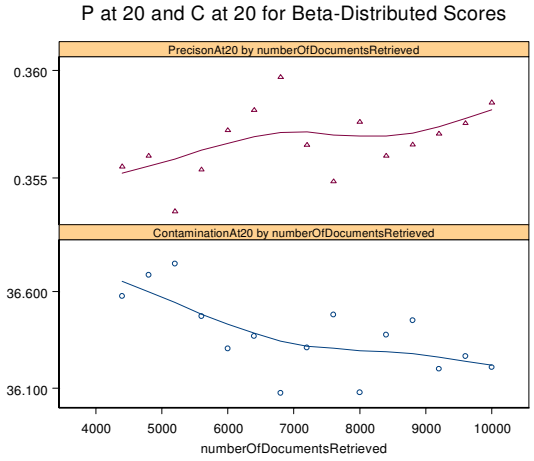

Figure 2: *P@20 and C@20 for the Beta-Beta case.*

$$\text{C@K} = M\{1 - \frac{\Gamma(N+1)}{\Gamma(N+1+\alpha)}\frac{\Gamma(N-K+1+\alpha)}{\Gamma(N-K+1)}\}.$$

PROOF. Omitted.

Figure 1 shows P@20 for $X_i \sim \text{Beta}(1,2)$ and $Y_i \sim \text{Beta}(5,1)$ and shows P@20 increasing as the number of ranked documents increases. This is consistent with the behavior observed at TREC. Figure 2 shows similar behavior for $X_i \sim \text{Beta}(1,1)$ and $Y_i \sim \text{Beta}(5,1)$ and also shows C@20. Monte Carlo simulations generated these Figures. Friedman's SuperSmoother fitted the curves to the simulations.

Manmatha *et al.* (2001) provided empirical evidence that for a number of different search engines a Gaussian distribution provides a reasonable fit to the scores of relevant documents (G) while an exponential distribution provides a reasonable fit to the irrelevant documents (F). Figure 3 presents P@20 and C@20 results for a Gaussian distribution with mean 1 and standard deviation 5 and an exponential distribution with rate

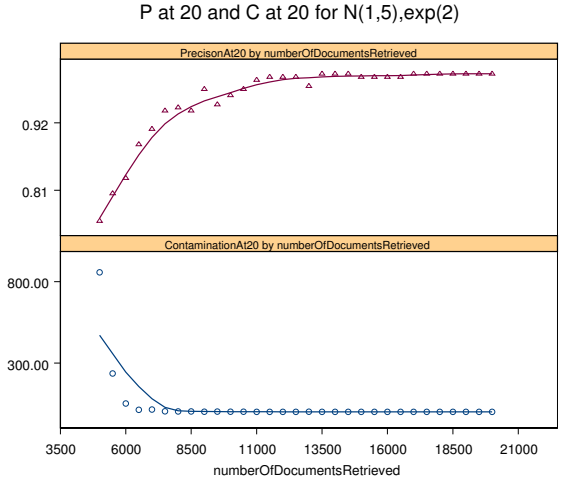

Figure 3: *P@20 and C@20 for the Normal-Exponential case. 1% relevant.*

2. Again, as the number of ranked documents increases, P@20 increases (and C@20 decreases). However, Figure 4 shows an example with a different Gaussian mean where P@20 does initially increase with collection size but ultimately decreases. In fact, P@20 ultimately goes to zero for the pairs of distributions considered in Figures 3 and 4.

4 Asymptotics

Within the framework described above it is possible to study limiting behavior of P@K and C@K for specific distributions. For example, it is straightforward to derive the following two results:

PROPOSITION 4.2. *For* $F(t) = t^\theta, G(t) = t^\eta, 0 \le t \le 1, \alpha = \theta/\eta$*:*

(1) If $N \to \infty$ *and* $\frac{M}{N} \to 0$*, then* $C@K \to 0$*.*

(2) If $N \to \infty$ *and* $\frac{M}{N} \to \lambda > 0$*, then* $\text{C@}K \to \alpha\lambda K$*.*

PROOF. For large N, Stirling's formula applied to C@K in Proposition 3.2 implies:

$$\text{C@K} \sim \alpha K M N^{-1}$$

and the result follows.

This result is intuitively expected—C@K is proportional to λ, namely, C@K is increasing if the number of irrelevant scores is increasing; proportional to α, namely, if the irrelevant scores are stochastically larger, then C@K is larger.

PROPOSITION 4.2. *For* $F(t) = 1 - e^{-\theta t}, G(t) = 1 - e^{-\eta t}, 0 \le t, \beta = \theta/\eta$ *(i.e., F and G exponential):*

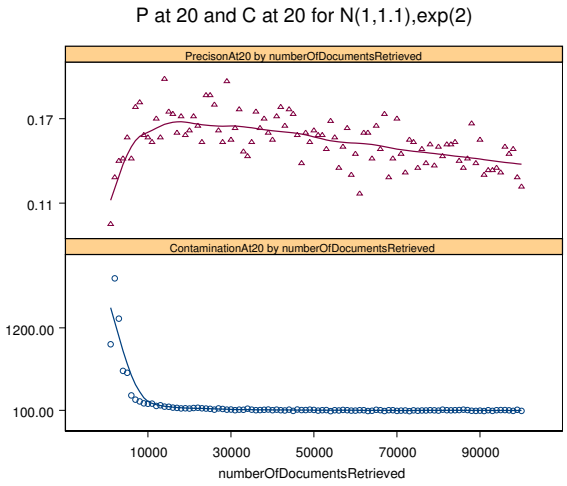

Figure 4: *P@20 and C@20 for the Normal-Exponential case. 1% relevant.*

(1′) If $N \to \infty$ *and* $\frac{M}{N^\beta} \to 0$, *then* $C@K \to 0$.

(2′) If $N \to \infty$ *and* $\frac{M}{N^\beta} \to \lambda > 0$, *then* $\mathrm{C@}K \to \lambda\Gamma(\beta+K)/\Gamma(K)$.

PROOF. As before, based on Proposition 3.1,

$$\mathrm{C@K} \sim MN^{-\beta}\Gamma(\beta+K)/\Gamma(K)$$

and the result follows.

These results for C@K, and similar but more involved results for P@K, show that for certain distributions, C@K and P@K can approach a constant limit as collection size increases.

4.1 A Mixture Model

Thus far, M and N are considered fixed. An alternative formulation instead assumes that the population of scores consists of two types with proportions p and $q = 1-p, 0 < p < 1$. Sampled scores $Z_1, \ldots, Z_n$ now arise from a mixture distribution:

$$Z \sim H_m = pG + qF.$$

We say that F and G are *tail-equivalent* if:

$$x^* := \sup\{x : F(x) < 1\} = \sup\{x : G(x) < 1\} \leq \infty.$$

and

$$\lim_{x \uparrow x^*} \frac{\bar{G}(x)}{\bar{F}(x)} = c, \quad 0 < c < \infty.$$

Without tail equivalence, as $n \to \infty$, $T_K \xrightarrow{P} 0$ or to K, and consequently P@K converges in probability to zero or one.

It is interesting to note that the Gaussian and exponential distributions are *not* tail equivalent. For exponential F and Gaussian G (i.e., the empirical model of Manmatha *et al.*, 2001):

$$\lim_{x \to \infty} \frac{\bar{G}(x)}{\bar{F}(x)} = 0$$

and consequently P@K converges in probability to zero. This is consistent with the non-mixture model results for P@20 in Figure 4. Extensive TREC evidence showing P@K continuing to increase even at a collection size of seven million documents, suggests that the Gaussian/Exponential model may be inadequate for analyzing tail behavior of score distributions.

4.2 A Modified P@K

The mixture model facilitates asymptotic analysis of an alternative "early precision" measure, $\mathrm{P^*@K}$. First define a threshold $u_n = H_m^{-1}(1 - \frac{K}{n})$. Let:

$$\begin{aligned} N_Z(u_n) &= \#\mathrm{of}\, Z_i > u_n, \\ N_Y &= T_{N_Z} = \#\mathrm{of}\, Z_i = Y_i > u_n, \mathrm{and} \\ N_X &= N_Z - N_Y. \end{aligned}$$

Then:

$$\begin{aligned} N_Z &\sim B(n, \frac{K}{n}), \\ N_Y &\sim B(n, p\bar{G}(u_n)), \mathrm{and} \\ N_X &\sim B(n, q\bar{F}(u_n)). \end{aligned}$$

In this context, $\mathrm{P^*@K}$ is the expected value of N_Y/N_Z, conditioned on $N_Z > 0$.

PROPOSITION 4.3. *As* $n \to \infty, (N_Y, N_X) \xrightarrow{D} (J_1, J_2)$ *where* J_1 *and* J_2 *are independent Poisson with parameters* p^*K *and* q^*K *respectively, and* $p^* = \frac{pc}{pc+q}$ and $q^* = 1 - p^*$.

PROOF. The marginal convergence in distribution is clear, since:

$$np\bar{G}(u_n) \to \frac{pcK}{pc+q},$$

and

$$nq\bar{F}(u_n) \to \frac{qK}{pc+q}.$$

To show the marginal independence, use the joint characteristic function:

$$E\exp\{itN_Y + isN_X\}$$

$$= (E[\exp\{it1_{\{Z=Y>u_n\}} + is1_{\{Z=X>u_n\}}\}])^n$$

$$= \{1 - \frac{K}{n} + p\bar{G}(u_n)e^{it} + q\bar{F}(u_n)e^{is}\}^n$$

$$\simeq \{1 + \frac{p^*K(e^{it}-1) + q^*K(e^{is}-1)}{n}\}^n$$
$$\rightarrow \exp\{p^*K(e^{it}-1)\} \cdot \exp\{q^*K(e^{is}-1)\}.$$

An immediate consequence is $N_Z \xrightarrow{D} J = J_1 + J_2 \sim Poisson(K)$. We also have this corollary:

COROLLARY 4.4. *Asymptotically, conditioned on* $N_Z > 0$, $\text{P}^*@K = p*$.

PROOF. The proof follows from the fact that if W, V are independent Poisson random variables and $T = W + V$, then, conditioned on T, W has a binomial distribution.

Thus, the asymptotic behavior of P*@K depends on the mixing proportion p, and on c, which describes the relative limiting behavior of $\bar{F}$ and $\bar{G}$. Finally, we present some specific examples.

1. *Exponential*

$$\left.\begin{array}{rcl} \bar{F}(x) & = & e^{-x}, (x > 0) \\ \bar{G}(x) & = & e^{-(x-\theta)}, (x > \theta) \end{array}\right\} \quad \frac{\bar{G}(x)}{\bar{F}(x)} = e^{\theta} \equiv c$$

2. *Pareto*

$$\left.\begin{array}{rcl} \bar{F}(x) & = & x^{-\alpha}, (x > 1) \\ \bar{G}(x) & = & (\theta x)^{-\alpha}, (x > 1/\theta) \end{array}\right\} \quad \frac{\bar{G}(x)}{\bar{F}(x)} = \theta^{-\alpha} \equiv c$$

3. *Gaussian*

$$\left.\begin{array}{l} X_i \sim N(0,1) \\ Y_i \sim N(\theta,1) \end{array}\right\} \quad \frac{\bar{G}(x)}{\bar{F}(x)} \sim e^{-\theta^2/2} e^{x\theta}$$
$$\stackrel{x\rightarrow\infty}{\rightarrow} \left\{\begin{array}{ll} \infty & \theta > 0 \\ 1 & \theta = 0 \\ 0 & \theta < 0 \end{array}\right.$$

4. *Gaussian*

$$\left.\begin{array}{l} X_i \sim N(0,1) \\ Y_i \sim N(0,\theta^2) \end{array}\right\} \quad \frac{\bar{G}(x)}{\bar{F}(x)} \sim \frac{1}{\theta} e^{\frac{x^2}{2}(1-\frac{1}{\theta^2})}$$
$$\stackrel{x\rightarrow\infty}{\rightarrow} \left\{\begin{array}{ll} \infty & \theta > 1 \\ 1 & \theta = 1 \\ 0 & \theta < 1 \end{array}\right.$$

5. *Beta*

$$\left.\begin{array}{l} X_i \sim \beta(\alpha_1, \beta_1) \\ Y_i \sim \text{ is } \beta(\alpha_2, \beta_2) \end{array}\right\} \quad \frac{\bar{G}(x)}{\bar{F}(x)} \sim \frac{C_2}{C_1} x^{\alpha_2 - \alpha_1} (1-x)^{\beta_2 - \beta_1}$$
$$\stackrel{x\rightarrow 1}{\rightarrow} \left\{\begin{array}{ll} 0 & \beta_1 < \beta_2 \\ \frac{C_2}{C_1} & \beta_1 = \beta_2 \\ \infty & \beta_1 > \beta_2 \end{array}\right.$$

where:

$$C_i = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}, i = 1, 2.$$

Again, we see that generally c approaches either 0 or ∞, and consequently P*@K approaches either 0 or K. However, for certain parameter settings (e.g. when F and G are tail-equivalent), P*@K does approach a constant.

5 Discussion

TREC VLC participants have reported substantially higher P@20 scores for full collections compared with 10% samples. Our probabilistic analysis sheds some light on this phenomenon. In particular, in the case of a mixture model, if the tail of the distribution of the relevant documents dominates the tail of the distribution of the irrelevant documents (in the sense outlined in Section 4), then P@K will tend to one as the collection size increases. For tail equivalent score distributions, the limiting behavior depends on the specifics of the distributions.

Our work has some limitations. First, our analysis assumes that document scores are independent and identically distributed draws from some distribution (H_m for the mixture model; F and G otherwise). Reality for most text retrieval applications is probably more complex. Second, while Manmatha *et al.* (2001) presented evidence in favor of Gaussian/Exponential model for document scores, work is needed to more fully characterize useful distributional models.

Finally it would be useful to have a general closed-form approximation for P@K and C@K.

Acknowledgements

We thank David D. Lewis and Steve Fienberg for helpful discussions and the reviewers for constructive comments.

References

Hawking, D., Thistlewaite, P., and Harman, D. (1999). Scaling up the TREC Collection. *Information Retrieval*, **1**, 115–137.

Manmatha, R., Feng, F., and Rath, T. (2001). Using models of score distributions in information retrieval. In: *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval*, 267–275.