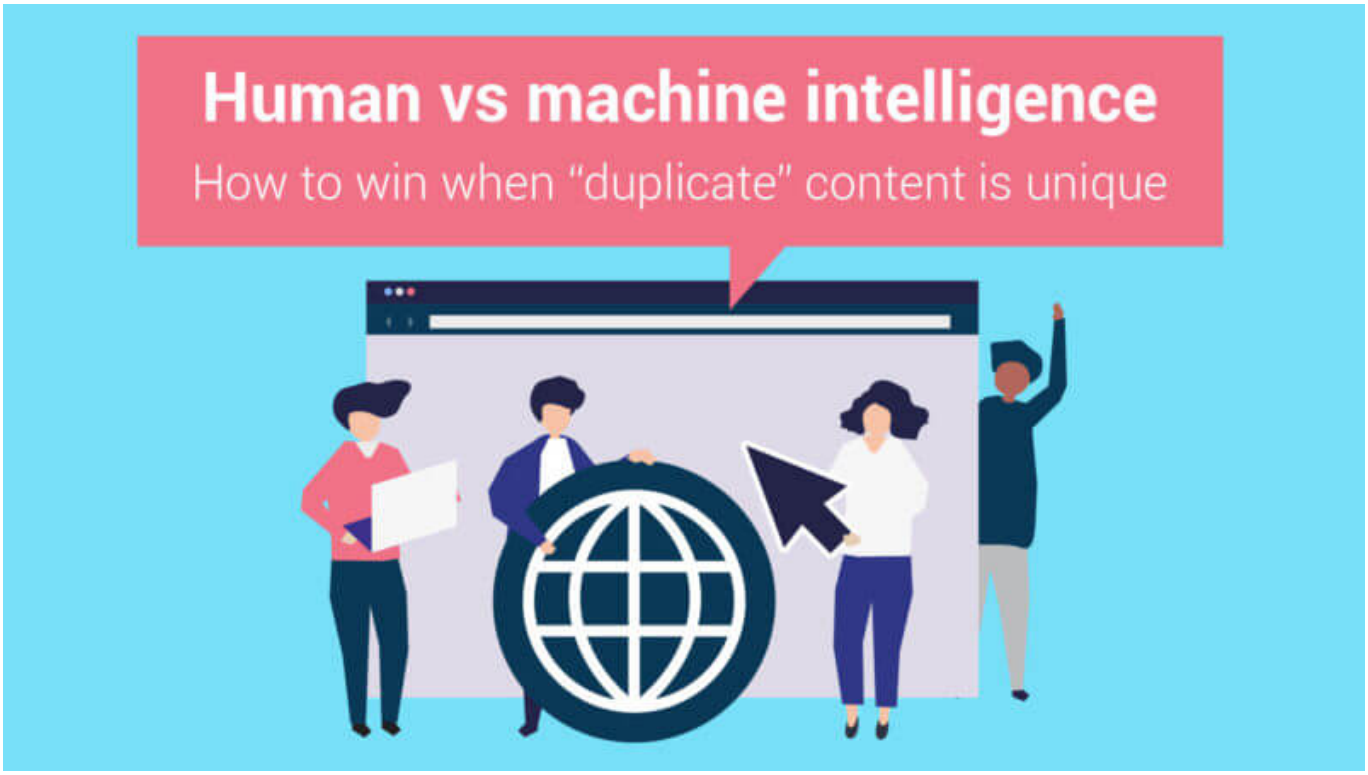
Human vs machine intelligence: how to win when ‘duplicate’ content is unique



As impressive as machine learning and algorithm-based intelligence can be, they often lack something that comes naturally to humans: common sense.

It’s common knowledge that putting the same content on multiple pages produces duplicate content. But what if you create pages about similar things, with differences that matter? Algorithms flag them as duplicates, though humans have no problem telling pages like these apart:

- E-commerce: similar products with multiple variants or critical differences
- Travel: hotel branches, destination packages with similar content
- Classifieds: exhaustive listings for identical items
- Business: pages for local branches offering the same

services in different regions

How does this happen? How can you spot issues? What can do you about it?

The danger of duplicate content

Duplicate content interferes with your ability to make your site visible to search users through:

- Loss of ranking for unique pages that unintentionally compete for the same keywords
- Inability to rank pages in a cluster because Google chose one page as a canonical
- Loss of site authority for large quantities of thin content

How machines identify duplicate content

Google uses algorithms to determine whether two pages or parts of pages are duplicate content, which Google defines as content that is “appreciably similar“.

Google’s similarity detection is based on their patented Simhash algorithm, which analyzes blocks of content on a web page. It then calculates a unique identifier for each block, and composes a hash, or “fingerprint”, for each page.

Because the number of webpages is colossal, scalability is key. Currently, Simhash is the only feasible method for finding duplicate content at scale.

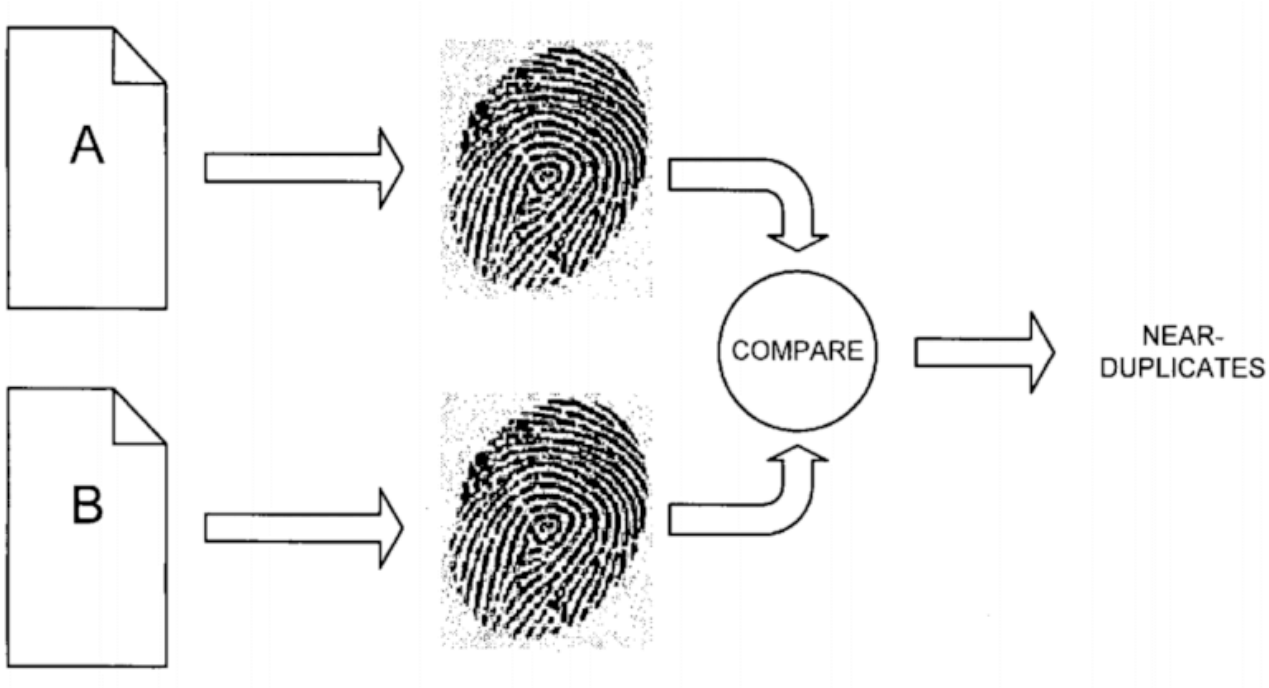
Simhash fingerprints are:

- Inexpensive to calculate. They are established in a single crawl of the page.
- Easy to compare, thanks to their fixed length.

- Able to find near-duplicates. They equate minor changes on a page with minor changes in the hash, unlike many other algorithms.

This last means that the difference between any two fingerprints can be measured algorithmically and expressed as a percentage. To reduce the cost of evaluating every single pair of pages, Google employs techniques such as:

- Clustering: by grouping sets of sufficiently similar pages together, only fingerprints within a cluster need to be compared, since everything else is already classified as different.
- Estimations: for exceptionally large clusters, an average similarity is applied after a certain number of fingerprint pairs are calculated.



Comparing page fingerprints. Source: Near-duplicate document detection for web crawling (Google patent)

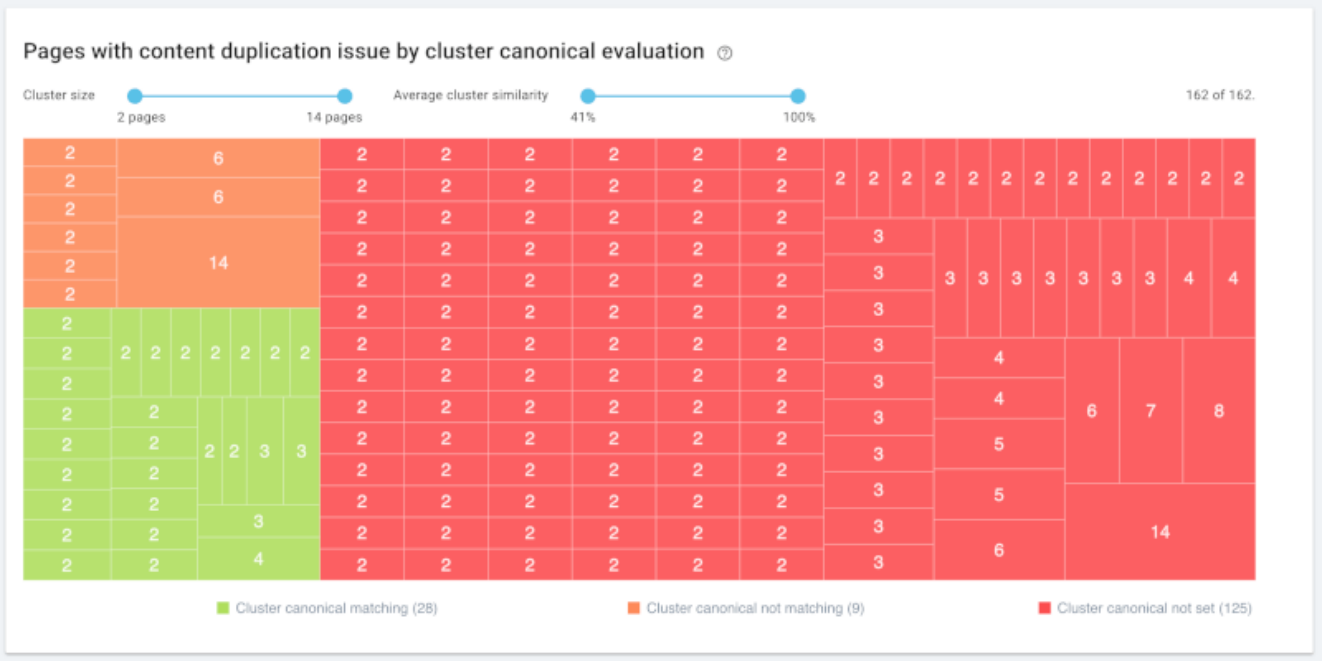
Finally, Google uses a weighted similarity rate that excludes certain blocks of identical content (boilerplate: header, navigation, sidebars, footer; disclaimers…). It takes into account the subject of the page using n-gram analysis to

determine which words on the page occur most frequently, and – in the context of the site – are most important.

Analyzing duplicate content with Simhash

We'll be looking at a map of content clusters flagged as similar using Simhash. This chart from OnCrawl overlays an analysis of your duplicate content strategy on clusters of duplicate content.

OnCrawl's content analysis also includes similarity ratios, content clusters, and n-gram analysis. OnCrawl is also working on an experimental heatmap indicating similarity per content block that can be overlaid on a webpage.



Mapping a website by content similarity. Each block represents a cluster of similar content. Colors indicate the coherence of the canonicalization strategy for each cluster. Source: OnCrawl.

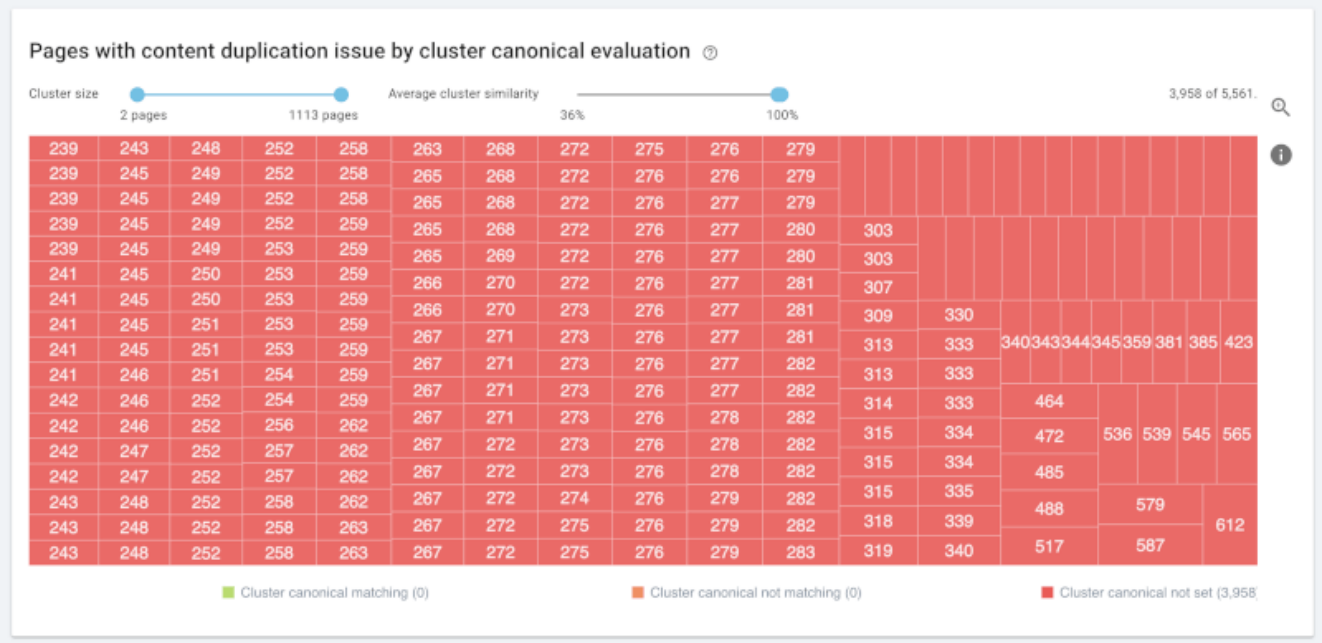
Validating clusters with canonicals

Using canonical URLs to indicate the main page in a group of similar pages is a way of intentionally clustering pages. Ideally, the clusters created by canonicals and those established by Simhash should be identical.



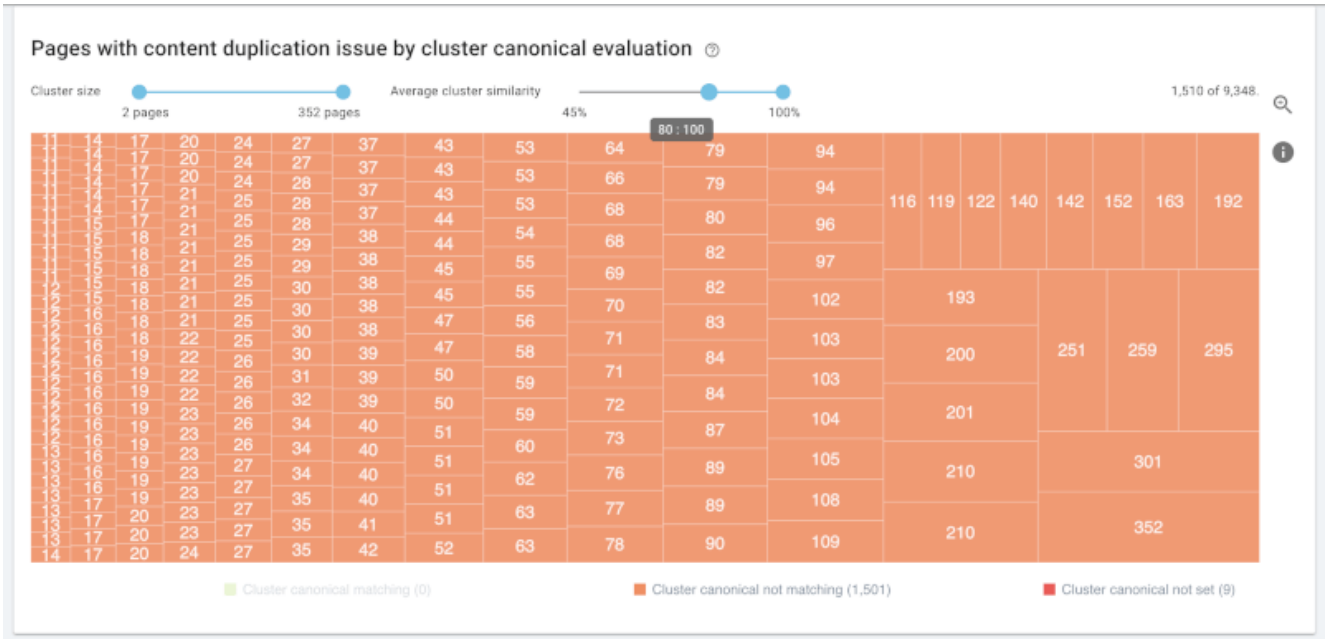
Canonical clusters matching similarity clusters (in green). Highlighted: 6 pages that are 100% similar. Your canonical policy and Google’s Simhash analysis treat them in the same way.

When this isn’t the case, it’s often because there is no canonical policy in place on your website:



No canonical declarations: clusters of hundreds of pages each, with an average similarity rate of 99-100%. Google may impose canonical URLs. You have no control over which pages will rank and which won't.

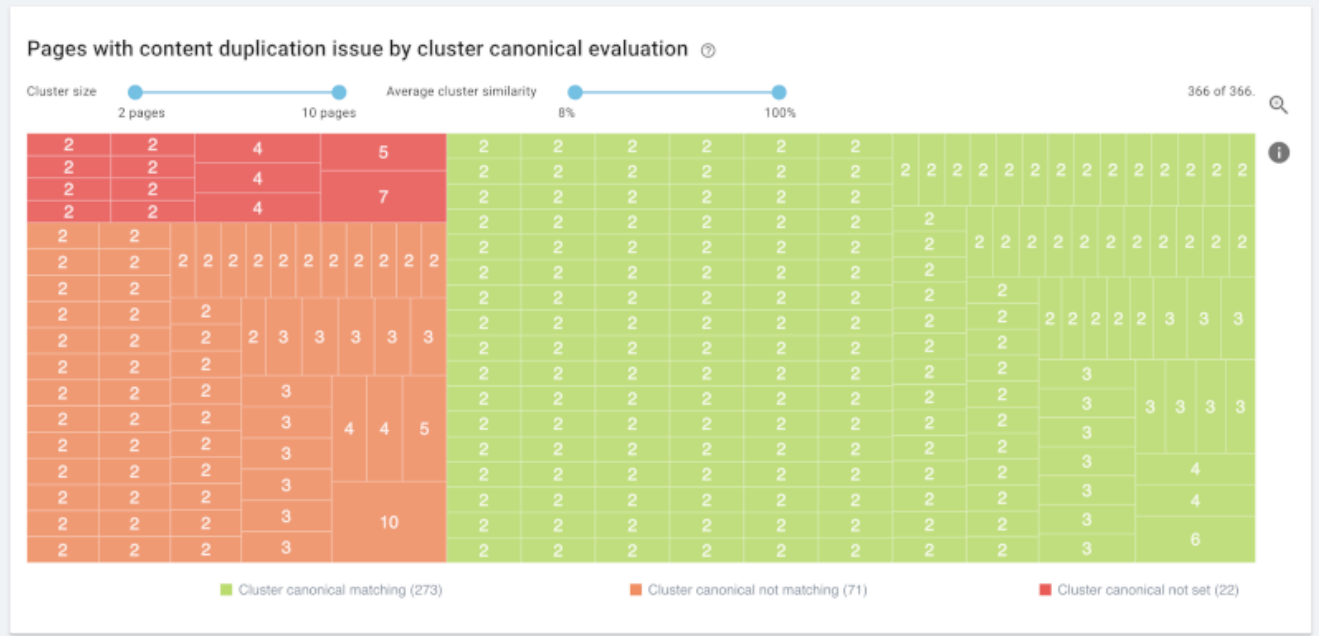
Or because there are conflicts between your canonical strategy and the methods Google uses to group similar content:



Problems with canonicals: large clusters with over 80% similarity and multiple canonical URLs per cluster. Google will either impose its own canonical URLs, or index duplicate

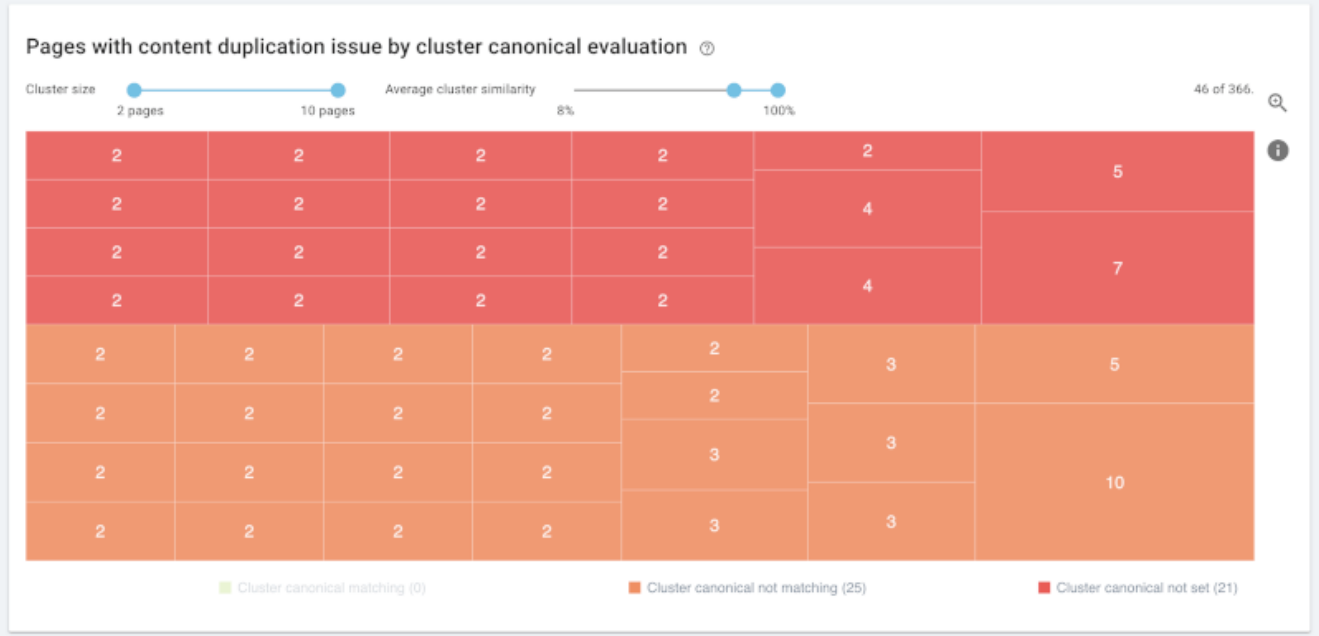
pages you wanted to keep out of the index.

Your site's clusters don't look like the ones above. You've already followed best practices for duplicate content. URLs that contain the same content — such as printable/mobile versions, or alternate URLs generated by a CMS — declare the correct canonical URL.



Mapping similarity clusters after canonicalization.

Filter out the duplicate content that is correctly handled by your canonical strategy. The remaining non-canonicalized URLs are pages you want to rank.



The previous mapping, after removing validated (green) clusters and clusters with less than 80% similarity. Most of the remaining 46 clusters only have 2 pages.

URLs that still appear in clusters based on Simhash and semantic analysis are URLs you and Google disagree on.

Solving duplicate content problems for unique content

There's no satisfying trick to correct a machine's view of unique pages that appear to be duplicates: we can't change how Google identifies duplicate content. However, there are still solutions to align your perception of unique content and Google's… while still ranking for the keywords you need.

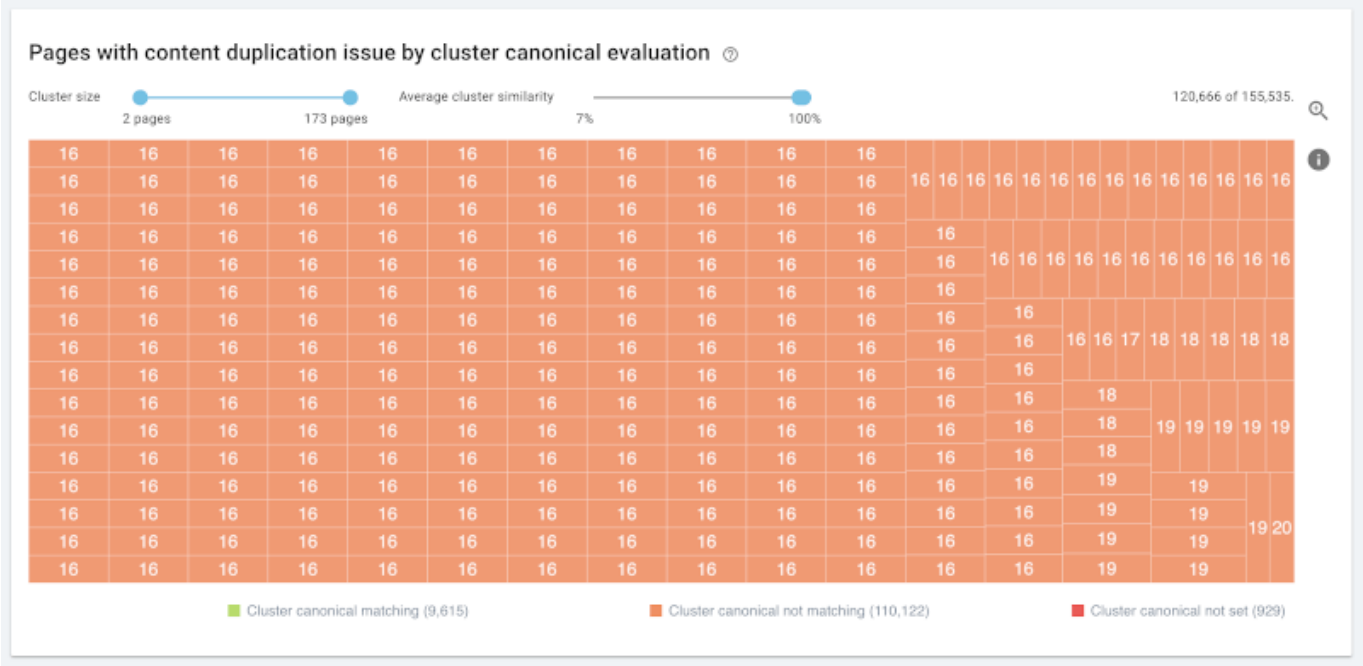
Here are five strategies to adapt to your site.

Resolve edge cases

Start by looking at the edge cases: clusters with very low or very high similarity rates.



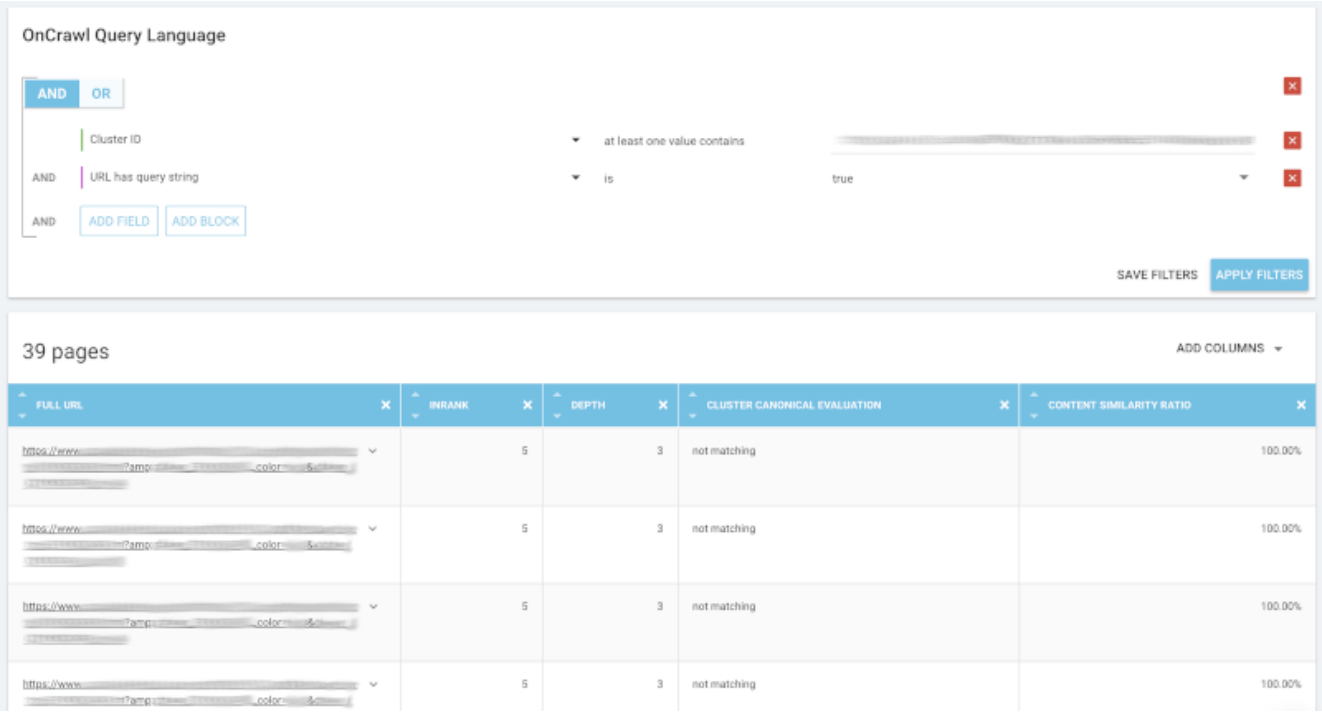
- Under 20% similarity: similar, but not too similar. You can signal Google to treat them as different pages by linking between the pages in the cluster, using distinct anchor text for each page.



- Maximum similarity: find the underlying issue. You will need to either enrich the content to differentiate the pages or merge the pages into one.

Reduce the number of facets

If your duplicate pages are related to facets, you may have an indexing issue. Maintain the facets that already rank, and limit the number of facets you allow Google to index.



Cluster composed of identical pages based on sortable facets. Source: OnCrawl.

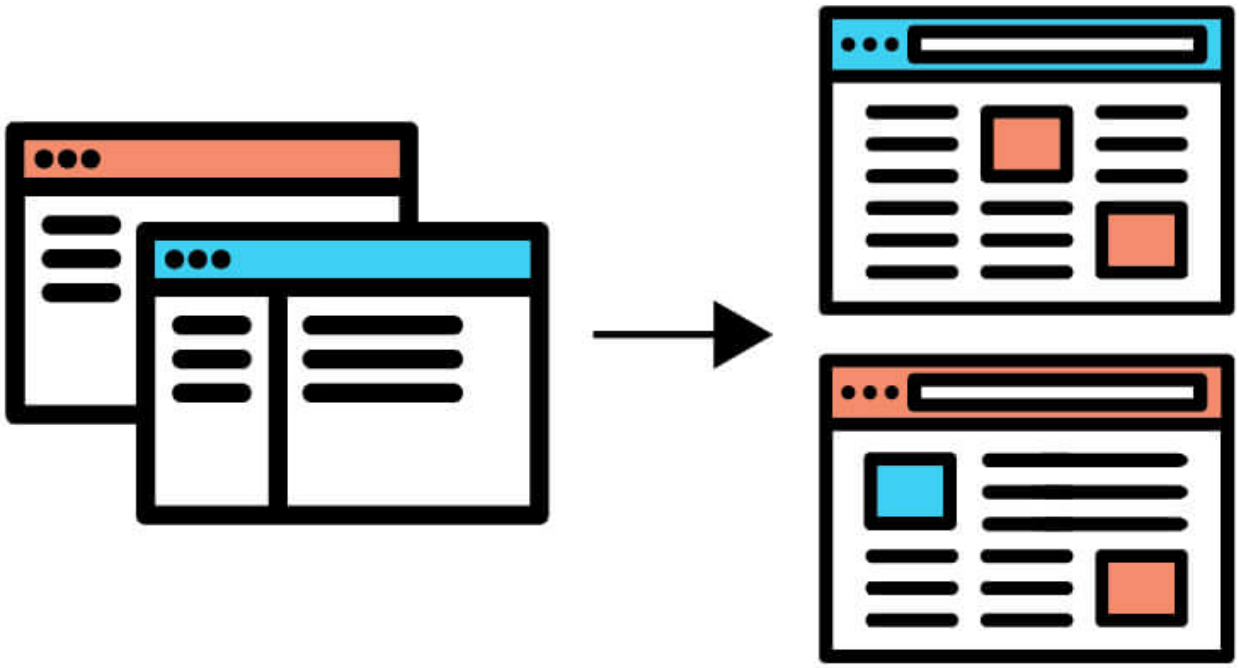
Make pages (more) unique

Remember: minor differences in content create minor differences in Simhash fingerprints. You need to make significant changes to the content on the page rather than small adjustments.

Enrich page content:

- Add text content to the pages.
    - Add different descriptions of images.
    - Include full customer reviews (If the reviews apply to multiple pages, merge the pages!).
    - Add additional information.

        - Add related information.
    - Use different images.
    - Test using very different anchor text for links to the different pages.
    - Reduce the amount of source code in common between the similar pages.
    - Improve semantic density on the pages.
        - Increase the vocabulary related to the subject and reduce the filler.



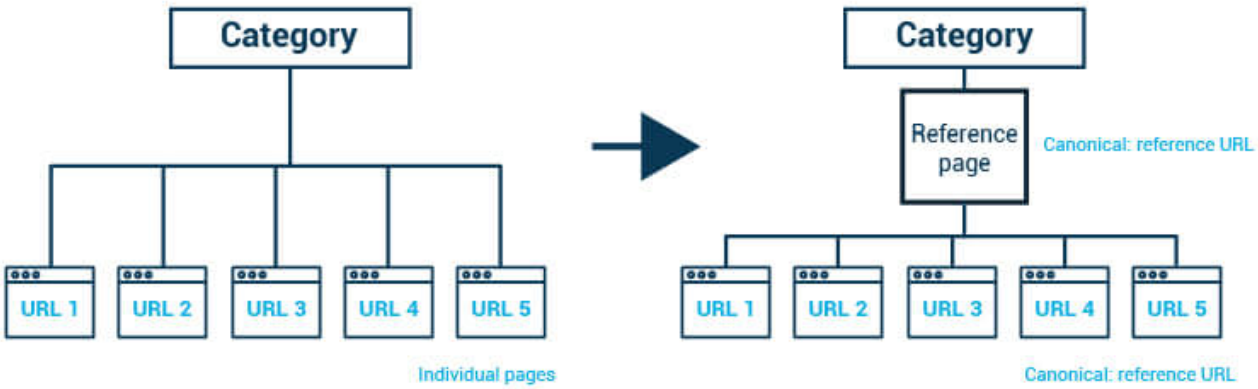
Create ranking reference pages

If enriching your pages isn’t possible or appropriate, consider creating a single reference page that ranks in place of all the “duplicate” pages. This strategy uses the same principle as content hubs to promote a main page for multiple keywords. It’s particularly useful when you have multiple versions of a product that you need to maintain as separate pages.

This strategy can be used to create pages targeting a need or a seasonal opportunity. It can improve families of pages by

providing stronger semantics and rankings.

It can also benefit classifieds websites, job offer sites, and other sites with many, often-similar listings. Reference pages should group listings by a single characteristic; location (city) is often used successfully.



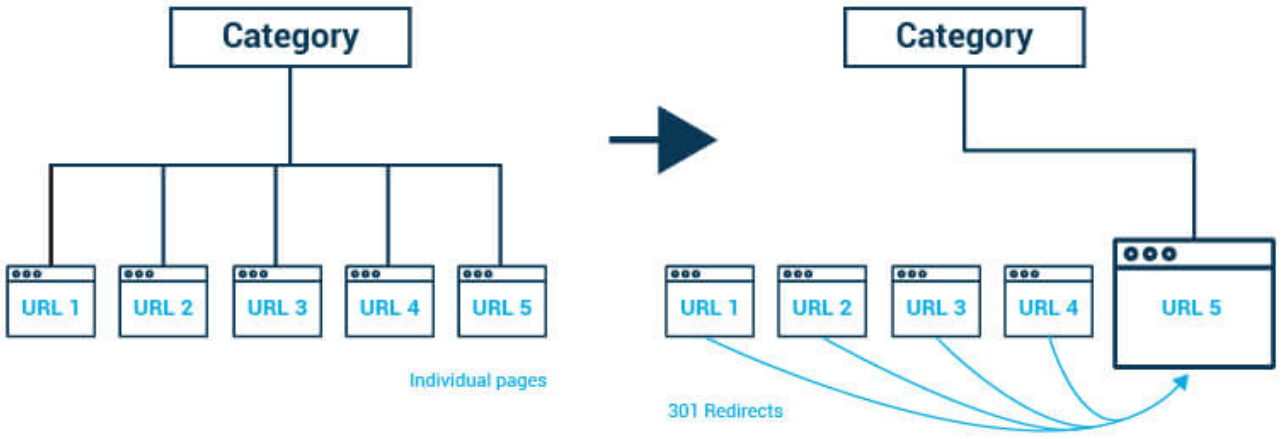
What to do:

1. Create a reference page that brings together the semantic content of all of the “duplicate” product pages. It should promote all of the keywords you want to use and link to all of the “duplicate” pages.
2. Set the canonical URL for each “duplicate” page to the reference page, and the reference page’s canonical URL as itself.
3. Link between the “duplicate” pages.
4. Optimize site navigation to promote the reference page.

Strengthened by links from the “duplicate” pages, canonical declarations, and combined content, reference pages are easy to rank.

Combine your pages

You keep trying to enrich pages with the same content? You can’t explain why you want to keep them all? It may be time to combine them.



If you decide to combine your pages into one:

- Keep the URL that performs the best.
- Redirect (301) pages that you’re getting rid of to the one you’re keeping.
- Add content from the pages you’re getting rid of to the page you’re keeping and optimize it to rank for all of the cluster’s keywords.

The future of duplicate content

Google’s ability to understand the content of a page is constantly evolving. With the increasingly precise ability to identify boilerplate and to differentiate between intent on web pages, unique content identified as a duplicate should eventually become a thing of the past.

Until then, understanding why your content looks like

duplicates to Google, and adapting it to convince Google otherwise, are the keys to successful SEO for similar pages.

About The Author



OnCrawl is an award-winning technical SEO platform that helps you make smarter SEO decisions. OnCrawl combines your content, log files and search data at scale so that you can open Google's blackbox and build a SEO strategy with confidence. Backed by a SEO crawler, a log analyzer and third-party integrations, OnCrawl currently works with over 800 clients in 66 countries including e-commerce websites, online publishers and agencies. OnCrawl produces actionable dashboards and reports to support your entire search engine optimization process and helps you improve your positions, traffic and revenues. Learn more about us at https://www.oncrawl.com.

http://feeds.searchengineland.com/~r/searchengineland/~3/AIW8Z5oDScY/human-vs-machine-intelligence-how-to-win-when-duplicate-content-is-unique-309098