APPENDIX 5A

FORMULA FOR DOCUMENT RANKING BASED ON PROBABILITY CONSIDERATIONS

by

G.H. STEARMAN

If a particular question at a particular coordination level results in the retrieval of a total of N documents, of which R documents are relevant, then the average time taken to find each of the relevant documents when a large number of searches is made can be determined on the basis of the following assumptions:-

(a) Each successive document is selected at random.

(b) The same time is taken to inspect each document for relevancy so that, for example, if a relevant document is found at the 3rd choice, three units of time are taken and if at the 7th choice, seven units and so on.

If one unit of time is assigned to each choice, then the value of the average as defined above can be taken as the rank of the relevant document in a simulated ordering of the N documents.

Let:-

Total number of documents retrieved be N

Total number of relevant documents be R

Order of N be S $(S = 1, 2 \cdots N)$

Order of R be K $(K = 1, 2 \cdots R)$

Then the problem is to find an expression for $P_{K,S}$, the probability that the Kth relevant document will be found at the Sth inspection, where N and R are given. Then if Q_K is the simulated ranking, its value is given by the weighted sum

$$Q_K = \sum_{S = 1, 2 \cdots N} S.P_{K,S}$$

${}^{A}C_{B}$ means the number of ways of choosing B items from A and is expressed as

$$\frac{A!}{B!\ (A-B)!}$$

The probability $P_{K,S}$ may be determined as the ratio of the number of configurations in which the Kth relevant document appears at the Sth inspection and the total number of ways in which R relevant documents may be arranged in N positions.

The typical layout would then be as shown:-

$$\left[\begin{array}{l}(S-1)\text{ choices including}\\(K-1)\text{ rel. docs.}\end{array}\right]\quad\left[\begin{array}{l}S\text{th choice giving the}\\K\text{th rel. doc.}\end{array}\right]\quad\left[\begin{array}{l}(N-S)\text{ choices including}\\(R-K)\text{ rel. docs.}\end{array}\right]$$

$$\xrightarrow[\text{TIME}]{\qquad\qquad\qquad\qquad}$$

The number of ways in which this layout can be formed is the numerator of the required ratio and is given by

$${}^{(S-1)}C_{(K-1)} \cdot {}^{(N-S)}C_{(R-K)}$$

The denominator is simply $\quad {}^{N}C_{R}$

$$\text{Thus } P_{K,S} = \frac{{}^{(S-1)}C_{(K-1)} \cdot {}^{(N-S)}C_{(R-K)}}{{}^{N}C_{R}}$$

Q_K can now be evaluated as indicated above for each value of K from 1 to R

$$\text{Thus } Q_K = \sum_{S=1}^{N} S \cdot \frac{{}^{(S-1)}C_{(K-1)} \quad {}^{(N-S)}C_{(R-K)}}{{}^{N}C_{R}}$$

$$= \sum_{S=1}^{N} \frac{S(S-1)!}{(K-1)!(S-K)!} \cdot \frac{{}^{(N-S)}C_{(R-K)}}{{}^{N}C_{R}}$$

$$= \frac{K}{{}^{N}C_{R}} \sum_{S=1}^{N} \frac{S!}{K!(S-K)!} \cdot {}^{(N-S)}C_{(R-K)}$$

$$= \frac{K}{{}^{N}C_{R}} \sum_{S=1}^{N} \left[{}^{S}C_{K} \cdot {}^{(N-S)}C_{(R-K)} \right]$$

The terms of the series vanish for $K > S > (K+N-R)$ so that the limits of S may be changed to give

$$Q_K = \frac{K}{{}^{N}C_{R}} \sum_{S=K}^{K+N-R} \left[{}^{S}C_{K} \cdot {}^{(N-S)}C_{(R-K)} \right] \qquad \text{Note that } K \leqslant R$$

The summation of this series is given by Schwatt ["Operations with series" - Chelsea, 2nd edn. page 47] in the form:-

$$\sum_{k=n}^{p-m} {}^{b}C_{n} \cdot {}^{p-k}C_{m} = {}^{p+1}C_{p-n-m}$$

Putting b=S, p=N, m=R-K, n=K we obtain

$$\sum_{S=K}^{N-R+K} {}^{S}C_{K} \cdot {}^{(N-S)}C_{(R-K)} = {}^{(N+1)}C_{(N-R)} = {}^{(N+1)}C_{(N+1-N+R)} = {}^{(N+1)}C_{(R+1)}$$

$$\therefore Q_{K} = \frac{K \cdot {}^{(N+1)}C_{(R+1)}}{{}^{N}C_{R}} = \underline{\underline{\frac{K(N+1)}{(R+1)}}}$$

This simple expression is the basis of the method used in Chapter 5; the above formal analysis shows it to be soundly based upon probability considerations.