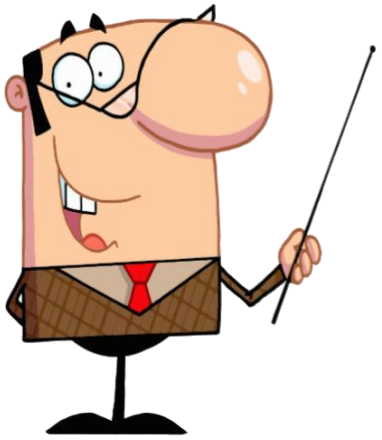
Search Engine

13

Web Technology



Asst. Prof. Manop Phankokkruad, Ph.D.
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang

Outline

1. What is Search Engine?
2. How Search Engines Work?
   A. Crawling
   B. Indexing
   C. Ranking

What is Search Engine?

Search engine is a software program that searches for sites based on the words that user designate as search terms.

- A user enters keywords or key phrases into a search engine and receives a list of Web content results in the form of websites, images, videos or other types of files.
- The list of content returned via a search engine to a user is known as a search engine results page (SERP).

Purpose of Search Engines

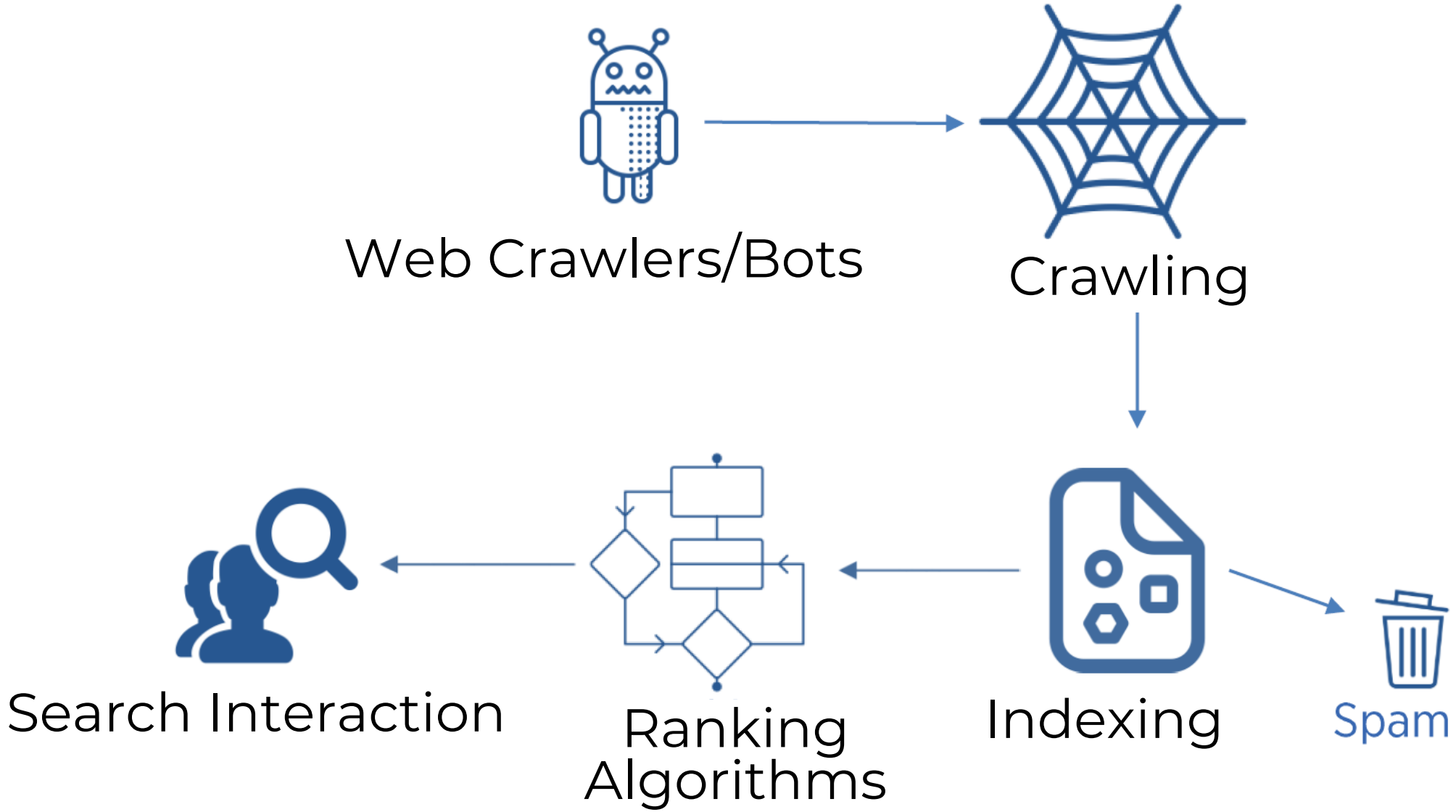
- Helping people find what they're looking for
  - Starts with an "information need"
  - Convert to a query
  - Gets results
- In the materials available
  - Web pages
  - Other formats
  - Deep Web

How Search Engines Work?

Search engines have three primary functions:

1. **Crawling**: Discover the Internet for content, looking over the code/content for each URL they find.
2. **Indexing**: Store and organize the content found during the crawling process. Once a page is in the index, it’s in the running to be displayed as a result to relevant queries.
3. **Ranking**: Provide the pieces of content that will best answer a searcher's query, which means that results are ordered by most relevant to least relevant.

How Search Engines Work?



A. Crawling

1. Search Engine collect information from selected web sites.
    - The engines find these pages, they decipher the code from them, and store selected pieces in massive databases, to be recalled later when needed for a search query.
    - The special software robots, called ***spiders***, to crawl web pages(**Web crawler**).
2. Spiders build lists of the words found in Web sites. When a spider is building its lists, the spider is Web crawling.

A. Crawling

3. Spiders store the lists in the engine's database.
   - the search engine companies have constructed datacenters all over the world.
   - These monstrous storage facilities hold thousands of machines processing large quantities of information very quickly.
   - The engine's indexing software builds an index of words.
   - The information is matched against query input and retrieved (processing algorithm).

A. Crawling

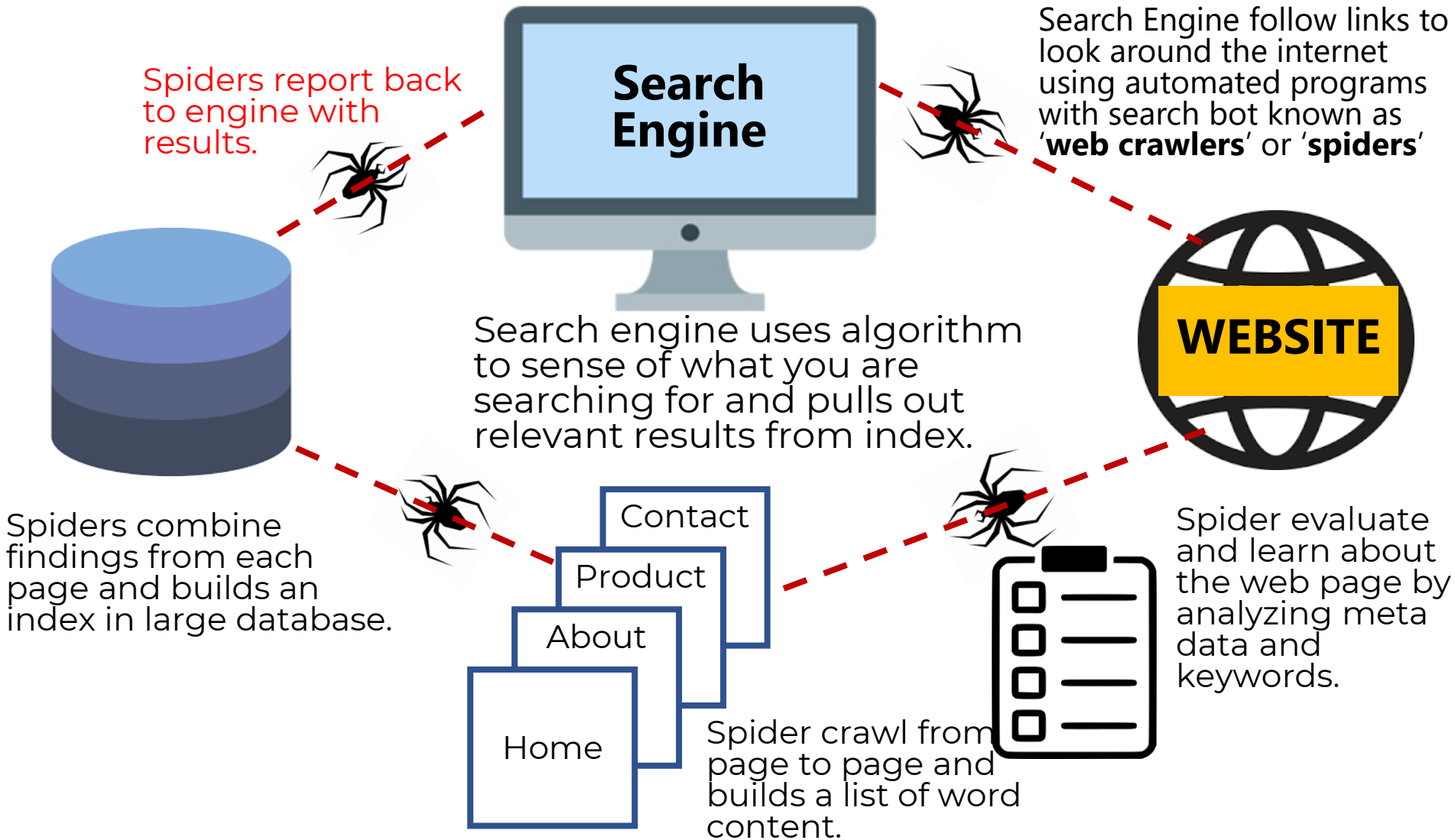


Figure : How Search Engines Work?

How Spiders and Crawlers Work?

- They begin with popular and heavily used web servers.
- They begin with a popular site, collect the words on its pages and follow every link found within the site. Spiders travel across pages and the most widely used portions of the Web.
- A dedicated server of URLs is built by a search engine company (e.g., Google) so that spiders collect information quickly.

How Spiders and Crawlers Work?

- More than one spider is used to crawl web pages at a time. Google uses 3-4 spiders and collect over 100 pages per second
- When no dedicated URL server is used, search engine company relies on ISP for the domain names (translated into addresses) to use for crawling the web.
  - Delay in gathering information
  - Delay in updating information
  - Lack of control over URL addresses

B. Indexing

Search engines process and store information, they find in an index, a huge database of all the content they’ve discovered and good enough to serve up to searchers.

What the Index Needs ?

- ❑ Basic information for document or record
  - File name / URL / record ID
  - Title or equivalent
  - Size, date, MIME type

B. Indexing

What the Index Needs ? *(cont.)*

- ❑ Full text of item
- ❑ More metadata
  - Product name, picture ID
  - Category, topic, or subject
  - Other attributes, for relevance ranking and display

B. Indexing

How Search Engines Store Words Indexed?

- The process varies among engines
- Words are stored with number of times they appear on a pages.
- Weight is assigned to each word.
- Words appearing near top of a page may have more weight than those appearing in sub-headings, in links, in meta tags, in title, etc.

B. Indexing

Bots give full web pages text to the indexer.

- Stop words like (for, in, at etc) and punctuation are ignored.
- The text is converted to lower case and stored.

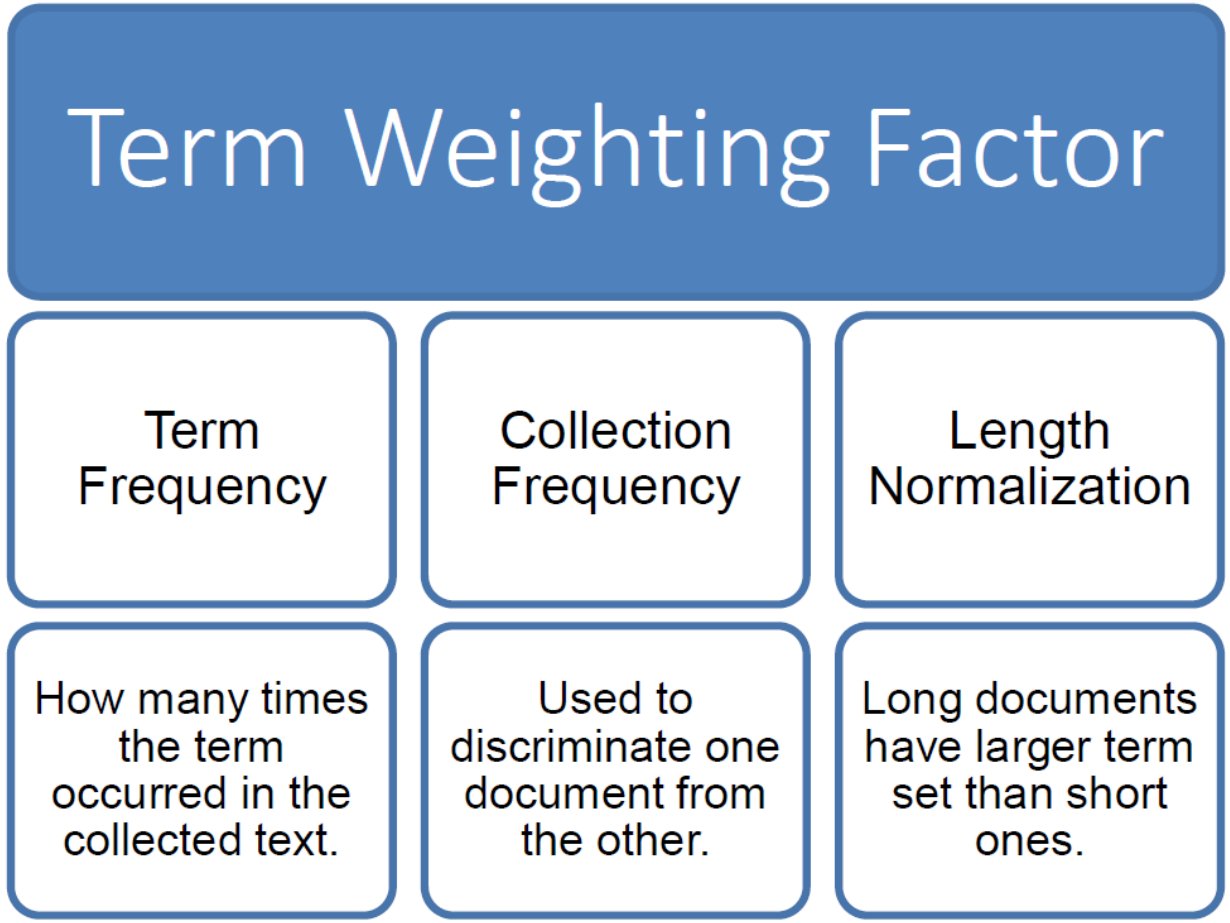


Figure : Term Weighting Factor

B. Indexing

How Search Engines Store Words Indexed? *(cont.)*

- ❑ Information is encoded to save space
- ❑ Information is indexed
  - An index of words is built by the automatic indexer (indexing software).
  - A hash table is created with an assigned weight or value for each word indexed.
  - Hashing allows for even the distribution of popular entries with those that are less popular for quick retrieval.

C. Ranking & Providing answers

When a person performs an online search, the search engine inspects its corpus of billions of documents and does two things:

1. It returns only those results that are relevant or useful to the searcher's query.
2. It ranks those results according to the popularity of the websites serving the information. It is both ***relevance and popularity*** that the process of search engine optimization (SEO) is meant to influence.

C. Ranking & Providing answers

How do search engines determine relevance and popularity?

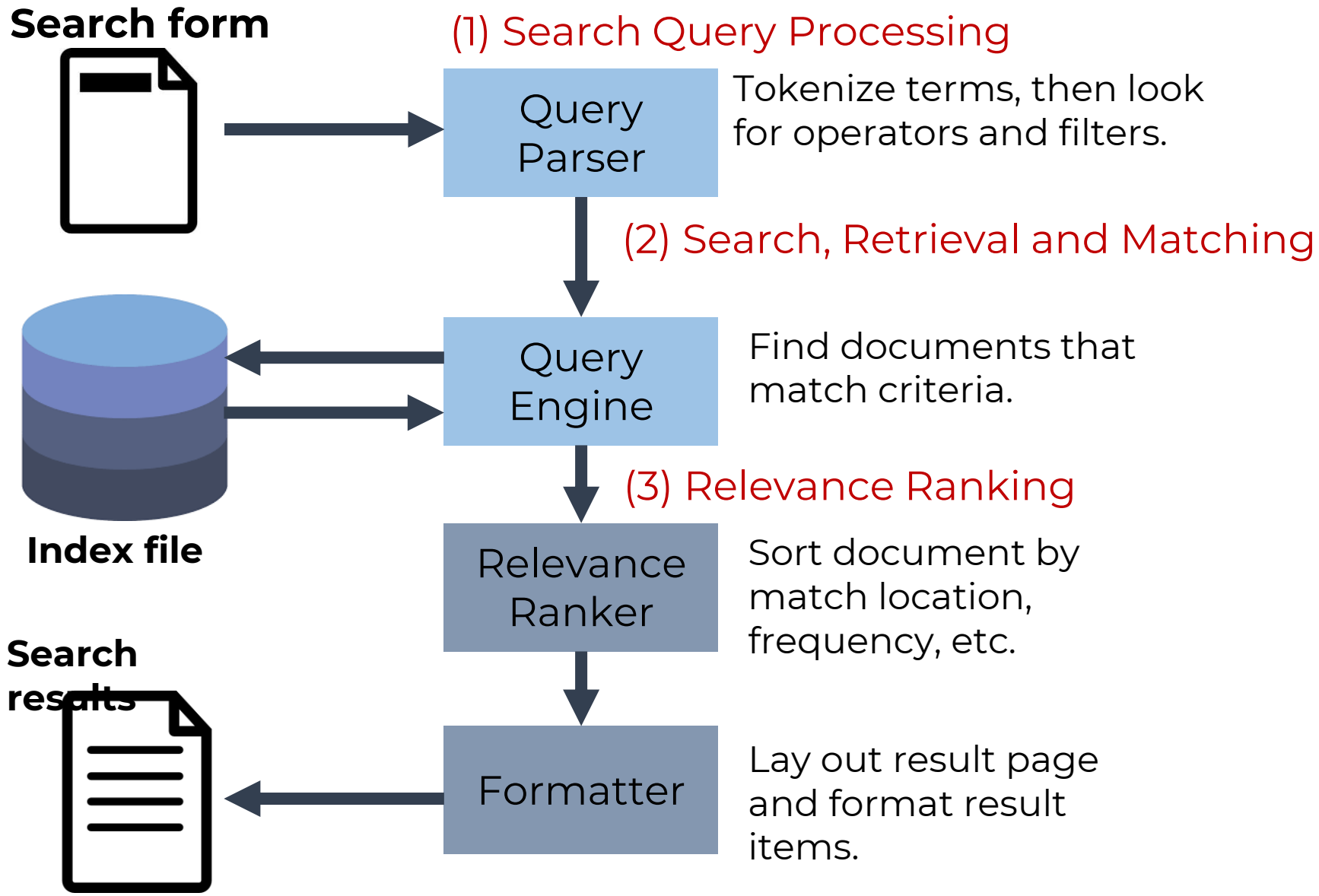
1. Relevance means more than finding a page with the right words. Hundreds of factors influence relevance.
2. Search engines typically *assume that the more popular a site*, page, or document, the more valuable the information it contains must be.

C. Ranking & Providing answers

How do search engines determine relevance and popularity? (cont.)

3. *Popularity and relevance* aren't determined manually. Instead, the engines employ mathematical equations (algorithms) to sort the relevance, and then to rank the popularity.
4. These algorithms often comprise hundreds of variables. We refer to them as "***ranking factors***."

Providing answers



(1) Search Query Processing

After users click the search button, and before retrieval starts. These processes will be done.

- Handle character set, maybe language
- Look for operators and organize the query
- Look for field names or metadata
- Extract words (just like the indexer)
- Deal with letter casing.

(2) Search, Retrieval and Matching

- Single-word queries
  - Find items containing that word
- Multi-word queries: combine lists
  - Any: every item with any query word
  - All: only items with every word
  - Phrases: find only items with all words in order
- Boolean and complex queries
  - Use algorithm to combine lists

(3) Relevance Ranking

- Theory : sort the matching items, so the most relevant ones appear first.
- Can't really know what the user wants.
- Relevance is hard to define and situational.
- Short queries tend to be deeply ambiguous , such as
  - What do people mean when they type “bank”?
- First 10 results are the most important.

(3) Relevance Ranking

- Heuristics are rules of thumb
  - Not algorithms, not math
- Search Relevance Ranking Heuristics
  - Documents containing all search words
  - Search words as a phrase
  - Matches in title tag
  - Matches in other metadata
- Based on real-word user behavior

More Information

- HOW SEARCH ENGINES WORK: CRAWLING, INDEXING, AND RANKING

https://moz.com/beginners-guide-to-seo/how-search-engines-operate

- Want More Traffic? Deindex Your Pages. Here's Why.

https://neilpatel.com/blog/deindex-your-pages/