Capture Recapture by Mark Levene

Assume the web has W pages and Google has G pages.

Further assume the proportion of G relative to W is:

$$P(G) = G/W \quad (1)$$

So if we can estimate P(G) and G:

$$W = G/P(G) \quad (2)$$

Now consider Bing, another search engine having B pages. Further if we sample from B and G we have, assuming independence:

$$P(G) = P(G \cap B \mid B)$$

Which simplifies to:

$$P(G) = \frac{P(G) \cdot P(B)}{P(B)} \quad (3)$$

So we compute $P(G \cap B \mid B)$ as follows.

Fire a set of carefully chosen queries to Bing and Google, collect the results from both, say QB and QG, and compute their intersection $QB \cap QG$.

So let us put this information together:

$$\frac{QB \cap QG}{QB} \quad (4)$$

is an estimate of $P(G) = P(G \cap B \mid B)$ and therefore:

$$P(G) \approx \frac{QB \cap QG}{QB} \quad (5)$$

where ≈ means approximately.

Plugging (5) into (1) we get:

$$\frac{QB \cap QG}{QB} \approx \frac{G}{W} \quad (6)$$

And so:

$$W \approx \frac{G \cdot QB}{QB \cap QG} \quad (7)$$

recalling that G is an estimate of Google's size.