Search Engine Optimization

The Index

The index is where the spider-collected data are stored. When you perform a search on a major search engine, you are not searching the web, but the cache of the web provided by that search engine's index.

Reverse Index

Search engines organize their content in what is called a *reverse index*. A reverse index sorts web documents by words. When you search Google and it displays 1-10 out of 143,000 websites, it means that there are approximately 143,000 web pages that either have the words from your search on them or have inbound links containing them. Also, note that search engines do not store punctuation, just words.

The following is an example of a reverse index and how a typical search engine might classify content. While this is an oversimplified version of the real thing, it does illustrate the point. Imagine each of the following sentences is the content of a unique page:

The dog ate the cat.

The cat ate the mouse.

Word	Document #	Position #
The	1,2	1-1, 1-4, 2-1, 2-4
Dog	1	2
Ate	1,2	1-3, 2-3
Cat	1,2	1-5, 2-2
Mouse	2	5

Storing Attributes

Since search engines view pages from their source code in a linear format, it is best to move JavaScript and other extraneous code to external files to help move the page copy higher in the source code.

Some people also use Cascading Style Sheets (CSS) or a blank table cell to place the page content ahead of the navigation. As far as how search engines evaluate what words are first, they look at how the words appear in the source code. I have not done significant testing to determine if it is worth the effort to make your unique

page code appear ahead of the navigation, but if it does not take much additional effort, it is probably worth doing. Link analysis (discussed in depth later) is far more important than page copy to most search algorithms, but every little bit can help.

Google has also hired some people from Mozilla and is likely working on helping their spider understand how browsers render pages. Microsoft published visually segmenting research that may help them understand what page content is most important.

As well as storing the position of a word, search engines can also store how the data are marked up. For example, is the term in the page title? Is it a heading? What type of heading? Is it bold? Is it emphasized? Is it in part of a list? Is it in link text?

Words that are in a heading or are set apart from normal text in other ways may be given additional weighting in many search algorithms. However, keep in mind that it may be an unnatural pattern for your keyword phrases to appear many times in bold and headings without occurring in any of the regular textual body copy. Also, if a page looks like it is aligned too perfectly with a topic (i.e., overly-focused so as to have an abnormally high keyword density), then that page may get a lower relevancy score than a page with a lower keyword density and more natural page copy.

Proximity

By storing where the terms occur, search engines can understand how close one term is to another. Generally, the closer the terms are together, the more likely the page with matching terms will satisfy your query.

If you only use an important group of words on the page once, try to make sure they are close together or right next to each other. If words also occur naturally, sprinkled throughout the copy many times, you do not need to try to rewrite the content to always have the words next to one another. Natural sounding content is best.

Stop Words

Words that are common do not help search engines understand documents. Exceptionally common terms, such as *the*, are called stop words. While search engines index stop words, they are not typically used or weighted heavily to determine relevancy in search algorithms. If I search for *the Cat in the Hat,* search engines may insert wildcards for the words *the* and *in,* so my search will look like ** cat * * hat.*

Index Normalization

Each page is standardized to a size. This prevents longer pages from having an unfair advantage by using a term many more times throughout long page copy. This also prevents short pages for scoring arbitrarily high by having a high

percentage of their page copy composed of a few keyword phrases. Thus, there is no magical page copy length that is best for all search engines.

The uniqueness of page content is far more important than the length. Page copy has three purposes above all others:

- To be unique enough to get indexed and ranked in the search result
- To create content that people find interesting enough to want to link to
- To convert site visitors into subscribers, buyers, or people who click on ads

Not every page is going to make sales or be compelling enough to link to, but if, in aggregate, many of your pages are of high-quality over time, it will help boost the rankings of nearly every page on your site.

Keyword Density, Term Frequency & Term Weight

Term Frequency (TF) is a weighted measure of how often a term appears in a document. Terms that occur frequently within a document are thought to be some of the more important terms of that document.

If a word appears in every (or almost every) document, then it tells you little about how to discern value between documents. Words that appear frequently will have little to no discrimination value, which is why many search engines ignore common stop words (like *the*, *and*, and *or*).

Rare terms, which only appear in a few or limited number of documents, have a much higher signal-to-noise ratio. They are much more likely to tell you what a document is about.

Inverse Document Frequency (IDF) can be used to further discriminate the value of term frequency to account for how common terms are across a corpus of documents. Terms that are in a limited number of documents will likely tell you more about those documents than terms that are scattered throughout many documents.

When people measure keyword density, they are generally missing some other important factors in information retrieval such as IDF, index normalization, word proximity, and how search engines account for the various element types. (Is the term bolded, in a header, or in a link?)

Search engines may also use technologies like latent semantic indexing to mathematically model the concepts of related pages. Google is scanning millions of books from university libraries. As much as that process is about helping people find information, it is also used to help Google understand linguistic patterns.

If you artificially write a page stuffed with one keyword or keyword phrase without adding many of the phrases that occur in similar natural documents you may not

show up for many of the related searches, and some algorithms may see your document as being less relevant. The key is to write naturally, using various related terms, and to structure the page well.

Multiple Reverse Indexes

Search engines may use multiple reverse indexes for different content. Most current search algorithms tend to give more weight to page title and link text than page copy.

For common broad queries, search engines may be able to find enough quality matching documents using link text and page title without needing to spend the additional time searching through the larger index of page content. Anything that saves computer cycles without sacrificing much relevancy is something you can count on search engines doing.

After the most relevant documents are collected, they may be re-sorted based on interconnectivity or other factors.

Around 50% of search queries are unique, and with longer unique queries, there is greater need for search engines to also use page copy to find enough relevant matching documents (since there may be inadequate anchor text to display enough matching documents).

Search Interface

The search algorithm and search interface are used to find the most relevant document in the index based on the search query. First the search engine tries to determine user intent by looking at the words the searcher typed in.

These terms can be stripped down to their root level (e.g., dropping *ing* and other suffixes) and checked against a lexical database to see what concepts they represent. Terms that are a near match will help you rank for other similarly related terms. For example, using the word swims could help you rank well for swim or swimming.

Search engines can try to match keyword vectors with each of the specific terms in a query. If the search terms occur near each other frequently, the search engine may understand the phrase as a single unit and return documents related to that phrase.

WordNet is the most popular lexical database. At the end of this chapter there is a link to a Porter Stemmer tool if you need help conceptualizing how stemming works.

Searcher Feedback

Some search engines, such as Google and Yahoo!, have toolbars and systems like Google Search History and My Yahoo!, which collect information about a user. Search engines can also look at recent searches, or what the search process was for similar users, to help determine what concepts a searcher is looking for and what documents are most relevant for the user's needs.

As people use such a system it takes time to build up a search query history and a click-through profile. That profile could eventually be trusted and used to

- aid in search personalization
- collect user feedback to determine how well an algorithm is working
- help search engines determine if a document is of decent quality (e.g., if many users visit a document and then immediately hit the back button, the search engines may not continue to score that document well for that query).

I have spoken with some MSN search engineers and examined a video about MSN search. Both experiences strongly indicated a belief in the importance of user acceptance. If a high-ranked page never gets clicked on, or if people typically quickly press the back button, that page may get demoted in the search results for that query (and possibly related search queries). In some cases, that may also flag a page or website for manual review.

As people give search engines more feedback and as search engines collect a larger corpus of data, it will become much harder to rank well using only links. The more satisfied users are with your site, the better your site will do as search algorithms continue to advance.

Real-Time versus Prior-to-Query Calculations

In most major search engines, a portion of the relevancy calculations are stored ahead of time. Some of them are calculated in real time.

Some things that are computationally expensive and slow processes, such as calculating overall inter-connectivity (Google calls this PageRank), are done ahead of time.

Many search engines have different data centers, and when updates occur, they roll from one data center to the next. Data centers are placed throughout the world to minimize network lag time. Assuming it is not overloaded or down for maintenance, you will usually get search results from the data centers nearest you. If those data centers are down or if they are experiencing heavy load, your search query might be routed to a different data center.

Search Algorithm Shifts

Search engines such as Google and Yahoo! may update their algorithm dozens of times per month. When you see rapid changes in your rankings, it is usually due to an algorithmic shift, a search index update, or something else outside of your control. SEO is a marathon, not a sprint, and some of the effects take a while to kick in.

Usually, if you change something on a page, it is not reflected in the search results that same day. Linkage data also may take a while to have an effect on search relevancy as search engines need to find the new links before they can evaluate them, and some search algorithms may trust links more as the links age.

The key to SEO is to remember that rankings are always changing, but the more you build legitimate signals of trust and quality, the more often you will come out on top.

Relevancy Wins Distribution!

The more times a search leads to desired content, the more likely a person is to use that search engine again. If a search engine works well, a person does not just come back, they also tell their friends about it, and they may even download the associated toolbar. The goal of all major search engines is to be relevant. If they are not, they will fade (as many already have).

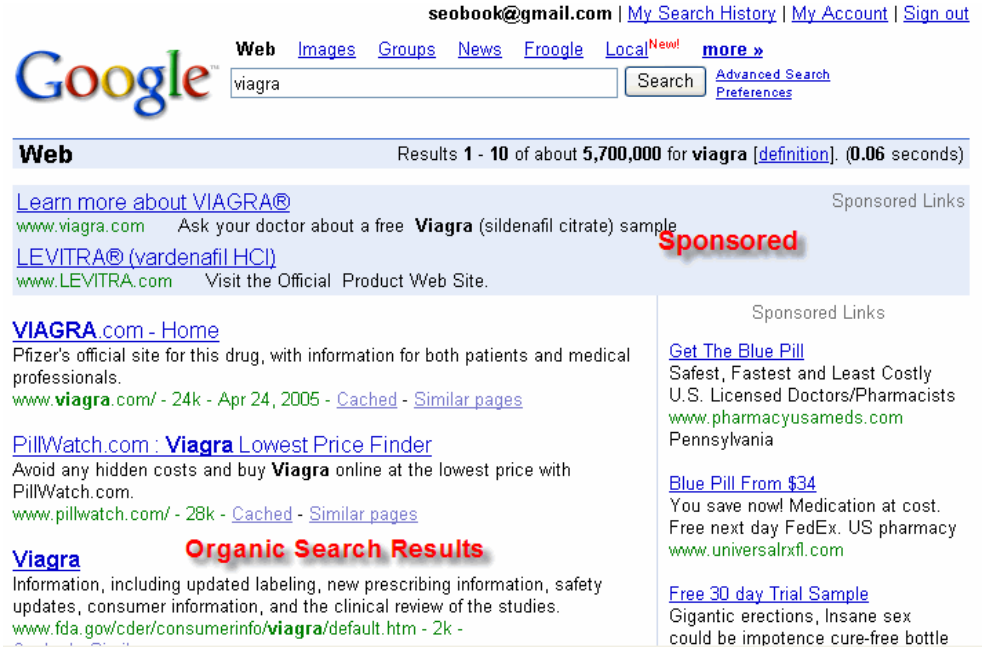
Search Engine Business Model

Search engines make money when people click on the sponsored advertisements. In the search result below you will notice that both Viagra and Levitra are bidding on the term *Viagra*. The area off to the right displays sponsored advertisements for the term Viagra. Google gets paid whenever a searcher clicks on any of the sponsored listings.

The white area off to the left displays the organic (free) search results. Google does not get paid when people click on these. Google hopes to make it hard for search engine optimizers (like you and I) to manipulate these results to keep relevancy as high as possible and to encourage people to buy ads.

Later in this e-book we will discuss both organic optimization and pay-per-click marketing.

Image of Search Results



Origins of the Web

The Web started off behind the idea of the free flow of information as envisioned by Tim Berners-Lee. He was working at CERN in Europe. CERN had a somewhat web-like environment in that many people were coming and going and worked on many different projects.

Tim created a site that described how the Web worked and placed it live on the first server at info.cern.ch. Europe had very little backing or interest in the Web back then, so U.S. colleges were the first groups to set up servers. Tim added links to their server locations from his directory known as the Virtual Library.

Current link popularity measurements usually show college web pages typically have higher value than most other pages do. This is simply a function of the following:

- The roots of the WWW started in lab rooms at colleges. It was not until the mid to late 1990s that the Web became commercialized.
- The web contains self-reinforcing social networks.
- Universities are pushed as sources of authority.
- Universities are heavily funded.
- Universities have quality controls on much of their content.

Early Search Engines

The Web did not have sophisticated search engines when it began. The most advanced information gatherers of the day primitively matched file names. You had to know the name of the file you were looking for to find anything. The first file that matched was returned. There was no such thing as search relevancy. It was this lack of relevancy that lead to the early popularity of directories such as Yahoo!.

Many search engines such as AltaVista, and later Inktomi, were industry leaders for a period of time, but the rush to market and lack of sophistication associated with search or online marketing prevented these primitive machines from having functional business models.

Overture was launched as a pay-per-click search engine in 1998. While the Overture system (now known as Yahoo! Search Marketing) was profitable, most portals were still losing money. The targeted ads they delivered grew in popularity and finally created a functional profit generating business model for large-scale general search engines.

Commercialized Cat & Mouse

Web = Cheap Targeted Marketing

As the Internet grew in popularity, people realized it was an incredibly cheap marketing platform. Compare the price of spam (virtually free) to direct mail (~ $1 each). Spam fills your inbox and wastes your time.

Information retrieval systems (search engines) must also fight off aggressive marketing techniques to keep their search results relevant. Search engines market their problems as spam, but the problem is that they need to improve their algorithms.

It is the job of search engines to filter through the junk to find and return relevant results.

There will always be someone out there trying to make a quick buck. Who can fault some marketers for trying to find holes in parasitic search systems that leverage others' content without giving any kickback?

Becoming a Resource

Though I hate to quote a source I do not remember, I once read that one in three people believe the top search result is the most relevant document relating to their search. Imagine the power associated with people finding your view of the world first. Whatever you are selling, someone is buying!