Is Compression A Google SEO Myth? via @sejournal, @martinibuster

I recently came across an SEO test that attempted to verify whether compression ratio affects rankings. It seems there may be some who believe that higher compression ratios correlate with lower rankings. Understanding compressibility in the context of SEO requires reading both the original source on compression ratios and the research paper itself before drawing conclusions about whether or not it's an SEO myth.

Search Engines Compress Web Pages

Compressibility, in the context of search engines, refers to how much web pages can be compressed. Shrinking a document into a zip file is an example of compression. Search engines compress indexed web pages because it saves space and results in faster processing. It's something that all search engines do.

Websites & Host Providers Compress Web Pages

Web page compression is a good thing because it helps search crawlers quickly access web pages which in turn sends the signal to Googlebot that it won't strain the server and it's okay to grab even more pages for indexing.

Compression speeds up websites, providing site visitors a high quality user experience. Most web hosts automatically enable

compression because it’s good for websites, site visitors and also good for web hosts because it saves on bandwidth loads. Everybody wins with website compression.

High Levels Of Compression Correlate With Spam

Researchers at a search engine discovered that highly compressible web pages correlated with low-quality content. The study called *Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages* (PDF) was conducted in 2006 by two of the world’s leading researchers, Marc Najork and Dennis Fetterly.

Najork currently works at DeepMind as Distinguished Research Scientist. Fetterly, a software engineer at Google, is an author of many important research papers related to search, content analysis and other related topics. This research paper isn’t just any research paper, it’s an important one.

What the 2006 research paper shows is that 70% of web pages that compress at a level of 4.0 or higher tended to be low quality pages with a high level of redundant word usage. The average compression level of sites was around 2.0.

Here are the averages of normal web pages listed by the research paper:

- **Compression ratio of 2.0:**
  The most frequently occurring compression ratio in the dataset is 2.0.
- **Compression ratio of 2.1:**
  Half of the pages have a compression ratio below 2.1, and half have a compression ratio above it.
- **Compression ratio of 2.11:**
  On average, the compression ratio of the pages analyzed is 2.11.

It would be an easy first-pass way to filter out the obvious content spam so it makes sense that they would do that to weed out heavy-handed content spam. But weeding out spam is more complicated than simple solutions. Search engines use multiple signals because it results in a higher level of accuracy.

The researchers from 2006 reported that 70% of sites with a compression level of 4.0 or higher were spam. That means that the other 30% were not spam sites. There are always outliers in statistics and that 30% of non-spam sites is why search engines tend to use more than one signal.

Do Search Engines Use Compressibility?

It's reasonable to assume that search engines use compressibility to identify heavy handed obvious spam. But it's also reasonable to assume that if search engines employ it they are using it together with other signals in order to increase the accuracy of the metrics. Nobody knows for certain if Google uses compressibility.

Impossible To Determine If Google's Using Compression

This article is about the fact that there is no way to prove that a compression ratio is an SEO myth or not.

Here's why:

1. If a site triggered the 4.0 compression ratio plus the other spam signals, what would happen is that those sites wouldn't be in the search results.

2. If those sites are not in the search results, there is no way to test the search results to see if Google is using compression ratio as a spam signal.

It would be reasonable to assume that the sites with high 4.0 compression ratios were removed. But we don't *know* that, it's not a certainty. So we can't prove that they were removed.

The only thing we do know is that there is this research paper out there that's authored by distinguished scientists.

Is Compressibility An SEO Myth?

Compressibility may not be an SEO myth. But it's probably not anything publishers or SEOs should be worry about as long as they're avoiding heavy-handed tactics like keyword stuffing or repetitive cookie cutter pages.

Google uses de-duplication which removes duplicate pages from their index and consolidates the PageRank signals to whichever page they choose to be the canonical page (if they choose one). Publishing duplicate pages will likely not trigger any kind of penalty, including anything related to compression ratios, because, as was already mentioned, search engines don't use signals in isolation.

https://www.searchenginejournal.com/is-googles-use-of-compressibility-an-seo-myth/541771/