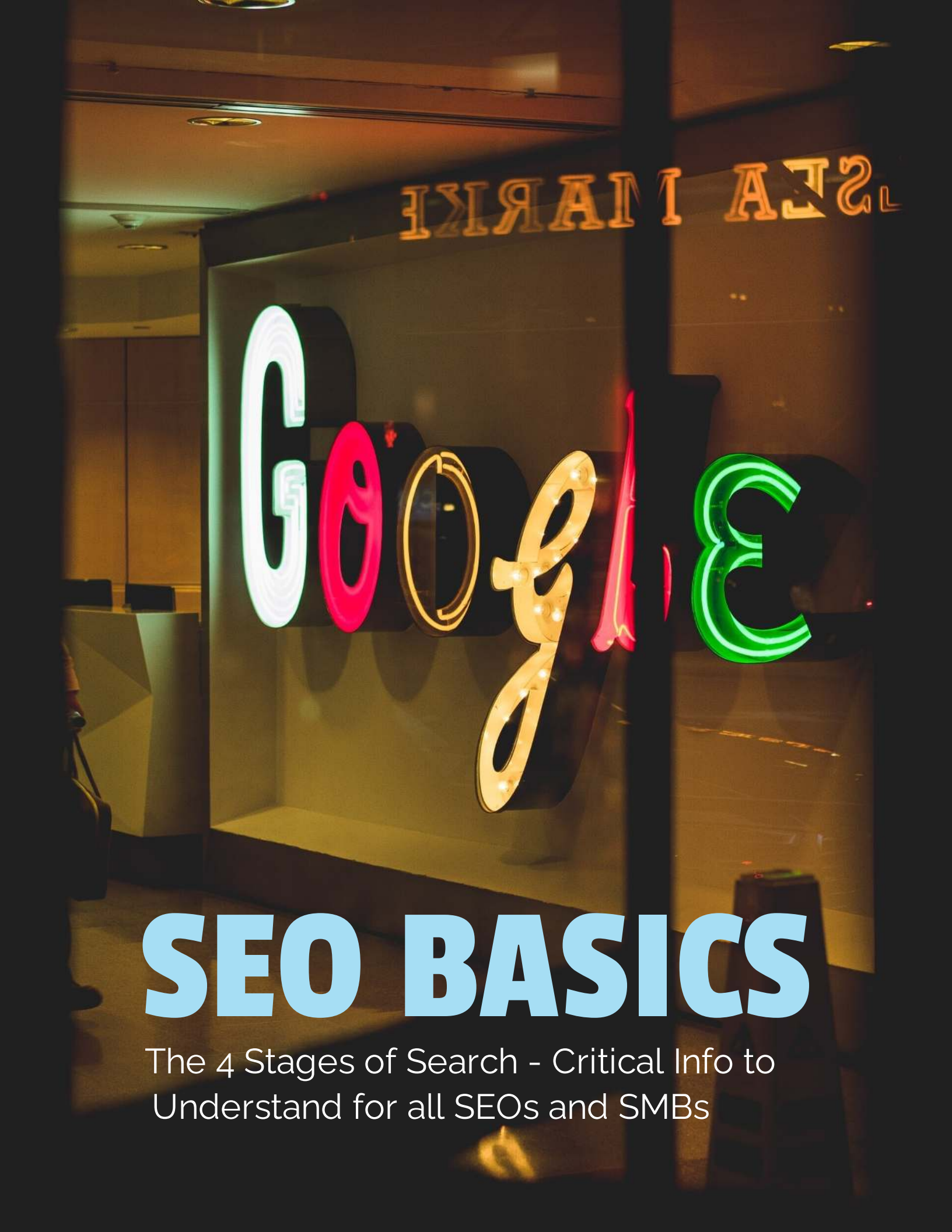


SEO BASICS

The 4 Stages of Search - Critical Info to Understand for all SEOs and SMBs

Table of Contents

Knowing how search engines work can help you adjust your website and increase your rankings and traffic.

Have you ever wondered, "What's the difference between crawling, rendering, indexing, and ranking?" If not, you should but after 1000's of strategic meetings with SEOs and agency owners, the process of the 4 stages of search is still widely misunderstood. For example, the question below is one we often get from users of the Signal Genesys platform and you can see, the answer from our Knowledgebase involves helping the user understand the 4 stages of search and why you don't have to see a page/URL in the search results in order for that page to have been crawled, rendered and indexed… meaning the signals were generated and received by the bot.

As foundational as it is and may seem, this common question demonstrates that it isn't uncommon for some SEOs and agency owners to confuse the basic stages of search and or get the process entirely wrong due to being misinformed. In this brief primer, you'll get a refresher on how search engines work and as we go over each stage of the process involved in rendering live search results. This is critical info to understand if you offer SEO or Local SEO to business clients.

How do search engines actually work? Well, there are so many websites and so much content on the internet that search engines need to use a system through which they can comb through different websites, and discover and understand the content on them. This is achieved using automated software known as **spiders**, **crawlers**, or **bots**.

These terms can be quite interchangeable and to a wide extent, they do a similar job. But they do have marginally different functions:

- **Spider**: A spider is a program run by a search engine to build a summary of a website's content. Spiders create a text-based summary of content and an address (URL) for each webpage.
- **Crawler**: A crawler visits each webpage on a website and determines all of the hyperlinks on every individual page.
- **Robot**: A robot is an automated computer program that visits websites and performs predefined tasks. Its job is to understand how to crawl and index pages on a given website.

Many different processes are involved in bringing the web's content into your search results. In some ways, it can be a gross oversimplification to say there are only a handful of discrete stages to make it happen.

Each of the four stages I cover here has several subprocesses that can occur within them.

Even beyond that, there are significant processes that can be asynchronous to these, such as:

- Types of spam policing.
- Incorporation of elements into the Knowledge Graph and updating of knowledge panels with the information.
- Processing of optical character recognition in images.
- Audio-to-text processing in audio and video files.
- Assessing and application of PageSpeed data.
- And more.

What follows are the primary stages of search required for getting web pages to appear in the search results.

Crawling

Crawling occurs when a search engine requests web pages from websites' servers.

Imagine that Google and Microsoft Bing are sitting at a computer, typing in or clicking on a link to a webpage in their browser window.

Thus, the search engines' machines visit web pages similar to how you do. Each time the search engine visits a webpage, it collects a copy of that page and notes all the links found on that page. After the search engine collects that webpage, it will visit the next link in its list of links yet to be visited.

This is referred to as "crawling" or "spidering" which is apt since the web is metaphorically a giant, virtual web of interconnected links.

The data-gathering programs used by search engines are called "spiders," "bots" or "crawlers."

Google's primary crawling program is "Googlebot" is, while Microsoft Bing has "Bingbot." Each has other specialized bots for visiting ads (i.e., GoogleAdsBot and AdIdxBot), mobile pages, and more.

This stage of the search engines' processing of web pages seems straightforward, but there is a lot of complexity in what goes on, just in this stage alone.

Think about how many web server systems there can be, running different operating systems of different versions, along with varying content management systems (i.e., WordPress, Wix, Squarespace), and then each website's unique customizations.

Many issues can keep search engines' crawlers from crawling pages, which is an excellent reason to study the details involved in this stage.

First, the search engine must find a link to the page at some point before it can request the page and visit it. (Under certain configurations, the search engines have been known to suspect there could be other, undisclosed links, such as one step up in the link hierarchy at a subdirectory level or via some limited website internal search forms.)

Search engines can discover webpages' links through the following methods:

- When a website operator submits the link directly or discloses a sitemap to the search engine.
- When other websites link to the page.
- Through links to the page from within its own website, assuming the website already has some pages indexed.
- Social media posts.
- Links found in documents.
- URLs found in written text and not hyperlinked.
- Via the metadata of various kinds of files.
- And more.

Robots.txt

Robots.txt files are located in the root directory of websites (ex. yourdomain.com/robots.txt) and suggest which parts of your site search engines should and shouldn't crawl, as well as the speed at which they crawl your site, via specific robots.txt directives.

How Googlebot treats robots.txt files

- If Googlebot can't find a robots.txt file for a site, it proceeds to crawl the site.
- If Googlebot finds a robots.txt file for a site, it will usually abide by the suggestions and proceed to crawl the site.
- If Googlebot encounters an error while trying to access a site's robots.txt file and can't determine if one exists or not, it won't crawl the site.

In some instances, a website will instruct the search engines not to crawl one or more web pages through its robots.txt file, which is located at the base level of the domain and web server.

Robots.txt files can contain multiple directives within them, instructing search engines that the website disallows crawling of specific pages, subdirectories or the entire website.

Instructing search engines not to crawl a page or section of a website does not mean that those pages cannot appear in search results. Keeping them from being crawled in this way can severely impact their ability to rank well for their keywords.

In yet other cases, search engines can struggle to crawl a website if the site automatically blocks the bots. This can happen when the website's systems have detected that:

- The bot is requesting more pages within a time period than a human could.
- The bot requests multiple pages simultaneously.
- A bot's server IP address is geolocated within a zone that the website has been configured to exclude.
- The bot's requests and/or other users' requests for pages overwhelm the server's resources, causing the serving of pages to slow down or error out.

However, search engine bots are programmed to automatically change delay rates between requests when they detect that the server is struggling to keep up with demand.

For larger websites and websites with frequently changing content on their pages, "crawl budget" can become a factor in whether search bots will get around to crawling all of the pages.

Essentially, the web is something of an infinite space of web pages with varying update frequency. The search engines might not get around to visiting every single page out there, so they prioritize the pages they will crawl.

Websites with huge numbers of pages, or that are slower responding might use up their available crawl budget before having all of their pages crawled if they have relatively lower ranking weight compared with other websites.

It is useful to mention that search engines also request all the files that go into composing the webpage as well, such as images, CSS and JavaScript.

Just as with the web page itself, if the additional resources that contribute to composing the webpage are inaccessible to the search engine, it can affect how the search engine interprets the webpage.

Tell search engines how to crawl your site

If you used Google Search Console or the "site:domain.com" advanced search operator and found that some of your important pages are missing from the index and/or some of your unimportant pages have been mistakenly indexed, there are some optimizations you can implement to better direct Googlebot how you want your web content crawled. Telling search engines how to crawl your site can give you better control of what ends up in the index.

Rendering

When the search engine crawls a webpage, it will then "render" the page. This involves taking the HTML, JavaScript and cascading stylesheet (CSS) information to generate how the page will appear to desktop and/or mobile users.

This is important in order for the search engine to be able to understand how the webpage content is displayed in context. Processing the JavaScript helps ensure they may have all the content that a human user would see when visiting the page.

The search engines categorize the rendering step as a subprocess within the crawling stage. I listed it here as a separate step in the process because fetching a webpage and then parsing the content in order to understand how it would appear composed in a browser are two distinct processes.

Google uses the same rendering engine used by the Google Chrome browser, called "Rendertron" which is built off the open-source Chromium browser system.

Bingbot uses Microsoft Edge as its engine to run JavaScript and render webpages. It's also now built upon the Chromium-based browser, so it essentially renders webpages very equivalently to the way that Googlebot does.

Google stores copies of the pages in their repository in a compressed format. It seems likely that Microsoft Bing does so as well (but I have not found documentation confirming this). Some search engines may store a shorthand version of web pages in terms of just the visible text, stripped of all the formatting.

Rendering mostly becomes an issue in SEO for pages that have key portions of content dependent upon JavaScript/AJAX.

Both Google and Microsoft Bing will execute JavaScript in order to see all the content on the page, and more complex JavaScript constructs can be challenging for the search engines to operate.

I have seen JavaScript-constructed webpages that were essentially invisible to the search engines, resulting in severely nonoptimal webpages that would not be able to rank for their search terms.

I have also seen instances where infinite-scrolling category pages on e-commerce websites did not perform well on search engines because the search engine could not see as many of the products' links.

Other conditions can also interfere with rendering. For instance, when there is one or more JaveScript or CSS files inaccessible to the search engine bots due to being in subdirectories disallowed by robots.txt, it will be impossible to fully process the page.

Googlebot and Bingbot largely will not index pages that require cookies. Pages that conditionally deliver some key elements based on cookies might also not get rendered fully or properly.

Indexing

Once a page has been crawled and rendered, the search engines further process the page to determine if it will be stored in an index or not, and to understand what the page is about.

The search engine index is functionally similar to an index of words found at the end of a book. There is more than one index. Read that again.

A book's index will list all the important words and topics found in the book, listing each word alphabetically, along with a list of the page numbers where the words/topics will be found.

A search engine index contains many keywords and keyword sequences, associated with a list of all the web pages where the keywords are found.

The index bears some conceptual resemblance to a database lookup table, which may have originally been the structure used for search engines. But the major search engines likely now use something a couple of generations more sophisticated to accomplish the purpose of looking up a keyword and returning all the URLs relevant to the word.

The use of the functionality to lookup all pages associated with a keyword is a time-saving architecture, as it would require excessively unworkable amounts of time to search all webpages for a keyword in real-time, each time someone searches for it.

Not all crawled pages will be kept in the search index, for various reasons. For instance, if a page includes a robots meta tag with a "noindex" directive, it instructs the search engine to not include the page in the index.

Similarly, a webpage may include an X-Robots-Tag in its HTTP header that instructs the search engines not to index the page.

In yet other instances, a webpage's canonical tag may instruct a search engine that a different page from the present one is to be considered the main version of the page, resulting in other, non-canonical versions of the page to be dropped from the index.

Google has also stated that webpages may not be kept in the index if they are of low quality (duplicate content pages, thin content pages, and pages containing all or too much irrelevant content).

There has also been a long history that suggests that websites with insufficient collective PageRank may not have all of their webpages indexed – suggesting that larger websites with insufficient external links may not get indexed thoroughly.

An insufficient crawl budget may also result in a website not having all of its pages indexed.

A major component of SEO is diagnosing and correcting when pages do not get indexed. Because of this, it is a good idea to thoroughly study all the various issues that can impair the indexing of web pages.

RankBrain

During the indexing stage, search engines use an algorithm to help them process and understand web information. The algorithm operates within a set of rules and it's a unique formula, and search engines use this formula to determine the significance of each individual web page.

Part of Google's algorithm is called **RankBrain**. RankBrain is a machine-learning artificial intelligence system that helps Google process some of its search results, in particular, rare or one-of-a-kind queries. Machine learning is where a computer teaches itself how to do something, rather than being taught by humans or following programming.

RankBrain was not introduced as a new way for Google to rank search results; it is simply part of Google's overall search algorithm, a computer program that's used to sort through the billions of pages it knows about and finds the ones deemed most relevant for particular queries.

Ranking

The ranking of web pages is the stage of search engine processing that is probably the most focused upon.

Once a search engine has a list of all the web pages associated with a particular keyword or keyword phrase, it then must determine how it will order those pages when a search is conducted for the keyword.

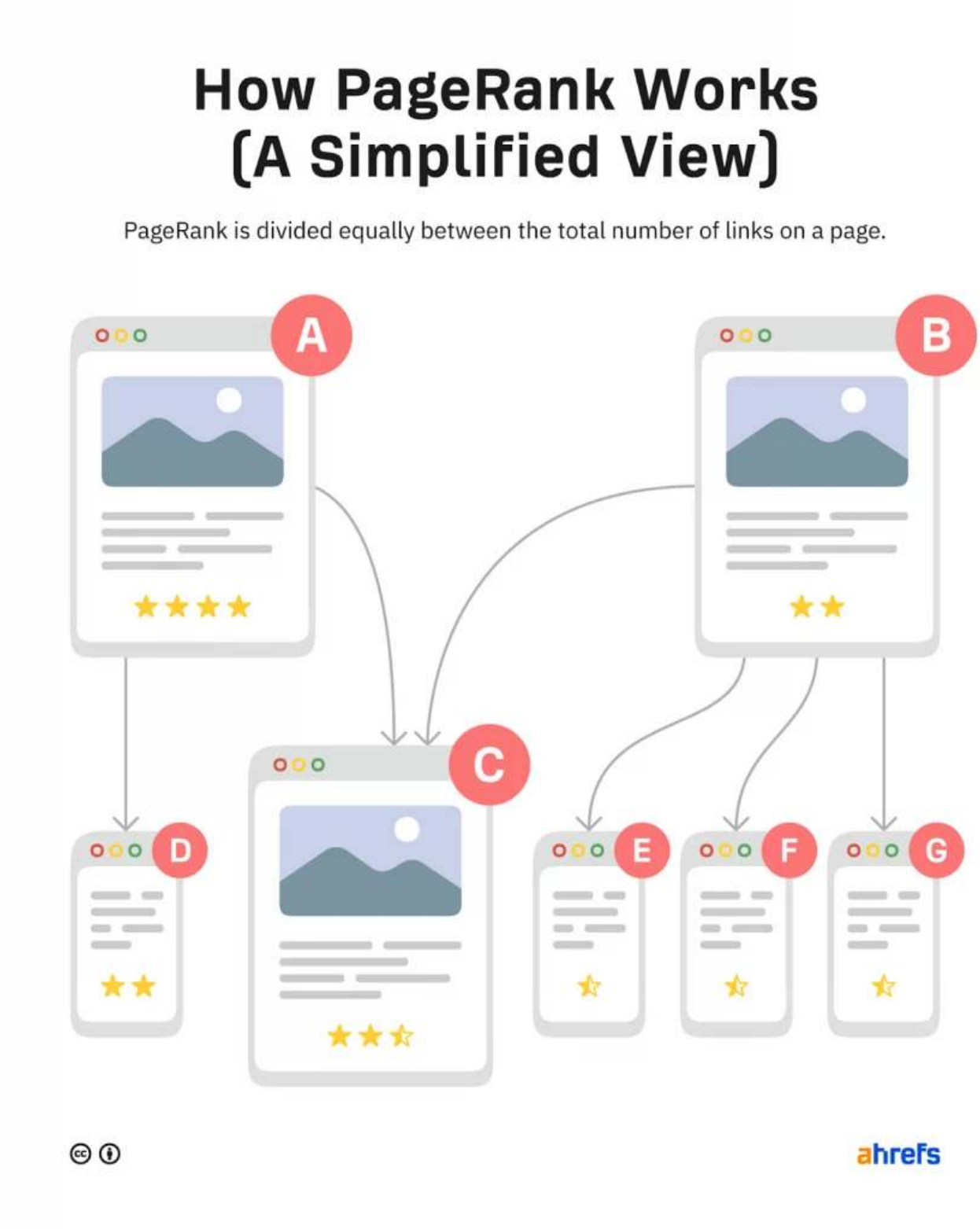
If you work in the SEO industry, you likely will already be pretty familiar with some of what the ranking process involves. The search engine's ranking process is also referred to as an "algorithm".

The complexity involved with the ranking stage of search is so huge that it alone merits multiple articles and books to describe.

There are a great many criteria that can affect a webpage's rank in the search results. Google has said there are **more than** 200 ranking factors used by its algorithm.

Within many of those factors, there can also be up to 50 "vectors" – things that can influence a single ranking signal's impact on rankings.

PageRank is Google's earliest version of its ranking algorithm invented in 1996. It was built off a concept that links to a webpage – and the relative importance of the sources of the links pointing to that webpage – could be calculated to determine the page's ranking strength relative to all other pages.


How PageRank Works
(A Simplified View)
PageRank is divided equally between the total number of links on a page.
A
B
C
D
E
F
G
ahrefs

A great visual on how PageRank works from AHREFS. Keep in mind that AHREFS does not measure PR, but rather their own algorithm that scores for DR or Domain Rating which is not the same as PR or PageRank.

A metaphor for this is that links are somewhat treated as votes, and pages with the most votes will win out in ranking higher than other pages with fewer links/votes.

Fast forward to 2022 and a lot of the old PageRank algorithm's DNA is still embedded in Google's ranking algorithm. That link analysis algorithm also influenced many other search engines that developed similar types of methods.

The old Google algorithm method had to process over the links of the web iteratively, passing the PageRank value around among pages dozens of times before the ranking process was complete. This iterative calculation sequence across many millions of pages could take nearly a month to complete.

Nowadays, new page links are introduced every day, and Google calculates rankings in a sort of drip method – allowing for pages and changes to be factored in much more rapidly without necessitating a month-long link calculation process.

Additionally, links are assessed in a sophisticated manner – revoking or reducing the ranking power of paid links, traded links, spammed links, non-editorially endorsed links and more.

Basics of Ranking/SEO

The internet is absolutely full of "experts" and "course creators" who claim guru status on all things SEO. Many folks claim to have info or a tool that they build up to be a magic bullet of sorts to get all the page one rankings you could imagine. In reality, there might be some good shortcuts here and there and ways to automate some of the processes but there is no such thing as a magic bullet. You have to put in the work. 90% of the results in SEO come from the first 10% of the process which means your research, strategy, and technical SEO phase. The rest of the process month over month in the long-term is targeted content generation, syndication, and amplification.

If you want a really, really good guide to learning the basics of SEO, I highly recommend this **SEO Guide** which covers:

- SEO Basics
- Keyword Research
- Search Engines
- Content Optimization
- Technical and on-page SEO
- Backlinks and link building (properly called link signals)

Can search engines follow your site navigation?

Just as a crawler needs to discover your site via links from other sites, it needs a path of links on your own site to guide it from page to page. If you've got a page you want search engines to find but it isn't linked to from any other pages, it's as good as invisible. Many sites make the critical mistake of structuring their navigation in ways that are inaccessible to search engines, hindering their ability to get listed in search results.

Common navigation mistakes that can keep crawlers from seeing all of your site:

- Having a mobile navigation that shows different results than your desktop navigation
- Any type of navigation where the menu items are not in the HTML, such as JavaScript-enabled navigations. Google has gotten much better at crawling and understanding Javascript, but it's still not a perfect process. The more surefire way to ensure something gets found, understood, and indexed by Google is by putting it in the HTML.
- Personalization, or showing unique navigation to a specific type of visitor versus others, could appear to be cloaking to a search engine crawler
- Forgetting to link to a primary page on your website through your navigation — remember, links are the paths crawlers follow to new pages!

This is why it's essential that your website has clear navigation and helpful URL folder structures.

CREDITS AND CITATIONS:

4 stages of search: Search Engine Land by Chris Silver Smith

How search engines work: Digital Marketing Institute

How search engines operate: Moz