PAC Over BitTorrent: From Scratch

April 8, 2013

The probability of finding a document, d, using PAC search in a system with n total documents is:

$$P(d) = 1 - (1 - \frac{r_i}{n})^{z_i} \tag{1}$$

$$z_i \log(1 - \frac{r_i}{n}) = \log(1 - P(d)) \tag{2}$$

where r_i is the document's replication at time i and z_i is the number of nodes queried for query i. We can also approximate using the following:

$$(1 - \frac{r_i}{n})^{z_i} = (1 + \frac{\frac{-r_i z_i}{n}}{z})^{z} \tag{3}$$

$$\approx \exp^{\frac{-r_i z_i}{n}} \tag{4}$$

Combining Equations (1) and (2) gives:

$$P(d) = 1 - \exp^{\frac{-r_i z_i}{n}} \tag{5}$$

$$-r_i z_i = n \log(1 - P(d)) \tag{6}$$

I'm not certain that this approximation is useful given that Equation (1) simplifies to a product that's only slightly more involved. We can fix $P(d)$ at some suitable value and therefore know r_i in terms of z_i alone and visa versa. When applying our BitTorrent extension we have another equation for r_i and z_i:

$$r_i = r_{i-1} + 1 + \frac{z_{i-1}}{P(d)}(1 - \frac{r_{i-1}n}{)} \tag{7}$$

I cannot find a closed form for r_i (i.e. remove the recursion), I thought I had it but I reached a dead end. The recursive form for r_i requires a base value at r_1 and, in fact, choice of this r_1 value dictates all remaining r_i and z_i (for fixed $P(d)$ that is). For each query, i, knowing r_i and z_i requires that you solve the following simultaneous equations:

$$\begin{align}
r_i = r_{i-1} + 1 + \frac{z_{i-1}}{P(d)}(1 - \frac{r_{i-1}n}{)} \tag{8} \\
z_{i-1}\log(1 - \frac{r_{i-1}}{n}) = \log(1 - P(d)) \tag{9}
\end{align}$$

Any solution of which must start with:

$$\begin{align}
r_1 &= c \tag{10} \\
z_1 &= \frac{\log(1 - \frac{c}{n})}{\log(1 - P(d))} \tag{11}
\end{align}$$

So we need to know what value is best for r_1; we can pick a small value and then require a very large z_1 or we can pick a large value and need a smaller z_1. The choice effects r_i and z_i for all i. We could find an optimal value by finding the minimal number of total requests made:

$$\min_{r_1, P(d), s}(r_1 + \sum_{i=1}^{s} \frac{z_i}{P(d)}) \tag{12}$$

where s is the total number of searches performed for the torrent over its lifetime. This might not be the best equation to minimise, we might want to factor in some bandwidth constraints.