

Compressibility – A Search Engine Trick for Thin Content

by @martinibuster

Compressibility is an interesting spam fighting trick from a long time ago. Many SEOs have never heard of it but it's worth knowing about. Compressibility refers to a way of identifying low quality content. The funny thing about compressibility is that search engineers discovered it by accident.

What I am about to describe may or may not be in use by a search engine. Yet it is still useful to understand. Knowing about compressibility could be useful for content planning and diagnosing why certain content might be considered thin.

Background on Compression

Search engines “compress” web page information so that they can fit more data on their hard drives. Ever shrink a file folder by turning it into a zip file? That's what compression is.

WinZip and GZip are compression algorithms. What they do is toss out repetitive data and replace them with code that represents that discarded information. That's how you get a smaller file size.

What search engineers noticed was that some web pages compressed at a higher level than other web pages. When they examined the web pages that compressed at a higher level they discovered that web pages with high compression ratios tended to have a lot of repetitive content.

When they looked closer they discovered that 70% of the high compression web pages were spam. They were thin pages that contained a lot of repetitive content. I am not saying that's the origin of the phrase, *thin pages*. But when you compress certain kinds of spam pages, that's what you're left with, thin pages.

Thin Pages Origins in Original Content

What SEOs were doing many years ago was an attempt to create original content. They used sets of unique paragraphs with blank spaces to add data like *City* and *State* information. A set of paragraphs were meant for the top of the page, another group of paragraphs were for the middle of the page and another set for the bottom of the page.

By randomly mixing and matching the paragraphs, every page was 100% unique. With enough paragraphs in each set, you could get a near infinite amount of page combinations. This technique was perfect for generating hundreds of thousands of pages to rank for city/state keyword combinations.

This technique worked for a long while!

Compression Redefines Unique Content

But compression is able to defeat that kind of content. Although the spammers could create twenty, forty, or more unique paragraphs for each set, the resulting web pages would still compress at a high ratio.

I don't know if search engines use compression for identifying thin content today. But it's a simple way to identify thin low value add content. Combine compression with other signals and

finding thin content pages becomes even easier.

Documentation of Compression

I first heard of compression in a research paper from 2006 titled, Detecting Spam Web Pages through Content Analysis. It's a Microsoft research paper investigating techniques for identifying spam by relying solely on content features. This was during the heyday of statistical analysis algorithms.

Here is a quote from the relevant section of that research paper:

"4.6 Compressibility

We measure the redundancy of web pages by the compression ratio, the size of the uncompressed page divided by the size of the compressed page.

The line graph, depicting the prevalence of spam, rises steadily towards the right of the graph. The graph gets quite noisy beyond a compression ratio of 4.0 due to a small number of sampled pages per range. However, in aggregate, 70% of all sampled pages with a compression ratio of at least 4.0 were judged to be spam."

Takeaway: How Compressibility is Useful Today

Compressibility is a useful thing to know because it gives you insight on why certain web pages may not be performing well. It may have been used by search engines back in the caveman days of spam fighting and SEO. It could still be useful today whether search engines use it or not.

If your website content compresses by a factor of four, then it may be useful to take a look at that content to be sure it is truly original and not redundant. It doesn't matter if

compression is used by search algorithms. It's still a useful thing to know about.

Images by Shutterstock, Modified by Author

Screenshots by Author

<http://tracking.feedpress.it/link/13962/9057436>