REVUE FRANÇAISE D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. SÉRIE VERTE

MARTIN KRAKOWSKI

Brèves communications : « Random file »

Revue française d'informatique et de recherche opérationnelle. Série verte, tome 5, n° V2 (1971), p. 111-112

<http://www.numdam.org/item?id=RO_1971__5_2_111_0>

R.I.R.O.
(5e année, V-2, 1971, p. 111-112)

Brèves communications

RANDOM FILE

par Martin KRAKOWSKI

A file contains N entries which are being requested with *known* relative frequencies Q_i, so that

$$\sum_1^N Q_i = 1. \tag{1}$$

The file is random in the following sense. If the Smith record is the one requested we keep on drawing folders until we succeed in retrieving the Smith entry. However, each time we draw a « non-Smith » record we replace it, shake up the file (in order to erase any memory of ordering resulting from the search) and continue the drawing process until the entry « Smith » is found.

Suppose now that the probabilities P_i of drawing the entries i are under our control. One thinkable arrangement is a roulette wheel where, for instance, $P_2 = 2P_1$ implies that entry # 2 subtends an angle twice that of entry # 1; each record is located in the corresponding angular sector. The roulette wheel is spun until the requested entry shows up.

Another arrangement involves multiple copies of some or of all entries so that, e.g. $P_2 = 2P_1$ means that entry # 2 has twice as many copies as has entry # 1. This method is limited to rational approximations of the P_i and good approximations may require huge files.

The problem is : Given the known relative demand frequencies Q_i, select the discretionary drawing probabilities P_i so as to minimize the expected number of draws (tries) per search of an entry.

Let

$$D_i = 1/P_i = \text{expected \# of tries to retrieve entry } i; \tag{2}$$

$$D = \sum_1^N Q_i/P_i = \text{expected \# of tries per search}; \tag{3}$$

$$\sum_1^N P_i - 1 = 0.$$

In order to minimize D we use the method of Lagrangian multipliers and we obtain (neglecting momentarily the requirement $P_i \geqslant 0$), α being the multiplier,

$$\frac{\partial D}{\partial P_i} = -Q_i/P_i^2 = -\alpha \ , \qquad \text{for each } i. \tag{4}$$

It follows that

$$P_i^2 = Q_i/\alpha \quad \text{and} \quad P_i \sim \sqrt{Q_i}, \tag{5}$$

so that the optimal P_i is proportional to the square root of Q_i.

Therefore,

$$P_i = \sqrt{Q_i}/\sum_1^N \sqrt{Q_j} \tag{6}$$

Thus the requirements $P_i \geqslant 0$ for each i are satisfied; since $\dfrac{\partial^2 D}{\partial P_i^2} > 0$ when (6) holds, the minimum value of (3) is

$$D^* = \sum_1^N Q_i/P_i = \left[\sum_1^N \sqrt{Q_i}\right]^2. \tag{7}$$

When $Q_i = 1/N$ for each i then $P_i = Q_i$ and $D^* = N$.

When $Q_1 = 1$ and $Q_i = 0$ for $i > 1$ then $P_i = 1$ and $P_i = 0$ for $i > 1$, and $D^* = 1$, as expected.

NOTE : The result that the optimal $P_i \sim \sqrt{Q_i}$ is counter-intuitive.

Most people venture the guess $P_i = Q_i$ for the optimal probabilities. The corresponding expected number of tries per search would be $D = N$, the number of file entries, irrespective of the assumed frequencies Q_i. This guess can be very bad. If $Q_1 \to 1$ and $Q_k \to 0$ for $k > 1$, then the optimal $D^* \to 1$, while the intuitive expectation is $D = N$.