How Search Engines Work

Instructor: Pam Ann Aungst, M.B.A.

Course Objectives

At the end of this course, you will be able to understand how search engines work, including:

- How search engines discover, collect, and sort (rank) web pages

The Mission of Search Engines

Larry Page and Sergey Brin, Founders of Google, in the garage where they started the company in the late 90s

Google, today's most popular search engine, says that their mission is "***to organize the world's information and make it universally accessible and useful***" and that "***the relentless search for better answers continues to be at the core of everything we do***."

Image credit and quote source: From the Garage to the Googleplex

The Mission of Search Engines

Search engines have one very simple objective:

Present users the information they are seeking.

Simple Concept, Complex Execution

Although the mission of search engines is simple, executing on that mission is not easy.

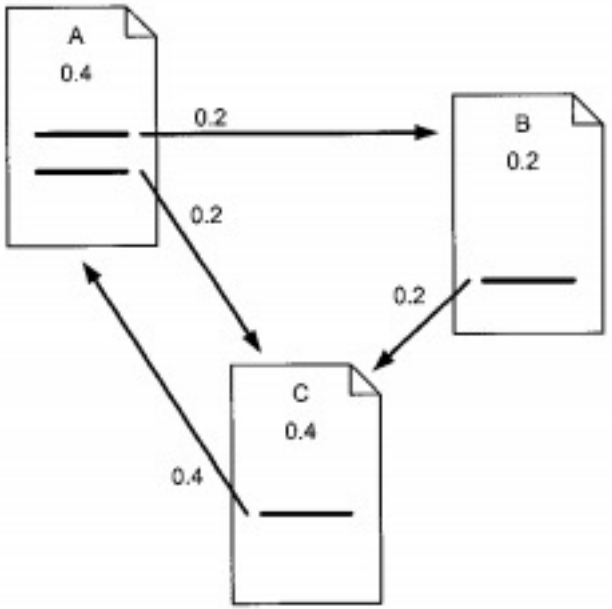
As of Google's latest estimation, there were over 130 trillion individual pages on the web. And that was over 2 years ago, so there's many more now!

Image credit and quote source: Search Engine Land

Simple Concept, Complex Execution

Therefore, search engines need to use very complex technology to analyze trillions of web pages and decide which ones provide the most accurate and trustworthy information for each search.

Although this technology is complex, we can break it down into 3 basic concepts.



A diagram depicting relationships between web pages, from Google's first patent, "Method for Node Ranking in a Linked Database."

Image credit: US Patent #6,285,999

Activity!

Try this simple activity to get a feel for just how many web pages Google processes for a single search, and how fast they do it.

- Using a desktop browser, visit https://www.google.com/ and search for your favorite food.

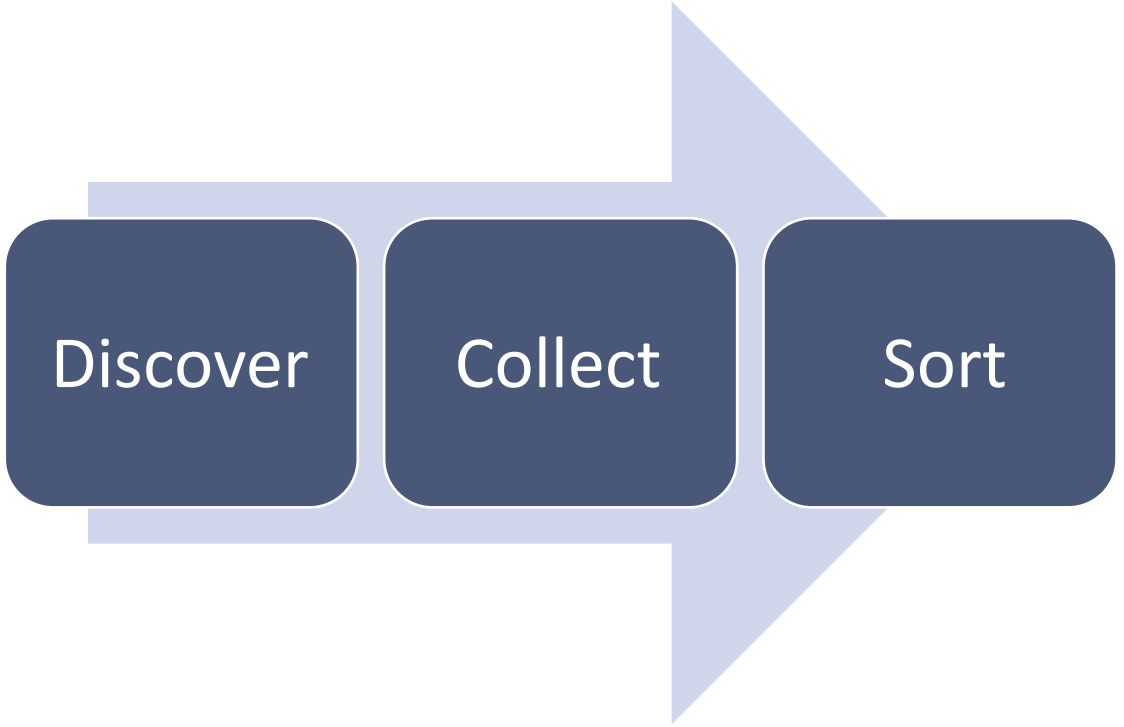
- Under the word "All," take note of how many results were provided and the time it took to provide them.

This simple activity helps demonstrate just how much content a search engine has to process for a single search, and how incredibly fast it happens!

Discover, Sort, Collect

Search engines need to perform 3 basic tasks in order to produce search results.

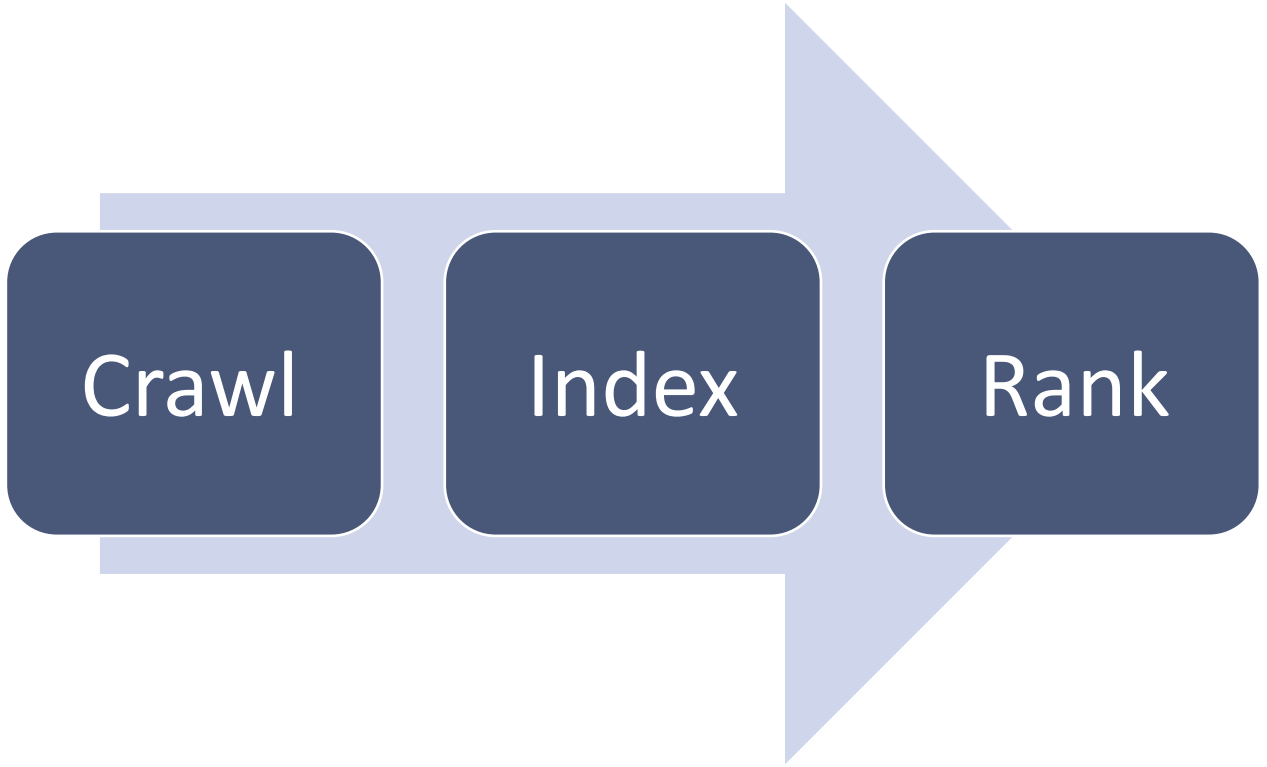
1. First, they need to **discover** web pages.
2. Then, they need to **collect** the content from the web pages.
3. Lastly, they need to **sort** the web pages by relevance to a search query.



Crawl, Index, Rank

The more technical terms for these steps are:

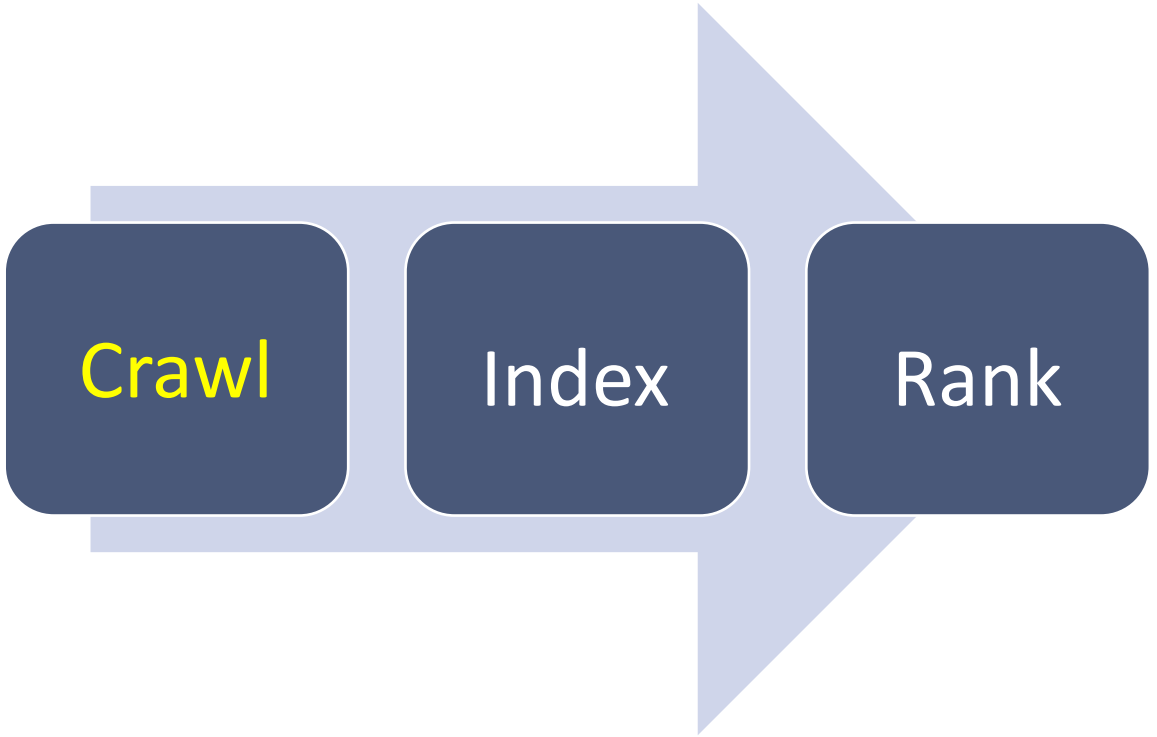
1. Discover = “Crawl”
2. Collect = “Index”
3. Sort = “Rank”



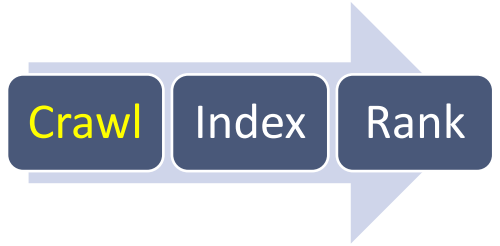
Crawl

The process of discovering available website content is called “crawling.”

- Search engines use automated computer software, often referred to as a *bot*, *crawler*, *robot*, or *spider*, to find web pages on the internet.
- Google's nickname for its crawler is "Googlebot," and Bing's is "Bingbot."



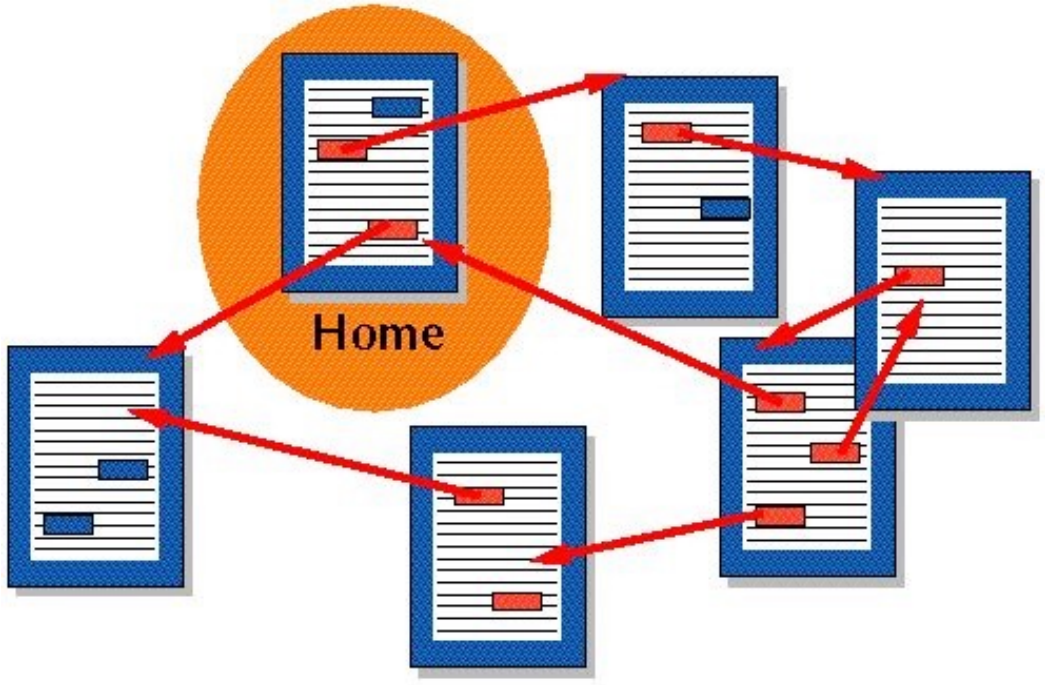
Crawl



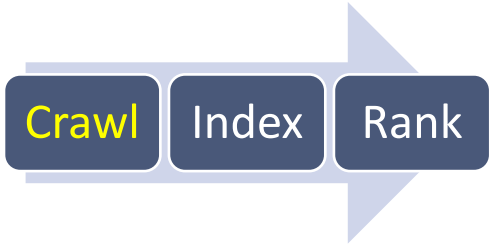
The crawler software starts by scanning a website and looking for links.

- Links are clickable pathways to either another page on the same website, or a page on a different website.

- The program then "follows" each link, visiting each internal and external page that the first page linked to.

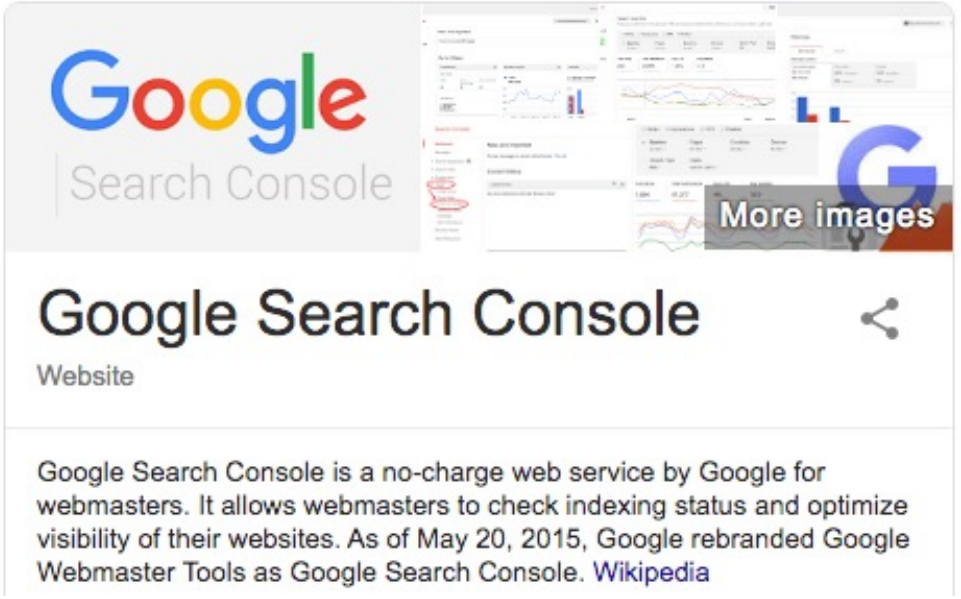


Crawl

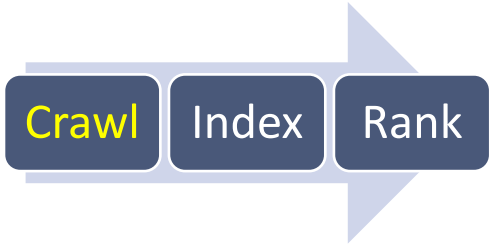


Through this process, search engines are able to find a majority of content that exists on the web.

- Most search engines also offer website owners a way to directly submit pages for crawling.
  - For Google and Bing, this can be done through Google Search Console and Bing Webmaster Tools.



Crawl: Recap



“Crawling” refers to the process of a search engine discovering content on the internet.

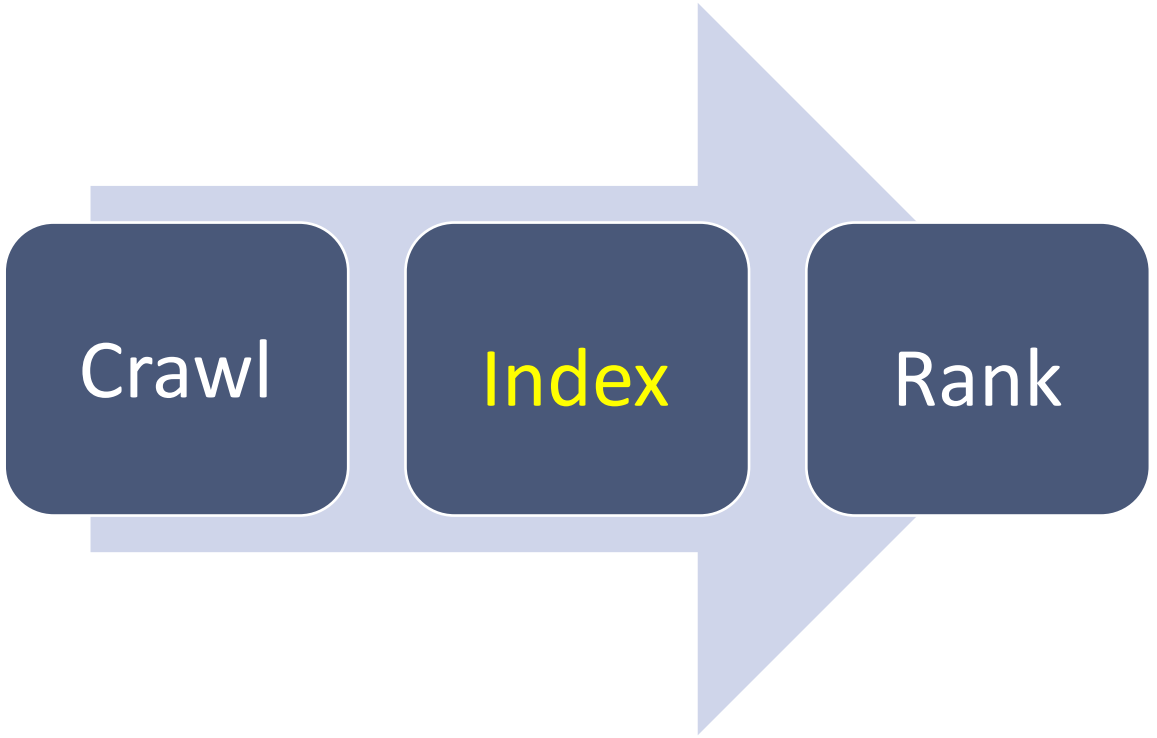
- Search engines discover content in two ways:
  - By following links within content it has already discovered
  - By users submitting pages directly to the search engine for crawling

RECAP

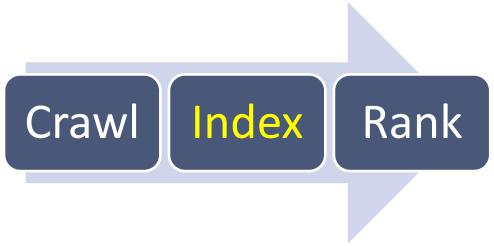
Index

As the search engine crawler proceeds from page to page, it also collects a copy of the information on that page. This is called “indexing.”

The intentions behind indexing are twofold.

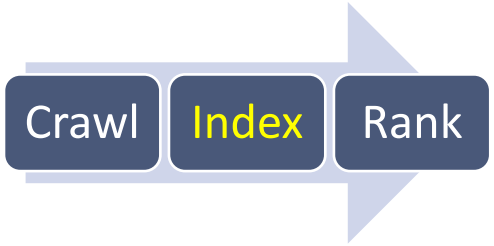


Index

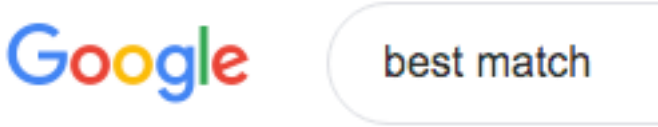


1. **First, search engines need to have a copy of all available web pages on their own servers in order to be able to process searches quickly.**
    - It would take too long to check every website in the world in real time for a match every time a user types in a search query.
    - It is much more efficient for search engines to process user search queries against their internal index (copy) of the internet's content.

Index

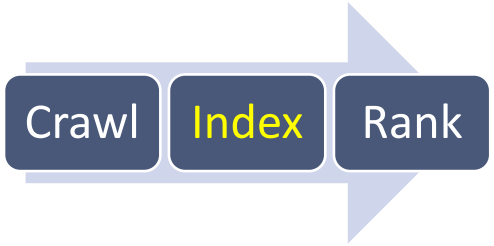


2. **Secondly, search engines need a way to sort through all of the available web page content in order to determine which pages are the best results for a search.**



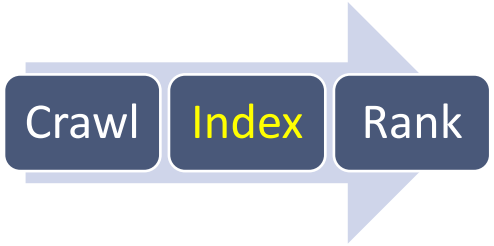
- Therefore, they need to have a copy of all web page content in one place in order to compare them all at the same time.

Index



Search engine crawlers will revisit each page in their index from time-to-time to retrieve the most recent version, in case something changed on the page, and to see if that page has begun to link to new pages that didn't formerly exist.

Index

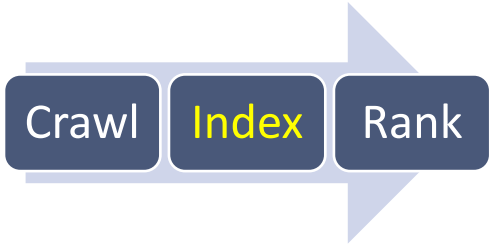


The frequency in which crawlers revisit a site varies and depends upon a variety of factors, including:

- How much content the site has
- How frequently the site tends to make updates to its content
- How popular search engines perceive the site to be

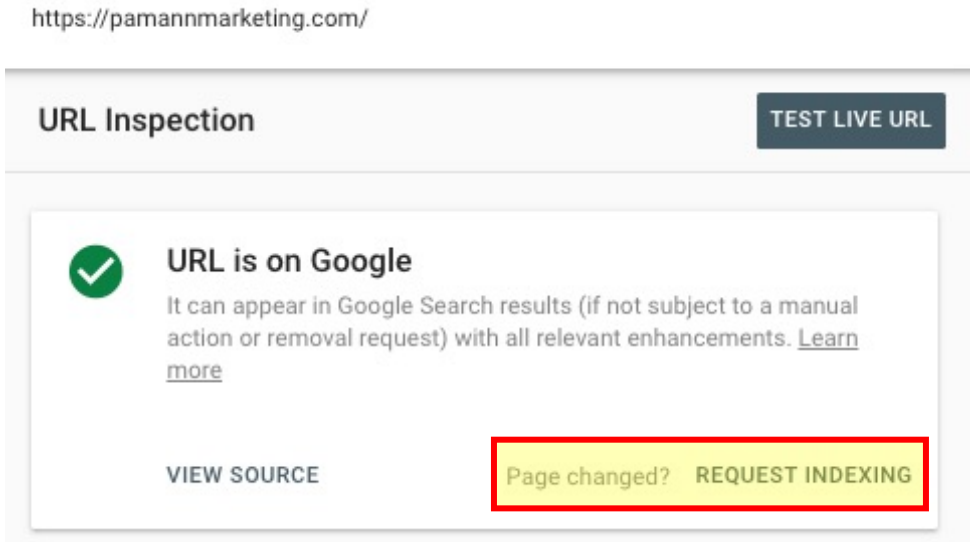
Generally, sites that are frequently updated tend to be re-crawled more often than those that aren’t.

Index

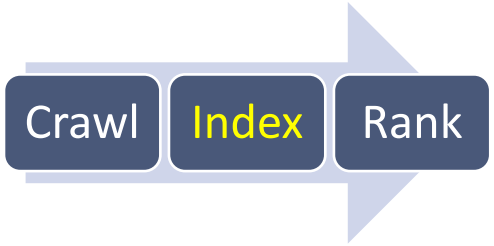


Most search engines also offer a way to manually request re-indexing of existing content or indexation of new content.

- For Google and Bing, this can be done through Google Search Console and Bing Webmaster Tools.



Index: Recap

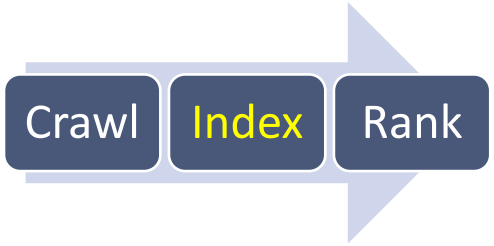


“Indexing” refers to the process of a search engine collecting a copy of the information on each page as it discovers (crawls) it.

- Indexing serves two purposes:
  - Makes it possible to perform searches very quickly.
  - Provides a way to compare all available web content at once for sorting it in the results.

RECAP

Index: Recap



Search engine crawlers revisit each page in their indexes from time to time to see if the content has changed or if new content has been added.

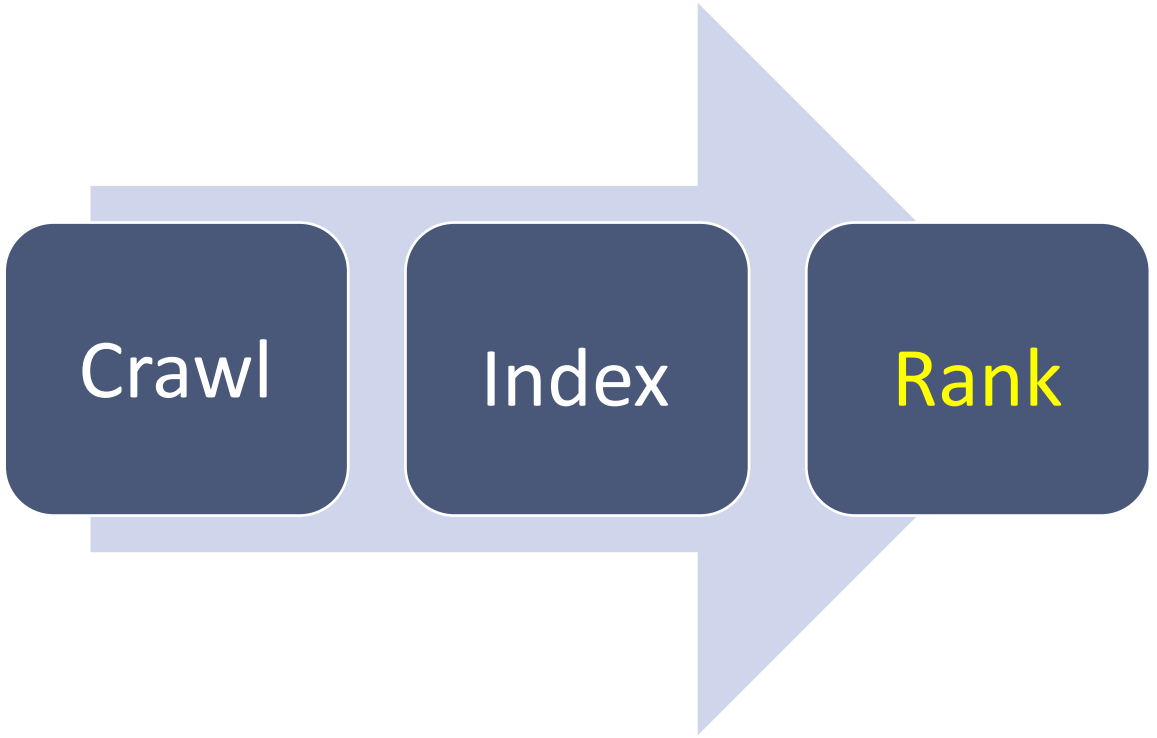
- They will do this more often for large sites, popular sites, and sites that update or add to their content frequently.

- Search engines also offer tools that allow website owners to request re-indexing of existing content or indexing of new content.

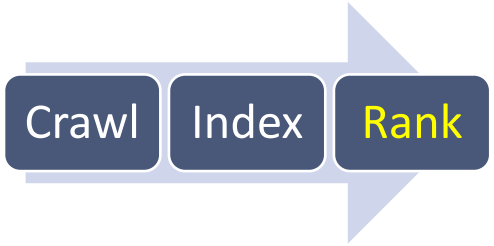
RECAP

Rank

Once a search engine discovers (crawls) and collects (indexes) content, it can then sort (rank) the content for each search query performed by a user.



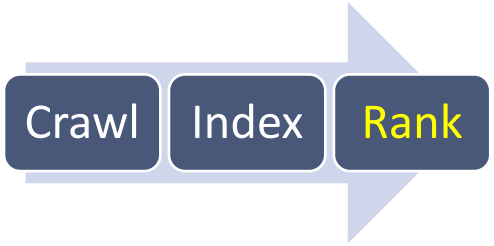
Rank



The process of comparing available web page content to determine the best matches, and then sorting those matches in order of most relevant to least relevant, is referred to as "ranking."

- To rank web pages, search engines use complex *algorithms* to decide which web pages are relevant to the search query, and then order them from most to least relevant.

Rank



An algorithm is essentially a set of "if this, then that" statements which check for the presence of certain factors.

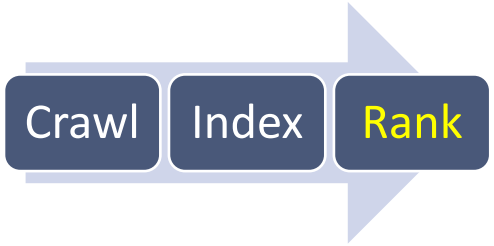
- For example, "If the web page content contains the word 'yellow,' then it is relevant to searches that contain the word 'yellow.'"
    - The presence of the keyword in the content is a factor that the statement is checking for.

al·go·rithm
/ˈalgəˌriT͟Həm/

noun

a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.
"a basic **algorithm for** division"

Rank

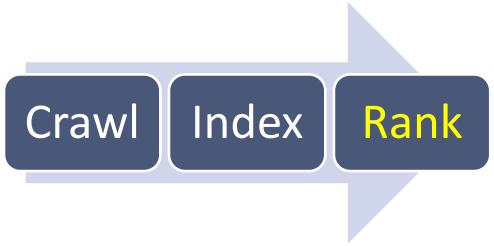


That's an example of what a single "if this, then that" statement may look like in an algorithm.

- This simple example is reflective of how early search engine technology worked.

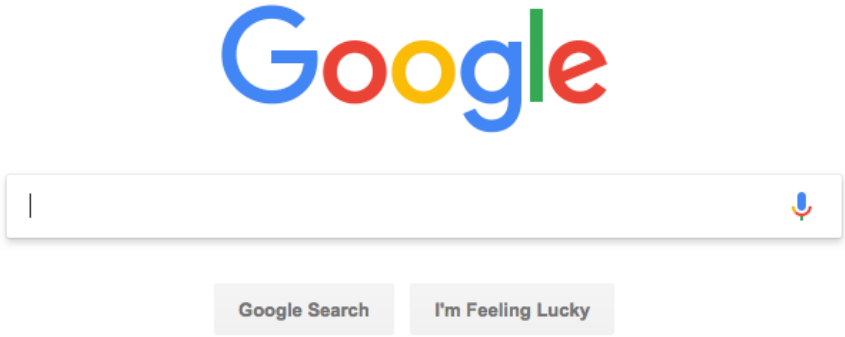
Screenshot of Google upon launch in 1998. Source: Mashable

Rank



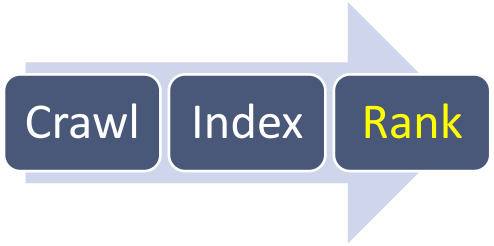
However, a modern search engine algorithm is comprised of a set of many different "if this, then that" statements all stacked on top of one another.

- These stacked statements check for hundreds of different factors at once, and consider different factors with different measures of importance.



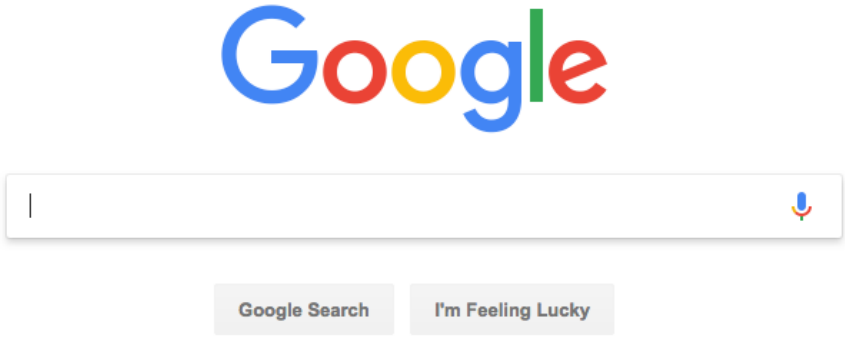
Screenshot of Google in 2018. Source: Google.com

Rank



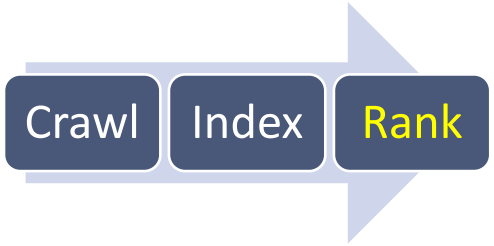
Search engines don't give out every detail about what factors are contained in their algorithms, nor how they are weighted for importance.

That is their intellectual property and they are very careful to protect it.



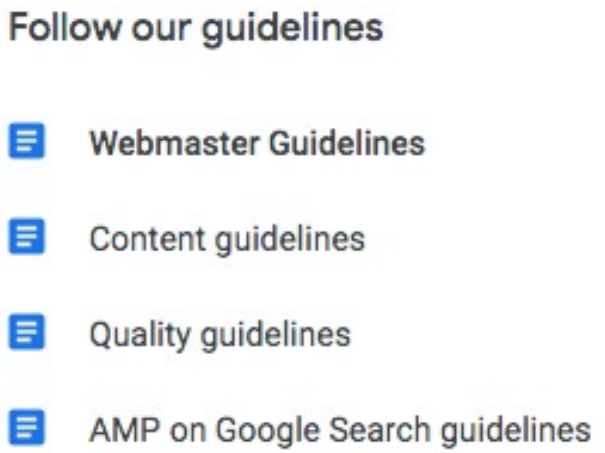
Screenshot of Google in 2018. Source: Google.com

Rank



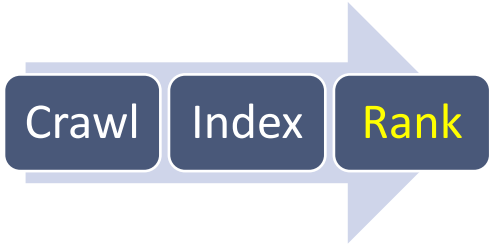
However, search engines do apply for patents for certain pieces of their technology, and they do tell us about best practices that they want us to follow via their "webmaster guidelines."

- Google and Bing each publish their own sets of webmaster guidelines, but they refer to many of the same encouraged and discouraged practices.



Screenshot of Google Webmaster Guidelines index.

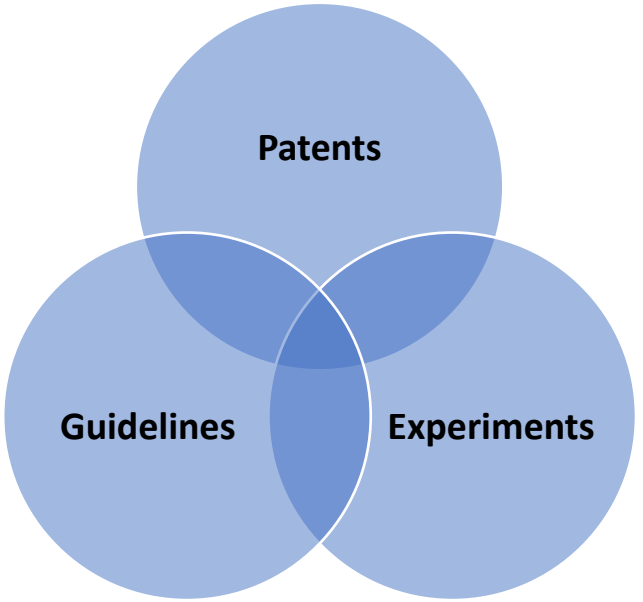
Rank



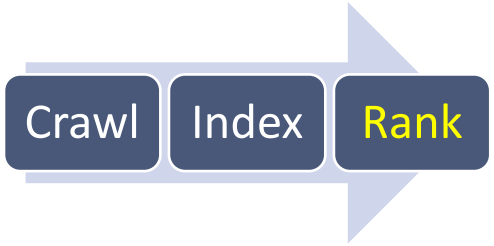
SEO agencies and other technology companies also conduct experiments to learn about search engine algorithm behavior.

- Therefore, from the patents, published guidelines, and experiments, we know that there are over 200 different factors that search engine algorithms check for when ranking content for a search query.

- We also know a lot about what the algorithms tend to favor and what they're designed to dislike.

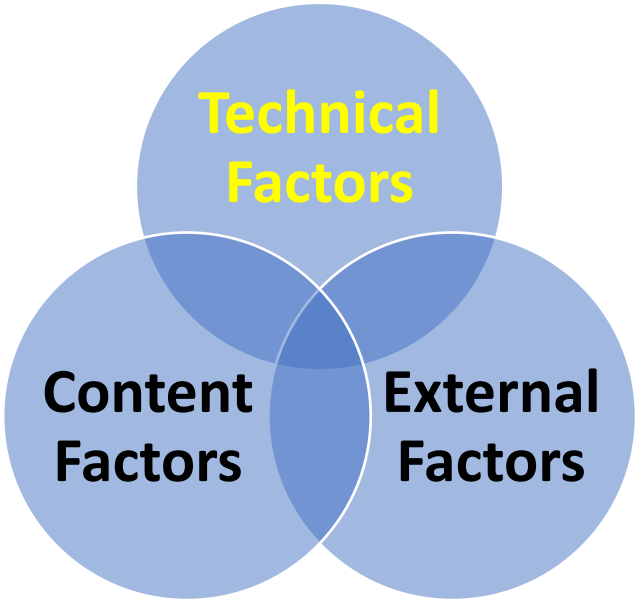


Rank

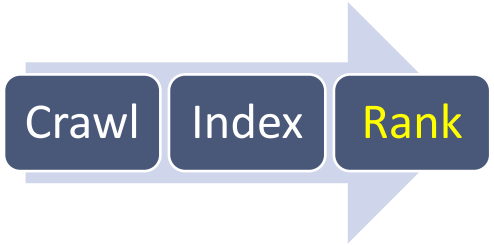


A few examples of factors that search engines check for in their algorithms include:

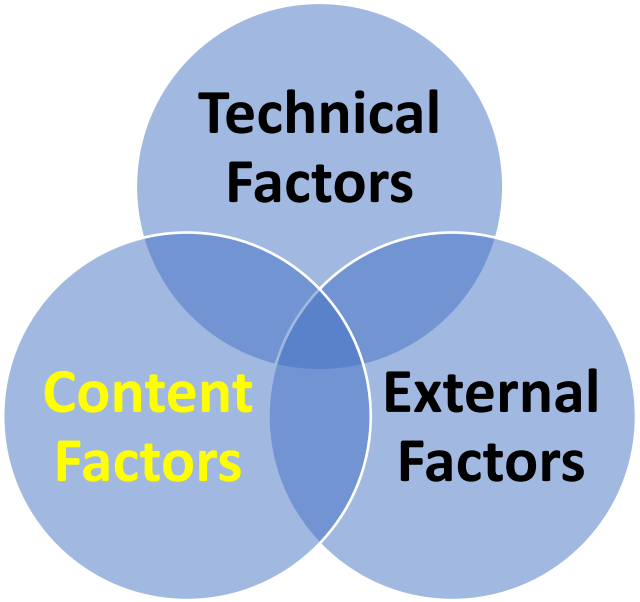
- **Technical Factors**, such as:
  - Whether or not the page is mobile-friendly
  - How fast the page loads
  - Meta tags, which are bits of code with information about the page



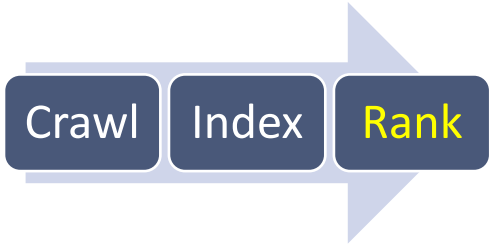
Rank



- **Content Factors,** such as:
  - The presence of keywords in the content on the page
  - The quality, trustworthiness, and authoritative nature of the content
  - How many other pages on the website link to that page
  - How many clicks it takes to get to that page on the website

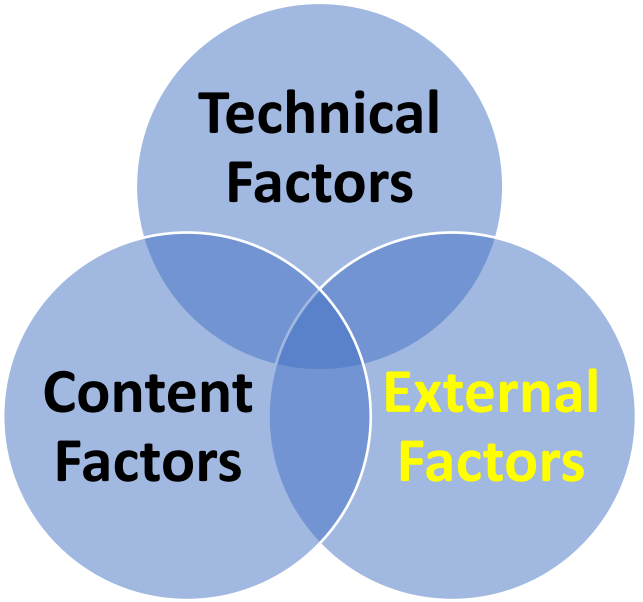


Rank

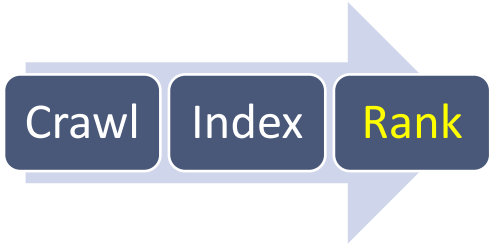


- **External Factors,** such as:
  - How many other **high-quality*** websites link to that page, and to the website the page is part of (often referred to as “inbound links.”)

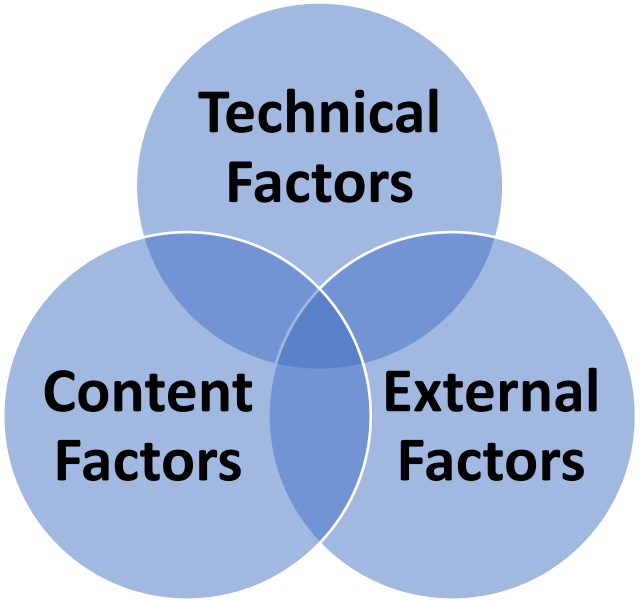
*It’s important to note that quality is more impactful than quantity when it comes to inbound links.



Rank

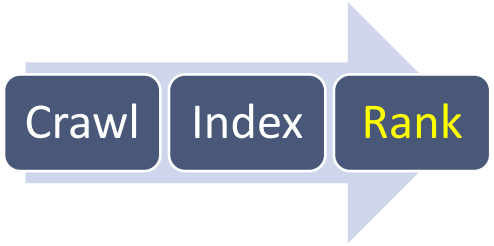


Those are just a few examples of the over 200 factors that search engine algorithms take into account when ranking a web page.



200+ Total Factors!

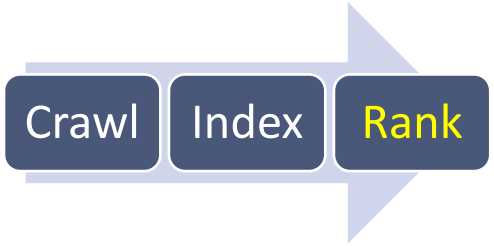
Rank: Recap



Ranking is the process of comparing available web page content to determine the best matches, and then sorting those matches in order of most relevant to least relevant

- To rank web pages, search engines use complex algorithms.
- An algorithm is essentially a set of "if this, then that" statements which check for the presence of certain factors.

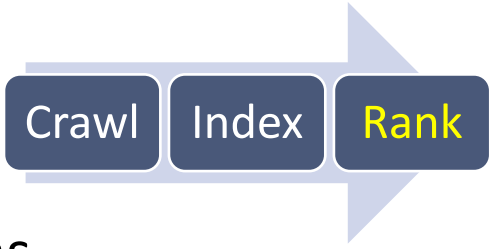
Rank: Recap



- Modern search engine algorithms are comprised of many different "if this, then that" statements all stacked on top of one another.
- These stacked statements check for hundreds of factors, and consider different factors with different measures of importance.

RECAP

Rank: Recap



- From search engine patents and guidelines, as well as SEO experiments, we know that there are over 200 factors that the algorithms consider.

- Examples of such factors include:

  - Technical factors such as **page load time**

  - Content factors such as **the presence of keywords on a page**

  - External factors such as **how many other high-quality websites link to that website**

RECAP