

Index Size and Web Search Relevance

Gordon Rios

January 22, 2003

Abstract

The effect of index size on web search relevance is discussed and two components of this effect are briefly explored.

The conceptual work for this note was done while I worked as Scientist for Inktomi Corp during the years

from 1999 to 2002. In forming my conclusions I benefited strongly from evidence developed by Dr. Jean Marc Langlois and myself using a large scale relevance testing system. I also benefited from discussions with other senior scientists at Inktomi most notably Prof. Hongyuan Zha, and Dr. Arkady Borkovsky.

1 Index Size and Relevance

In practice when we plot independent measures of relevance for a set of search engines against their index size we see a log shaped curve. This suggests that index size has a positive effect on relevance with diminishing marginal returns. Let's look at index size in more detail by decomposing it into two separate effects.

1.1 Content Effect

The effects of index size on the content side of the relevance equation can be seen by considering the case of finding a single document containing word w . Suppose that each document in the index has a probability p of containing the word w .

Let's look at the probability that the index contains *at least one* document containing w :

$$\begin{aligned} P_w(n) &= 1 - (1 - p)^n \\ &= 1 - q^n \end{aligned}$$

where n is the number of documents in the index and q is the probability of a document not containing w .

It's easy to see that $P_w(n)$ is increasing in n but at a decreasing rate:

$$\frac{d}{dn} P_w(n) = \frac{d}{dn} (1 - q^n) = -q^n \ln q > 0$$

$$\frac{d^2}{dn^2} P_w(n) = \frac{d^2}{dn^2} (1 - q^n) = -q^n \ln^2 q < 0$$

where the inequalities follow because $q < 1$.

In practice, when we plot independent relevance measures for a set of search engines against index size we see a concave curve with the above characteristics.

1.2 Ranking Effect

In the case where there are many documents containing a given search term (e.g. the term has some degree of popularity) and relevance varies across those documents then index size can reduce the effectiveness of an engine's *ranking power*.

Let's agree to measure the *ranking power* of a search engine by the rank correlation between the ranking assigned by human judges and that of the engine. Testing even the best search engines will show a maximum rank correlation in the neighborhood of 0.25 – in terms of statistical variance explained that's 6.25%. With this level of noise we can expect lot's errors as index size forces us to rank increasing numbers of documents.

One immediate conclusion is that for queries yielding lots of documents (e.g. popular queries) it's extremely important to find additional ranking factors such as click popularity, anchortext, etc. Whereas for more specific rare queries index size is absolutely critical to improve the chances of returning any useful documents at all.