CHAPTER 1

Understanding Search Engine Optimization

IN THIS CHAPTER

- Learning how search engines see websites
- Taking a look at popularity in SEO
- Considering the role of relevancy in SEO

At Google, search engineers talk about "80-20" problems.

They are describing situations where the last 20 percent of the problem is 80 percent of the work. Learning SEO is one of these problems. Eighty percent of the knowledge SEOs need is available online for free. Unfortunately, the remaining 20 percent takes the majority of the time and energy to find and understand. My goal with this book is to solve this problem by making the last 20 percent as easy to get as the first 80 percent. Though I don't think I will be able to cover the entire 20 percent (some of it comes from years of practice), I am going to write as much actionable advanced material as humanly possible.

This book is for those who already know the basics of SEO and are looking to take their skills to the next level. Before diving in, try reading the following list:

- robots.txt
- sitemap
- nofollow
- 301 redirect
- canonicalization

If you are not sure what any of the items in this list are, you should go over to the nearest computer and read the article "The Beginner's Guide to SEO" at

```
http://www.seomoz.org/article/beginners-guide-to-search-engine-
optimization
```

This free article can teach you everything you need to know to use this book to its fullest. Done with that? Great, now we can begin.

THE SECRETS OF POPULARITY

Once upon a time there were two nerds at Stanford working on their PhDs. (Now that I think about it, there were probably a lot more than two nerds at Stanford.) Two of the nerds at Stanford were not satisfied with the current options for searching online, so they attempted to develop a better way.

Being long-time academics, they eventually decided to take the way academic papers were organized and apply that to webpages. A quick and fairly objective way to judge the quality of an academic paper is to see how many times other academic papers have cited it. This concept was easy to replicate online because the original purpose of the Internet was to share academic resources between universities. The citations manifested themselves as hyperlinks once they went online. One of the nerds came up with an algorithm for calculating these values on a global scale, and they both lived happily ever after.

Of course, these two nerds were Larry Page and Sergey Brin, the founders of Google, and the algorithm that Larry invented that day was what eventually became PageRank. Long story short, Google ended up becoming a big deal and now the two founders rent an airstrip from NASA so they have somewhere to land their private jets. (Think I am kidding? See `http://searchengineland.com/your-guide-to-the-google-jet-12161`.)

RELEVANCE, SPEED, AND SCALABILITY

Hypothetically, the most relevant search engine would have a team of experts on every subject in the entire world—a staff large enough to read, study, and evaluate every document published on the web so they could return the most accurate results for each query submitted by users.

The fastest search engine, on the other hand, would crawl a new URL the very second it's published and introduce it into the general index immediately, available to appear in query results only seconds after it goes live.

The challenge for Google and all other engines is to find the balance between those two scenarios: To combine rapid crawling and indexing with a relevance algorithm that can be instantly applied to new content. In other words, they're trying to build *scalable relevance*. With very few exceptions, Google is uninterested in hand-removing (or hand-promoting) specific content. Instead, its model is built around identifying characteristics in web content that indicate the content is especially relevant or irrelevant, so that content all across the web with those same characteristics can be similarly promoted or demoted.

This book frequently discusses the benefits of content created with the user in mind. To some hardcore SEOs, Google's "think about the user" mantra is corny; they'd much prefer to know a secret line of code or server technique that bypasses the intent of creating engaging content.

While it may be corny, Google's focus on creating relevant, user-focused content really is the key to its algorithm of scalable relevance. Google is constantly trying to find ways to reward content that truly answers users' questions and ways to minimize or filter out content built for content's sake. While this book discusses techniques for making your content visible and accessible to engines, remember that means talking about content constructed with users in mind, designed to be innovative, helpful, and to serve the query intent of human users. It might be corny, but it's effective.

That fateful day, the Google Guys capitalized on the mysterious power of links. Although a webmaster can easily manipulate everything (word choice, keyword placement, internal links, and so on) on his or her own website, it is much more difficult to influence inbound links. This natural link profile acts as an extremely good metric for identifying legitimately popular pages.

NOTE Google's PageRank was actually named after its creator, Larry Page. Originally, the algorithm was named BackRub after its emphasis on backlinks. Later, its name was changed to PageRank because of its connections to Larry Page's last name and the ability for the algorithm to rank pages.

Larry Page's original paper on PageRank, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," is still available online. If you are interested in reading it, it is available on Stanford's website at `http://infolab.stanford.edu/~backrub/google.html`. It is highly technical, and I have used it on more than one occasion as a sleep aid. It's worth noting that the original PageRank as described in this paper is only a tiny part of Google's modern-day search algorithm.

Now wait a second—isn't this supposed to be a book for advanced SEOs? Then why am I explaining to you the value of links? Relax, there is a method to my madness. Before I am able to explain the more advanced secrets, I need to make sure we are on the same page.

As modern search engines evolved, they started to take into account the link profile of both a given page and its domain. They found out that the relationship between these two indicators was itself a very useful metric for ranking webpages.

Domain and Page Popularity

There are hundreds of factors that help engines decide how to rank a page. And in general, those hundreds of factors can be broken into two categories—relevance and popularity (or "authority"). For the purposes of this demonstration you will need to completely ignore relevancy for a second. (Kind of like the search engine Ask.com.) Further, within the category of popularity, there are two primary types—domain popularity and page popularity. Modern search engines rank pages by a combination of these two kinds of popularity metrics. These metrics are measurements of link profiles. To rank number one for a given query you need to have the highest amount of total popularity on the Internet. (Again, bear with me as we ignore relevancy for this section.)

This is very clear if you start looking for patterns in search result pages. Have you ever noticed that popular domains like `Wikipedia.org` tend to rank for everything? This is because they have an enormous amount of domain popularity. But what about those competitors who outrank me for a specific term with a practically unknown domain? This happens when they have an excess of page popularity. See Figure 1-1.

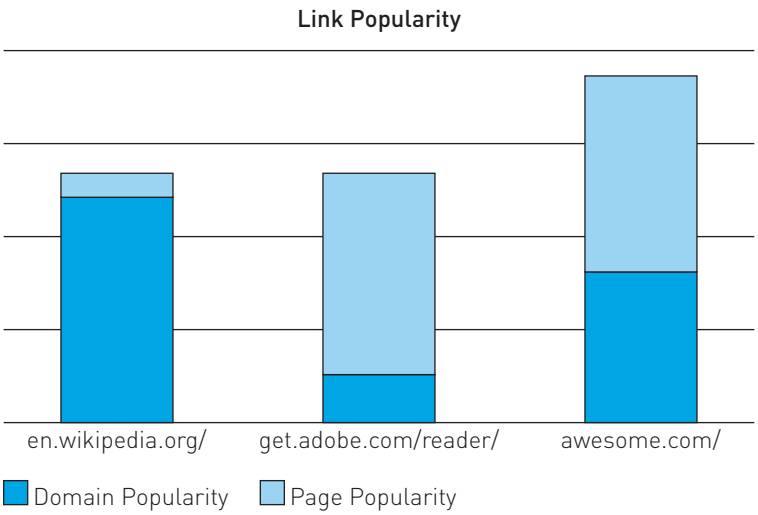

FIGURE 1-1: Graph showing different combinations of relevancy and popularity metrics that can be used to achieve high rankings

Although `en.wikipedia.org` has a lot of domain popularity and `get.adobe.com/reader/` has a lot of page popularity, `www.awesome.com` ranks higher because it has a higher *total* amount of popularity. This fact and relevancy metrics (discussed later in this chapter) are the essence of Search Engine Optimization. (Shoot! I unveiled it in the first chapter, now what am I going to write about?)

POPULARITY TOP TEN LISTS

The top 10 most linked-to domains on the Internet (at the time of writing) are:

- `Google.com`
- `Adobe.com`
- `Yahoo.com`
- `Blogspot.com`
- `Wikipedia.org`
- `YouTube.com`
- `W3.org`
- `Myspace.com`

continued

(continued)

- Wordpress.com
- Microsoft.com

The top 10 most linked-to pages on the Internet (at the time of writing) are:

- http://wordpress.org/
- http://www.google.com/
- http://www.adobe.com/products/acrobat/readstep2.html
- http://www.miibeian.gov.cn/
- http://validator.w3.org/check/referer
- http://www.statcounter.com/
- http://jigsaw.w3.org/css-validator/check/referer
- http://www.phpbb.com/
- http://www.yahoo.com/
- http://del.icio.us/post

Source: SEOmoz's Linkscape—Index of the World Wide Web

▶ Not only that, but at any given time, the TbPR (Toolbar PageRank) value you see may be up to 60–90 days older or more, and it's a single-digit representation of what's probably very a long decimal value.

Before I summarize I would like to nip the PageRank discussion in the bud. Google releases its PageRank metric through a browser toolbar. This is not the droid you are looking for. That green bar represents only a very small part of the overall search algorithm.

Just because a page has a PageRank of 5 does not mean it will outrank all pages with a PageRank of 4. Keep in mind that major search engines do not want you to reverse engineer their algorithms. As such, publicly releasing a definitive metric for ranking would be idiotic from a business perspective. If there is one thing that Google is not, it's idiotic.

▶ Google makes scraping (automatically requesting and distributing) its PageRank metric difficult. To get around the limitations, you need to write a program that requests the metric from Google and identifies itself as the Google Toolbar.

In my opinion, hyperlinks are the most important factor when it comes to ranking web pages. This is the result of them being difficult to manipulate. Modern search engines look at link profiles from many different perspectives and use those relationships to determine rank. The takeaway for you is that time spent earning links is time well spent. In the same way that a rising tide raises all ships, popular domains raise all pages. Likewise, popular pages raise the given domain metrics.

In the next section I want you to take a look into the pesky missing puzzle piece of this chapter: relevancy. I am going to discuss how it interacts with popularity, and I may or may not tell you another fairy tale.

THE SECRETS OF RELEVANCY

In the previous section, I discussed how popular pages (as judged by links) rank higher. By this logic, you might expect that the Internet's most popular pages would rank for everything. To a certain extent they do (think Wikipedia!), but the reason they don't dominate the rankings for *every* search result page is that search engines put a lot of emphasis on determining relevancy.

Text Is the Currency of the Internet

Relevancy is the measurement of the theoretical distance between two corresponding items with regards to relationship. Luckily for Google and Microsoft, modern-day computers are quite good at calculating this measurement for text.

By my estimations, Google owns and operates well over a million servers. The electricity to power these servers is likely one of Google's larger operating expenses. This energy limitation has helped shape modern search engines by putting text analysis at the forefront of search. Quite simply, it takes less computing power and is much simpler programmatically to determine relevancy between a text query and a text document than it is between a text query and an image or video file. This is the reason why text results are so much more prominent in search results than videos and images.

As of this writing, the most recent time that Google publicly released the size of its indices was in 2006. At that time it released the numbers shown in Table 1-1.

TABLE 1-1: Size of Google Indices

DATA	SIZE IN TERABYTES
Crawl Index	800
Google Analytics	200
Google Base	2
Google Earth	70
Orkut	9
Personalized Search	4

This is especially true until Google finds better ways to interpret and grade non-textual media

So what does this emphasis on textual content mean for SEOs? To me, it indicates that my time is better spent optimizing text than images or videos. This strategy will likely have to change in the future as computers get more powerful and energy efficient, but for right now text should be every SEO's primary focus.

But Why Content?

The most basic structure a functional website could take would be a blank page with a URL. For example purposes, pretend your blank page is on the fake domain `www.WhatIsJessicaSimpsonThinking.com`. (Get it? It is a blank page.) Unfortunately for the search engines, clues like top-level domains (`.com`, `.org`, and so on), domain owners (WHOIS records), code validation, and copyright dates are poor signals for determining relevancy. This means your page with the dumb domain name needs some content before it is able to rank in search engines.

The search engines must use their analysis of content as their primary indication of relevancy for determining rankings for a given search query. For SEOs, this means the content on a given page is essential for manipulating—that is, earning—rankings. In the old days of AltaVista and other search engines, SEOs would just need to write "Jessica Simpson" hundreds times on the site to make it rank #1 for that query. What could be more relevant for the query "Jessica Simpson" than a page that says Jessica Simpson 100 times? (Clever SEOs will realize the answer is a page that says "Jessica Simpson" 101 times.) This metric, called *keyword density*, was quickly manipulated, and the search engines of the time diluted the power of this metric on rankings until it became almost useless. Similar dilution has happened to the keywords meta tag, some kinds of internal links, and H1 tags.

Despite being more sophisticated, modern-day search engines still work essentially the same way they did in the past—by analyzing content on the page.

Hey, Ben Stein, thanks for the history lesson, but how does this apply to modern search engines? The funny thing is that modern-day search engines still work essentially the same way they did back in the time of keyword density. The big difference is that they are now much more sophisticated. Instead of simply counting the number of times a word or phrase is on a webpage, they use natural language processing algorithms and other signals on a page to determine relevancy. For example, it is now fairly trivial for search engines to determine that a piece of content is about Jessica Simpson if it mentions related phrases like "Nick Lachey" (her ex-husband), "Ashlee Simpson" (her sister), and "Chicken of the Sea" (she is infamous for thinking the tuna brand "Chicken of the Sea" was made from chicken). The engines can do this for a multitude of languages and with astonishing accuracy.

Don't believe me? Try going to Google right now and searching `related:www.jessicasimpson.com`. If your results are like mine, you will see websites about her movies, songs, and sister. Computers are amazing things.

In addition to the words on a page, search engines use signals like image meta information (alt attribute), link profile and site architecture, and information hierarchy to determine how relevant a given page that mentions "Jessica" is to a search query for "The Simpsons."

Link Relevancy

As search engines matured, they started identifying more metrics for determining rankings. One that stood out among the rest was link relevancy.

The difference between link relevancy and link popularity (discussed in the previous section) is that link relevancy does not take into account the power of the link. Instead, it is a natural phenomenon that works when people link out to other content.

Let me give you an example of how it works. Say I own a blog where I write about whiteboard markers. (Yes, I did just look around my office for an example to use, and yes, there are actually people who blog about whiteboard markers. I checked.) Ever inclined to learn more about my passion for these magical writing utensils, I spend part of my day reading online what other people have to say about whiteboard markers.

On my hypothetical online reading journey, I find an article about the psychological effects of marker color choice. Excited, I go back to my website to blog about the article so (both of) my friends can read about it. Now here is the critical takeaway. When I write the blog post and link to the article, *I get to choose the anchor text*. I could choose something like "click here," but more likely I choose something that it is relevant to the article. In this case I choose "psychological effects of marker color choice." Someone else who links to the article might use the link anchor text "marker color choice and the effect on the brain."

People have a tendency to link to content using the anchor text of either the domain name or the title of the page. Use this to your advantage by including keywords you want to rank for in these two elements.

This human-powered information is essential to modern-day search engines. These descriptions are relatively unbiased and produced by real people. This metric, in combination with complicated natural language processing, makes up the lion's share of relevancy indicators online.

Other important relevancy indicators are link sources and information hierarchy. For example, the search engines can also use the fact that I linked to the color choice article from a blog about whiteboard markers to supplement their understanding of relevancy. Similarly, they can use the fact that the original article was located at the URL `www.example.com/vision/color/` to determine the high-level positioning and relevancy of the content. As you read later in this book (Chapter 2 specifically), these secrets are essential for SEOs to do their job.

Beyond specific anchor text, proximal text—the certain number of characters preceding and following the link itself—have some value. Something that's logical,

but annoying is when people use a verb as anchor text, such as "Frank said . . . " or "Jennifer wrote . . .", using "said" or "wrote" as the anchor text pointing back to the post. In a situation like that, engines have figured out how to apply the context of the surrounding copy to the link.

Tying Together Popularity and Relevancy

So far in this chapter I have discussed both popularity and relevancy. These two concepts make up the bulk of Search Engine Optimization theory. They have been present since the beginning of search engines and undoubtedly will be important in the future. The way they are determined and the relationship between them changes, but they are both fundamental to determining search results.

▶ Popularity and relevancy are the two concepts that make up the bulk of Search Engine Optimization theory.

This fact is critical to SEOs. We have very little control over how the major search engines operate, yet somehow we are supposed to keep our jobs. Luckily, these immutable laws of popularity and relevance govern search engines and provide us with some job security.

SUMMARY

In this chapter, I explained the concepts of popularity and relevancy in relation to modern search engines. This information, along with your prior SEO experience, will make up the foundation for all of the SEO secrets and knowledge that you learn throughout the rest of the book. You no doubt have some questions. I'll start answering many of your questions in the next chapter, but you will likely form many more. Welcome to the mindset of a Professional SEO. Prepare to be questioning and Googling things for the rest of your life.