

CHAPTER TWO

Search Engine Basics

In this chapter, we will begin to explore how search engines work. Building a strong foundation on this topic is essential to understanding the SEO practitioner’s craft.

As we discussed in [Chapter 1](#), people have become accustomed to receiving nearly instantaneous answers from search engines after they have submitted a search query. In [Chapter 1](#) we also discussed the volume of queries (more than 7,500 per second). As early as 2008, Google knew about 1 trillion pages on the Web.¹ At SMX Advanced in Seattle in 2014, Google’s Gary Illyes stated that Google now knows about 30,000 trillion pages on the Web. The scale of the Internet/Web (sometimes called the Interwebs) is growing fast!

Underlying the enormous problem of processing all these pages is the complex nature of the Web itself. Web pages include text, video, images, and more. It’s easy for humans to understand these and to transition seamlessly between them, but software lacks the intelligence we take for granted. This limitation and others affect how search engines understand the web pages they come across. We’ll discuss some of these limitations in this chapter.

Of course, this is an ever-changing landscape. The search engines continuously invest in improving their ability to process the content of web pages. For example, advances in image and video search have enabled search engines to inch closer to human-like understanding, a topic that will be explored more in the section [“Vertical Search Engines” on page 122](#).

¹ Google Official Blog, “We Knew the Web Was Big...”, July 25, 2008, <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>.

Understanding Search Engine Results

In the search marketing field, the pages the engines return to fulfill a query are referred to as *search engine results pages* (SERPs). Each engine returns results in a slightly different format, and these may include *vertical results*—results that can be derived from different data sources or presented on the results page in a different format, which we'll illustrate shortly.

Understanding the Layout of Search Results Pages

Figure 2-1 shows the SERPs in Google for the query *stuffed animals*.

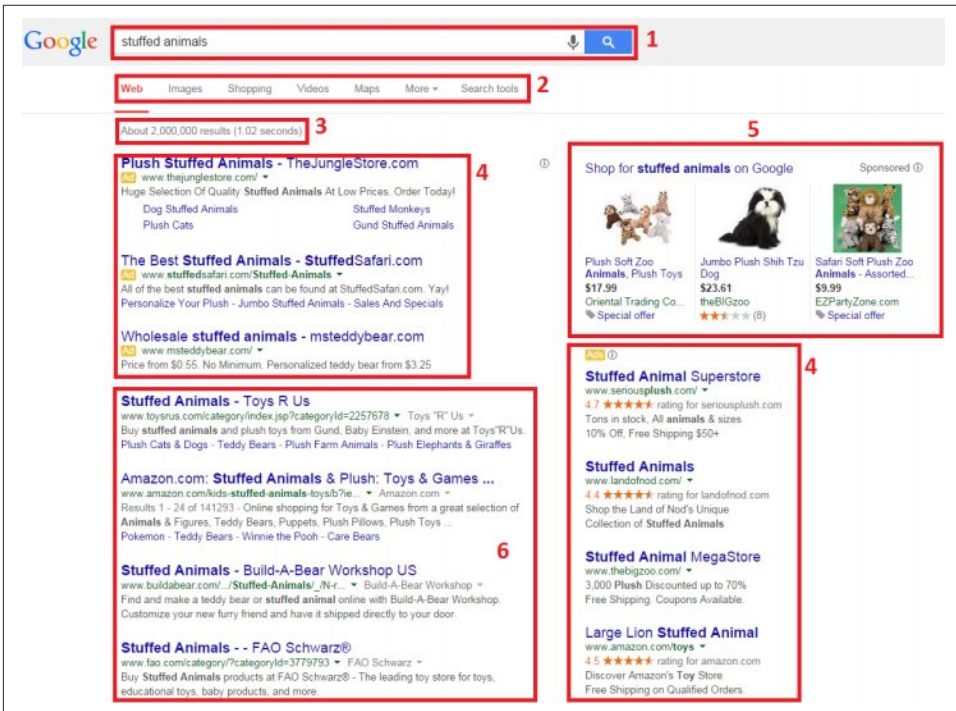


Figure 2-1. Layout of Google search results

The various sections outlined in the Google search results are as follows:

- Search query box (1)
- Vertical navigation (2)
- Results information (3)
- PPC advertising (4)
- Google product search results (5)

- [Natural/organic/algorithmic results \(6\)](#)

Even though Yahoo! no longer does its own crawl of the Web or provides its own search results information (it sources them from Bing), it does format the output differently. [Figure 2-2](#) shows Yahoo!'s results for the same query.

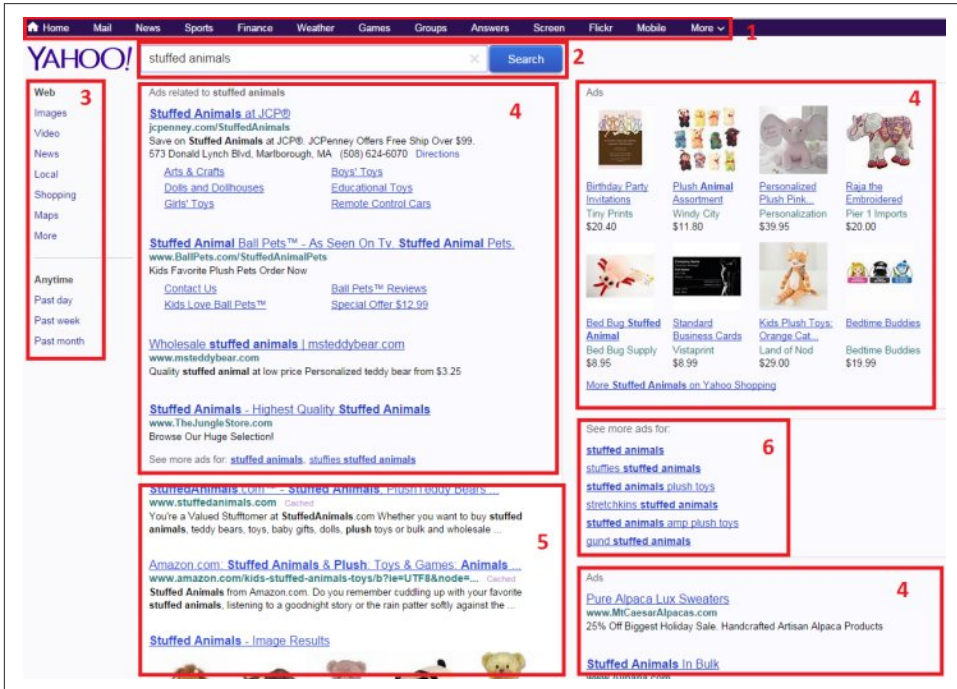


Figure 2-2. Layout of Yahoo! search results

The sections in the Yahoo! results are as follows:

- Vertical navigation (1)
- Search query box (2)
- Horizontal navigation (3)
- PPC advertising (4)
- Natural/organic/algorithmic results (5)
- Navigation to more advertising (6)

[Figure 2-3](#) shows the layout of the results from Microsoft's Bing for *stuffed animals*.

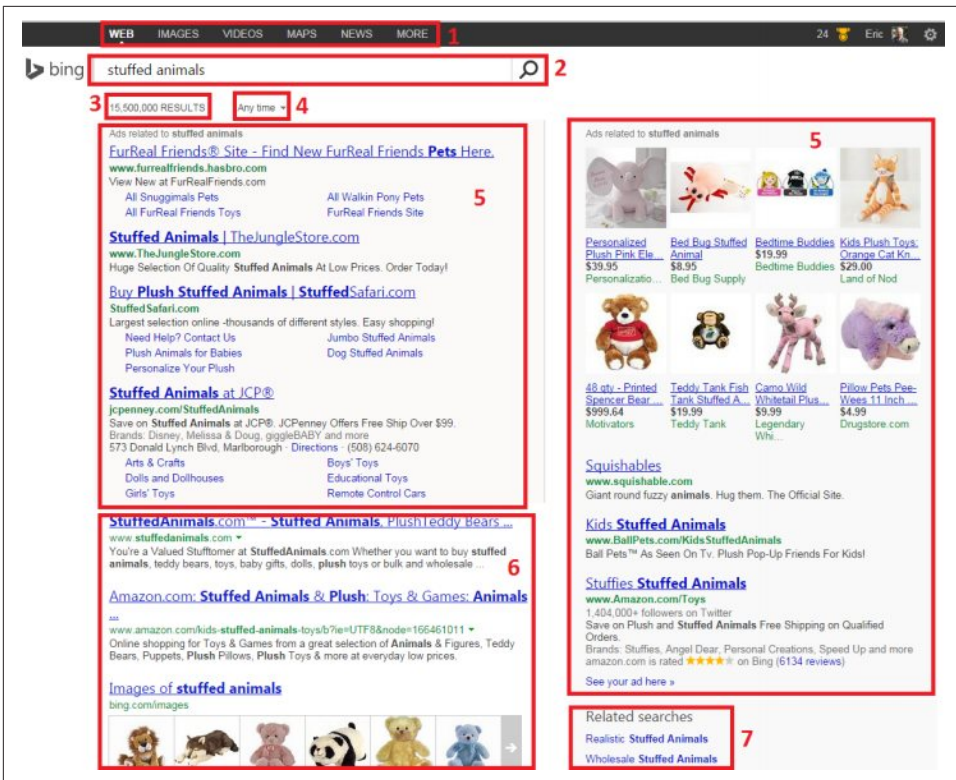


Figure 2-3. Layout of Bing search results

The sections in Bing’s search results are as follows:

- Vertical navigation (1)
- Search query box (2)
- Results information (3)
- Time-based refinement options (4)
- PPC advertising (5)
- Natural/organic/algorithmic results (6)
- Query refinement options (7)

Each unique section represents a snippet of information provided by the engines. Here are the definitions of what each piece is meant to provide:

Vertical navigation

Each engine offers the option to search different verticals, such as images, news, video, or maps. Following these links will result in a query with a more limited

[index](#). In [Figure 2-3](#), for example, you might be able to see news items about stuffed animals or videos featuring stuffed animals.

Horizontal navigation

All three engines used to have some form of horizontal navigation, but as of June 2015 only Yahoo! continues to include it.

Search query box

All of the engines show the query you've performed and [allow you to edit or reenter a new query](#) from the search results page. If you begin typing, you may notice that [Google gives you a list of suggested searches](#) below. This is the Google autocomplete suggestions feature, and it can be incredibly useful for targeting keywords. Next to the search query box, the engines also offer links to the advanced search page, the features of which we'll discuss later in the book. In addition, you will also see a [microphone icon in the right of the search box](#) that allows you to speak your query. In [Google image search](#), this shows up as a camera icon that allows you to upload an image or get similar images back.

Results information

This section provides a [small amount of meta-information about the results](#) that you're viewing, including an estimate of the number of pages relevant to that particular query (these numbers can be, and frequently are, [wildly inaccurate and should be used only as a rough comparative measure](#)).

PPC (a.k.a. paid search) advertising

The text ads are [purchased by companies that use either Google AdWords or Bing](#). The results are ordered by a variety of factors, including relevance (for which click-through rate, use of searched keywords in the ad, and relevance of the landing page are factors in Google) and bid amount (the ads require a maximum bid, which is then compared against other advertisers' bids).

Natural/organic/algorithmic results

These results are [pulled from the search engines' primary indices of the Web and ranked in order of relevance and importance](#) according to their complex algorithms. This area of the results is the primary focus of this section of the book.

Query refinement suggestions

Query refinements are offered by Google, Bing, and Yahoo!. The goal of these links is to [let users search with a more specific and possibly more relevant query](#) that will satisfy their intent.

In March 2009, Google enhanced the refinements by implementing Orion Technology, based on technology Google acquired in 2006. The goal of this enhancement is to provide a wider array of refinement choices. For example, a search on

principles of physics may display refinements for the Big Bang, angular momentum, quantum physics, and special relativity.

Navigation to more advertising

Only Yahoo! shows this in the search results. Clicking on these links will bring you to additional paid search results related to the original query.

Be aware that the SERPs are always changing as the engines test new formats and layouts. Thus, the images in **Figure 2-1** through **Figure 2-3** may be accurate for only a few weeks or months until Google, Yahoo!, and Bing shift to new formats.

Understanding How Vertical Results Fit into the SERPs

These “standard” results, however, are certainly not all that the engines have to offer. For many types of queries, search engines show *vertical* results, or *instant answers*, and include more than just links to other sites to help answer a user’s questions. These types of results present many additional challenges and opportunities for the SEO practitioner.

Figure 2-4 shows an example of these types of results. The query in **Figure 2-4** brings back a business listing showing an address and the option to get directions. This result attempts to provide the user with the answer he is seeking directly in the search results.

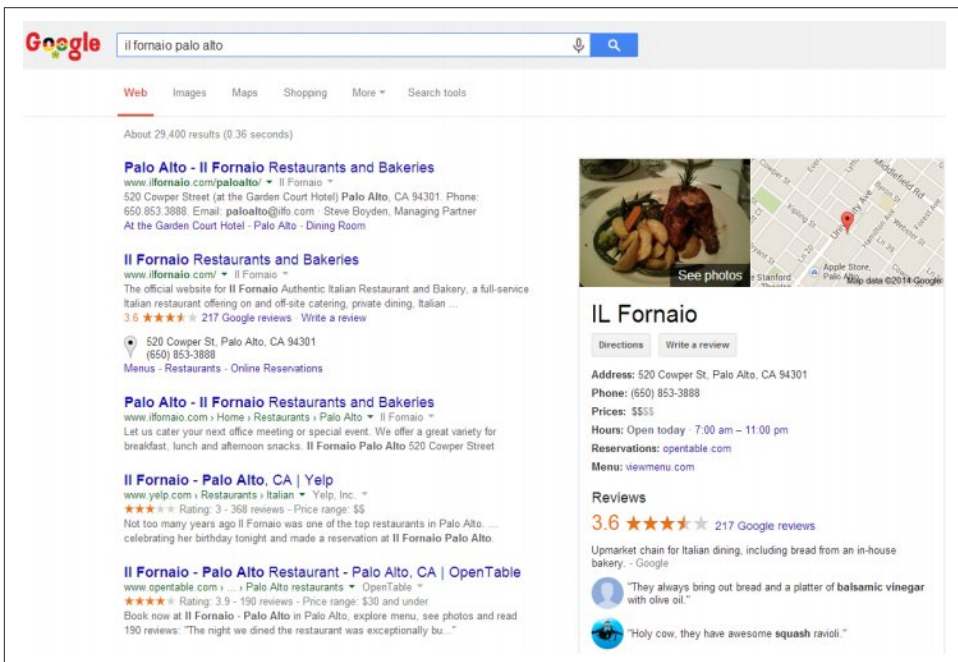


Figure 2-4. Local search result for a business

Figure 2-5 shows another example. The Google search in Figure 2-5 for weather plus a city name returns a direct answer. Once again, the user may not even need to click on a website if all she wanted to know was the temperature.

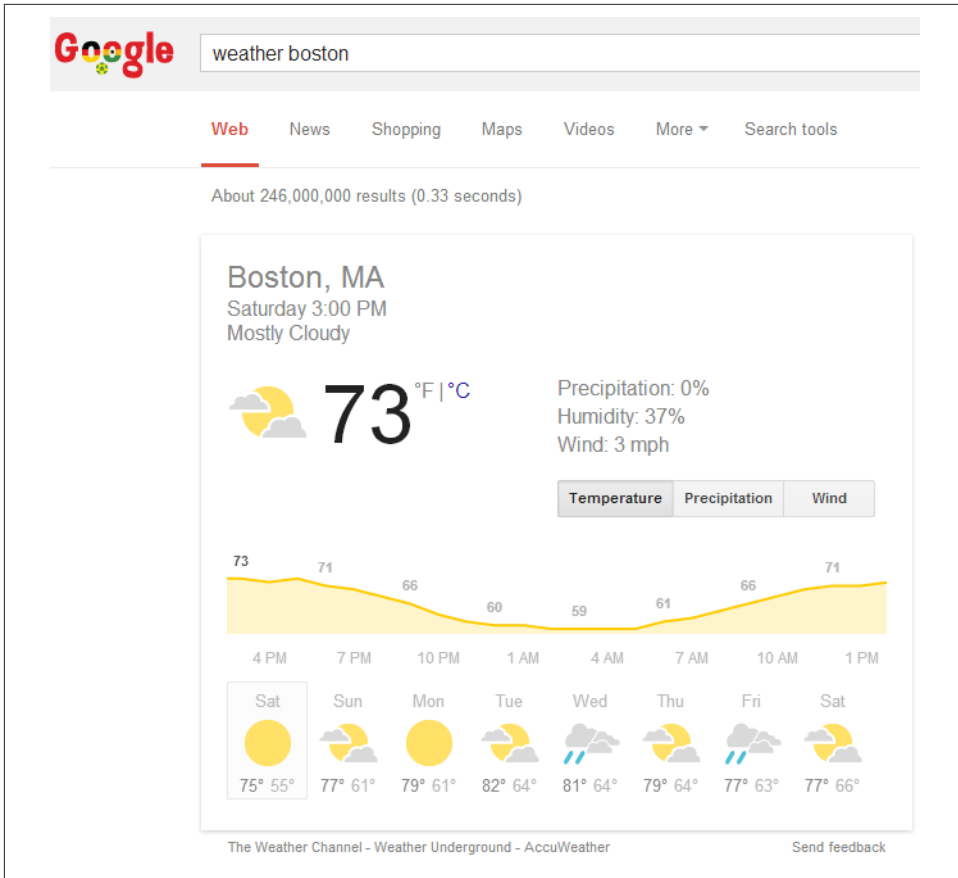


Figure 2-5. Weather search on Google

Figure 2-6 is an example of a search for a well-known painter. A Google search for the famous painter *Edward Hopper* returns image results of some of his most memorable works (shown in the lower-right of the screenshot). This example is a little different from the “instant answers” type of result shown in Figure 2-4 and Figure 2-5. If the user is interested in the first painting shown, he may well click on it to see the painting in a larger size or to get more information about it. For the SEO practitioner, getting placed in this vertical result could be a significant win.

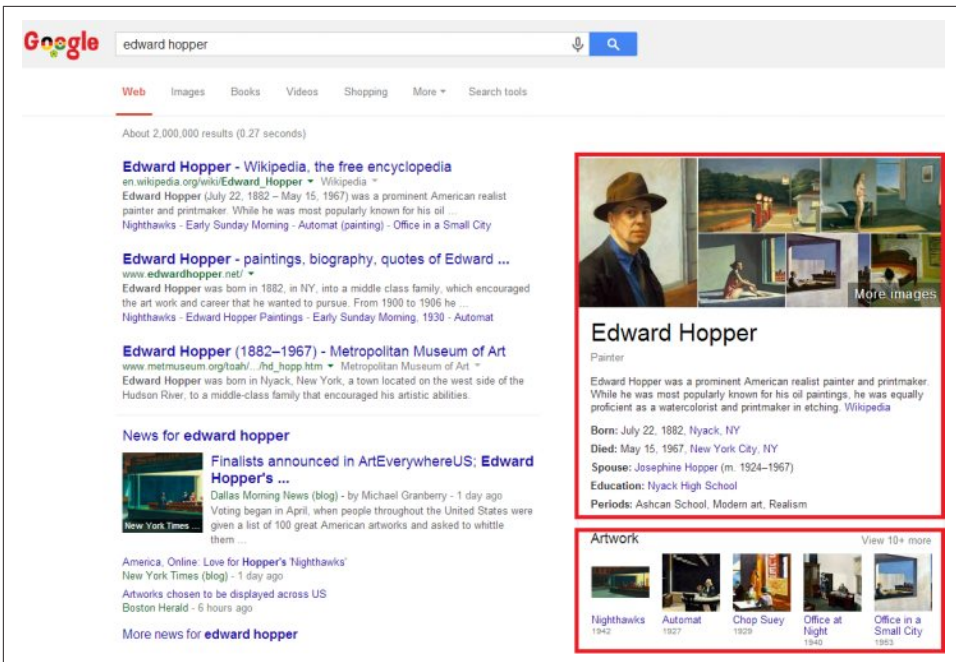


Figure 2-6. Google search on an artist's name

Figure 2-7 shows an example from Yahoo!. A query on Yahoo! for *chicago restaurants* brings back a list of popular dining establishments from Yahoo!'s local portal. High placement in these results has likely been a good thing for Lou Malnati's Pizzeria.

Figure 2-8 is an example of a celebrity search on Bing.

The results in Figure 2-8 include a series of images of the famous actor Charlie Chaplin. As a last example, Figure 2-9 is a look at the Bing search results for videos with Megan Fox.

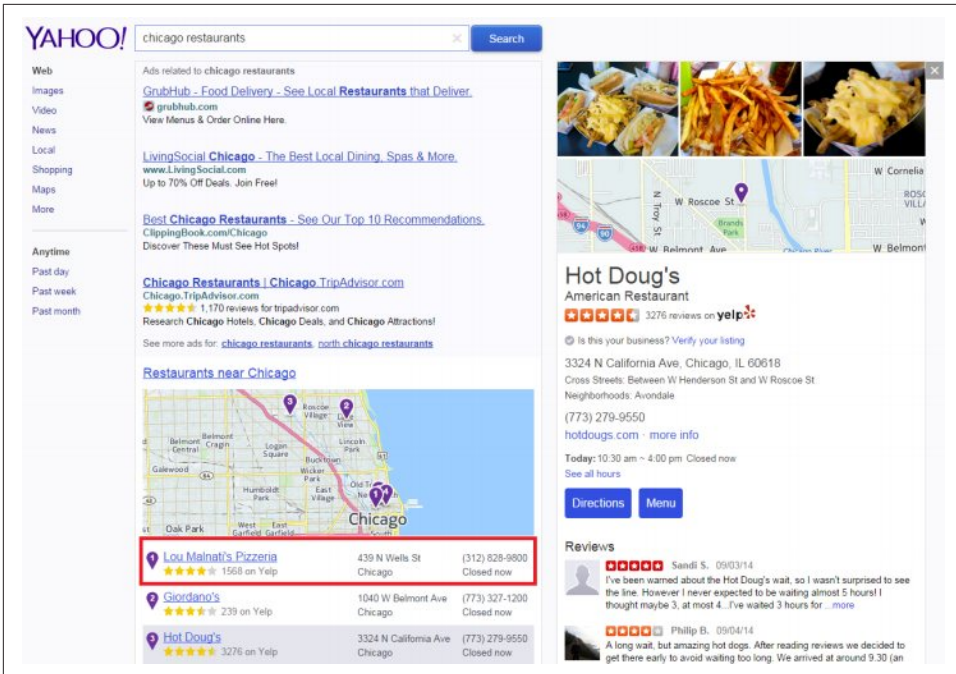


Figure 2-7. Yahoo! search for Chicago restaurants

At the top of the search results in Figure 2-9, you're provided with a series of popular videos. Click on a video in the results, and it begins playing right there in the search results.

As you can see, the vast variety of vertical integration into search results means that for many popular queries you can expect to receive significant amounts of information in the SERPs themselves. Engines are competing by providing more relevant results and more targeted responses to queries that they feel are best answered by vertical results, rather than web results.

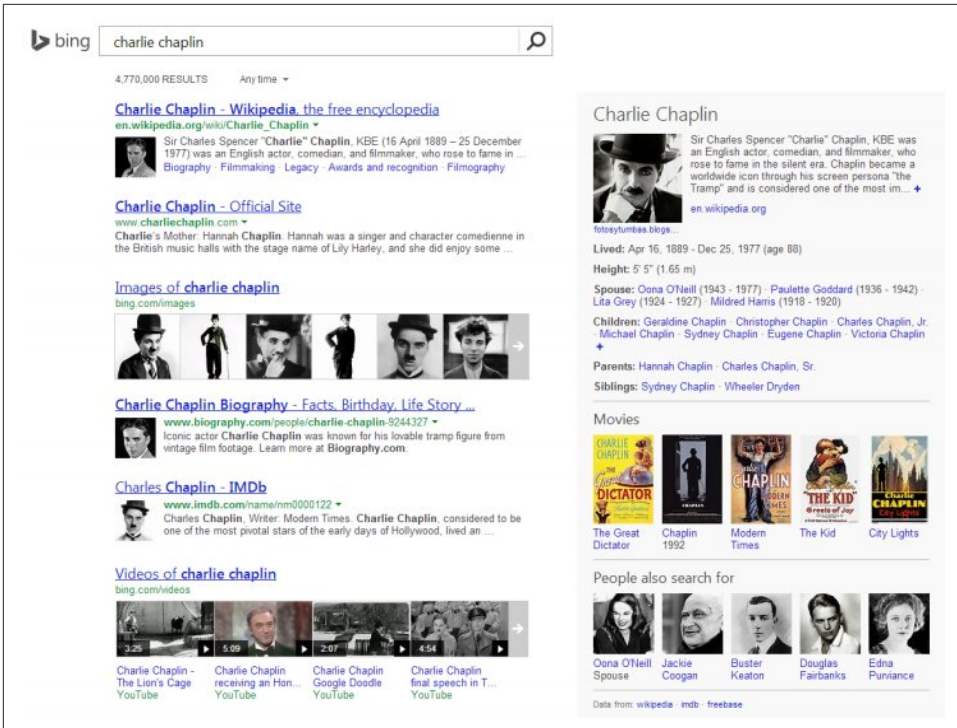


Figure 2-8. Bing result for Charlie Chaplin

As a direct consequence, site owners and web marketers must take into account how this incorporation of vertical search results may impact their rankings and traffic. For many of the searches shown in the previous figures, a high ranking—even in position #1 or #2 in the algorithmic/organic results—may not produce much traffic because of the presentation of the vertical results above them.

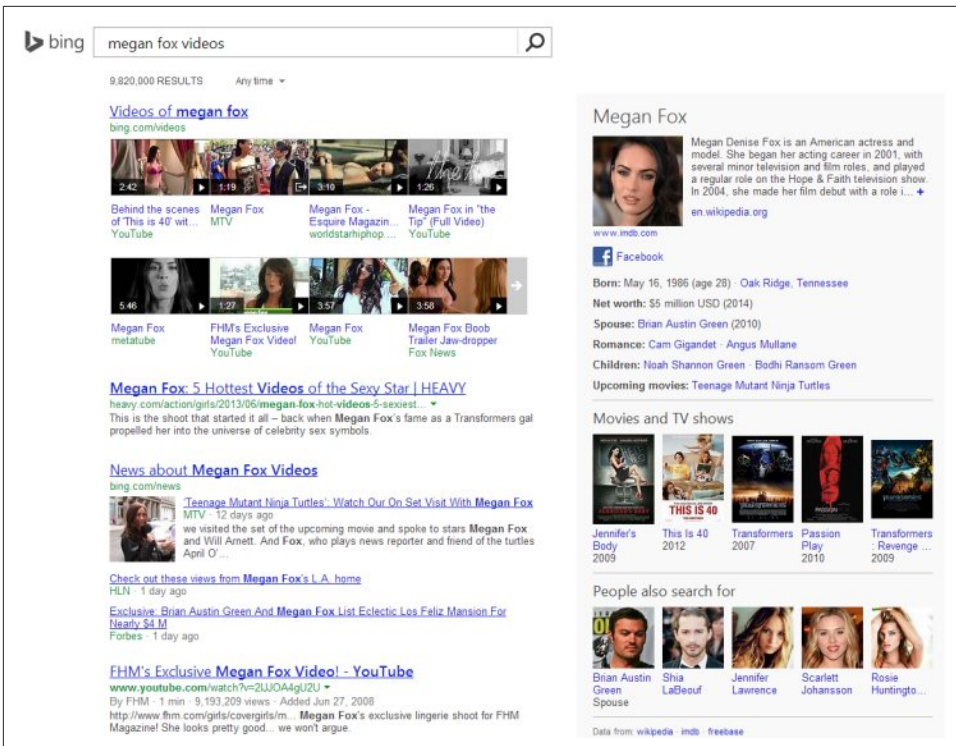


Figure 2-9. Bing result for Megan Fox videos

The vertical results also signify an opportunity, as listings are available in services from images to local search to news and products. We will cover how to get included in these results in [Chapter 11](#).

Google's Knowledge Graph

The search engines are actively building structured databases of information that allow them to show answers to questions that are not simply links to web pages. In

[Figure 2-6](#), the information on the upper right is an example of this. Google provides direct answers in the result, including Edward Hopper's birth date, place of birth, the date and place of his death, his spouse, and more. In [Figure 2-9](#), Bing provides similar information for Megan Fox.

Not only is additional information shown, but it is not just a data dump: it shows that the search engines are working to develop their own knowledge of the relationships between people and things. In the case of [Figure 2-6](#), we can see that Google understands that:

- Edward Hopper is the name of a person.
- People have dates and places of birth.
- People have dates and places of death.
- People might have spouses.

The search engines are actively mapping these types of relationships as part of their effort to offer more complete information directly in the search results themselves.

Algorithm-Based Ranking Systems: Crawling, Indexing, and Ranking

Understanding how crawling, indexing, and ranking works is useful to SEO practitioners, as it helps them determine what actions to take to meet their goals. This section primarily covers the way Google and Bing operate, and does not necessarily apply to other search engines that are popular in other countries, such as Yandex (Russia), Baidu (China), Seznam (Czech Republic), and Naver (Korea).

The search engines must execute many tasks very well to provide relevant search results. Put simplistically, you can think of these as:

- Crawling and indexing trillions of documents (pages and files) on the Web (note that they ignore pages that they consider to be “insignificant,” perhaps because the pages are perceived as adding no new value or are not referenced at all on the Web).
- Responding to user queries by providing lists of relevant pages.

In this section, we’ll walk through the basics of these functions from a nontechnical perspective. This section will start by discussing how search engines find and discover content.

Crawling and Indexing

To offer the best possible results, search engines must attempt to discover all the public pages on the World Wide Web and then present the ones that best match up with the user’s search query. The first step in this process is crawling the Web. The search engines start with a seed set of sites that are known to be very high quality, and then visit the links on each page of those sites to discover other web pages.

The link structure of the Web serves to bind together all of the pages that were made public as a result of someone linking to them. Through links, search engines’ automated robots, called *crawlers* or *spiders*, can reach the many trillions of interconnected documents.

In [Figure 2-10](#), you can see the home page of [USA.gov](#), the official U.S. government website. The links on the page are outlined in red. Crawling this page would start with loading the page, analyzing the content, and then seeing what other pages [USA.gov](#) links to.

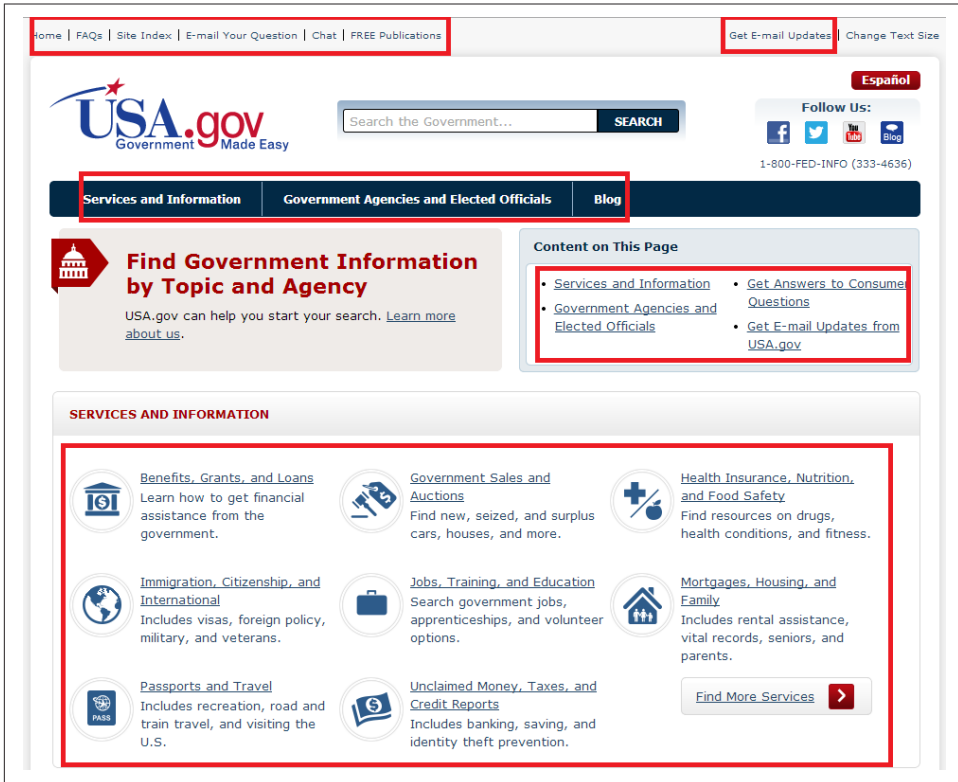


Figure 2-10. *Crawling the U.S. government website*

The search engine would then load those other pages and analyze that content as well. This process repeats over and over again until the crawling process is complete. This process is enormously complex, as the Web is a large and complex place.

— NOTE —

Search engines do not attempt to crawl the entire Web every day. In fact, they may become aware of pages that they choose not to crawl because those pages are not likely to be important enough to return in a search result. We will discuss the role of *importance* in “[Retrieval and Ranking](#)” on [page 80](#).

The first step in this process is to build an *index* of terms. This is a massive database that catalogs all the significant terms on each page crawled by the search engine.

A lot of other data is also recorded, such as a map of all the pages that each page links to, the clickable text of those links (known as the *anchor text*), whether or not those links are considered ads, and more.

To accomplish the monumental task of holding data on hundreds of trillions of pages that can be accessed in a fraction of a second, the search engines have constructed massive data centers to deal with all this data.

One key concept in building a search engine is deciding where to begin a crawl of the Web. Although you could theoretically start from many different places on the Web, you would ideally begin your crawl with a trusted seed set of websites.

Starting with a known, trusted set of websites enables search engines to measure how much they trust the other websites that they find through the crawling process. We will discuss the role of trust in search algorithms in more detail in [“How Links Historically Influenced Search Engine Rankings”](#) on page 421.

Retrieval and Ranking

For most searchers, the quest for an answer begins as shown in [Figure 2-11](#).

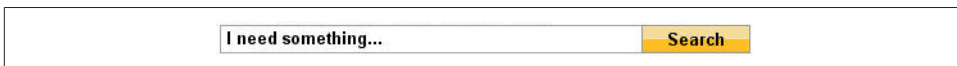


Figure 2-11. [Start of a user's search quest](#)

The next step in this quest occurs when the search engine returns a list of relevant pages on the Web in the order it believes is most likely to satisfy the user. This process requires the search engines to scour their corpus of hundreds of billions of documents and do two things: first, return only the results that are related to the searcher's query; and second, rank the results in order of perceived importance (taking into account the trust and authority associated with the site). It is both relevance and importance that the process of SEO is meant to influence.

Relevance is the degree to which the content of the documents returned in a search matches the user's query intention and terms. The relevance of a document increases if the page contains terms relevant to the phrase queried by the user, or if links to the page come from relevant pages and use relevant anchor text.

You can think of relevance as the first step to being “in the game.” If you are not relevant to a query, the search engine does not consider you for inclusion in the search results for that query. We will discuss how relevance is determined in more detail in [“Determining Searcher Intent and Delivering Relevant, Fresh Content”](#) on page 92.

Importance refers to the relative importance, measured via citation (the act of one work referencing another, as often occurs in academic and business documents), of a given document that matches the user's query. The importance of a given document increa-

ses with every other document that references it. In today's online environment, citations can come in the form of links to the document or references to it on social media sites. Determining how to weight these signals is known as *citation analysis*.

You can think of importance as a way to determine which page, from a group of equally relevant pages, shows up first in the search results, which is second, and so forth. The relative authority of the site, and the trust the search engine has in it, are significant parts of this determination. Of course, the equation is a bit more complex than this, and not all pages are equally relevant. Ultimately, it is the combination of relevance and importance that determines the ranking order.

So, when you see a search results page such as the one shown in [Figure 2-12](#), you can surmise that the search engine (in this case, Bing) believes [the Superhero Stamps page on eBay](#) has the highest combined score for relevance and importance for the query *marvel superhero stamps*.

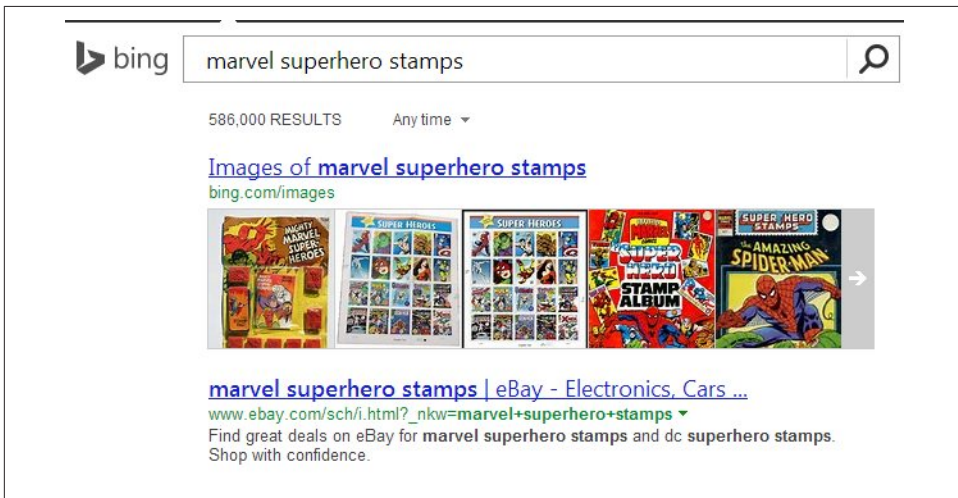


Figure 2-12. Sample search result for “marvel superhero stamps”

Importance and relevance aren't determined manually (those trillions of man-hours would require Earth's entire population as a workforce). Instead, the engines craft careful, mathematical equations—*algorithms*—to sort the wheat from the chaff and then rank the wheat in order of quality. These algorithms often comprise hundreds of components. In the search marketing field, they are often referred to as *ranking factors* or *algorithmic ranking criteria*.

We discuss ranking factors or signals (*signals* is the term Google prefers) in more detail in [“Analyzing Ranking Factors” on page 108](#).

Evaluating Content on a Web Page

Search engines place a lot of weight on the content of each web page. After all, it is this content that defines what a page is about, and the search engines do a detailed analysis of each web page they find during their crawl to help make that determination.

You can think of this as the search engine performing a detailed analysis of all the words and phrases that appear on a web page, and then building a map of that data for it to consider showing your page in the results when a user enters a related search query. This map, often referred to as a *semantic map*, seeks to define the relationships between those concepts so that the search engine can better understand how to match the right web pages with user search queries.

If there is no semantic match of the content of a web page to the query, the page has a much lower possibility of showing up. Therefore, the words you put on the page, and the “theme” of that page, play a huge role in ranking.

Figure 2-13 shows how a search engine will break up a page when it looks at it, using a page on the *Forbes* website.

The navigational elements of a web page are likely similar across the many pages of a site. These navigational elements are not ignored, and they do play an important role, but they do not help a search engine determine what the unique content is on a page. To do that, the search engine focuses on the part of Figure 2-13 that is labeled “Unique Page Content.”

Determining the unique content on a page is an important part of what the search engine does. The search engine uses its understanding of unique content to determine the types of search queries for which the web page might be relevant. Because site navigation is generally not unique to a single web page, it does not help the search engine with that task.

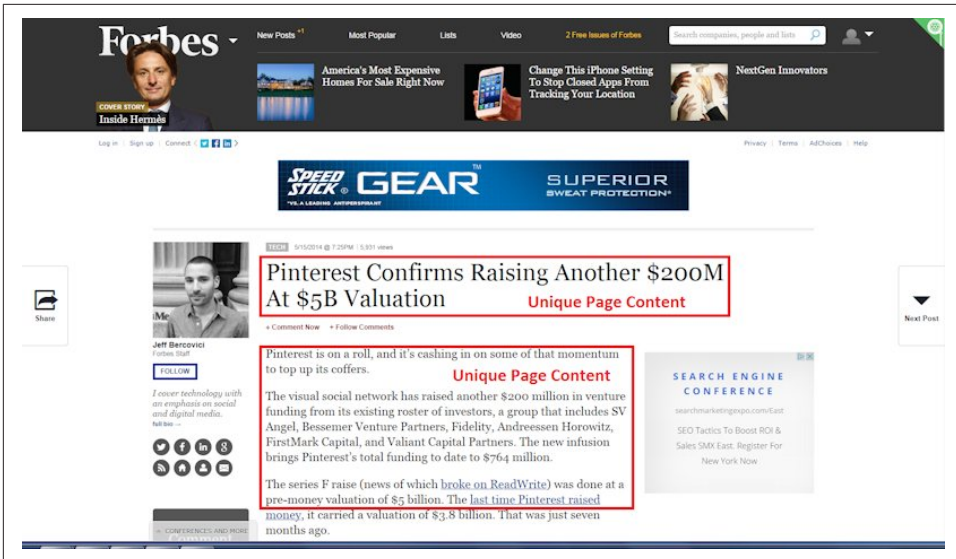


Figure 2-13. Breaking up a web page

This does not mean navigation links are not important—they most certainly are; however, they simply do not count when a search engine is trying to determine the unique content of a web page, as they are shared among many web pages.

One task the search engines face is judging the value of content. Although evaluating how the community responds to a piece of content using link analysis is part of the process, the search engines can also draw some conclusions based on what they see on the page.

For example, is the exact same content available on another website? Is the unique content the search engine can see two sentences long or 500 words long? Does the content repeat the same keywords excessively? These are a few examples of factors the search engine can evaluate when trying to determine the value of a piece of content.

Understanding What Content Search Engines Can “See” on a Web Page

Search engine crawlers and indexing programs are basically software programs. These programs are extraordinarily powerful. They crawl hundreds of trillions of web pages, analyze the content of all these pages, and analyze the way all these pages link to one another. Then they organize this into a series of databases that can respond to a user search query with a highly tuned set of results in a few tenths of a second.

This is an amazing accomplishment, but it has its limitations. Software is very mechanical, and it can understand only portions of most web pages. The [search engine crawler](#)

analyzes the raw HTML form of a web page. If you want to see what this looks like, you can do so by using your browser to view the source.

Figure 2-14 shows how to do that in Chrome, and Figure 2-15 shows how to do that in Firefox. Typically you can access it most easily by right-clicking with your mouse on a web page to access a hidden menu.



Figure 2-14. Viewing source in Chrome: right-click on the web page to access the menu

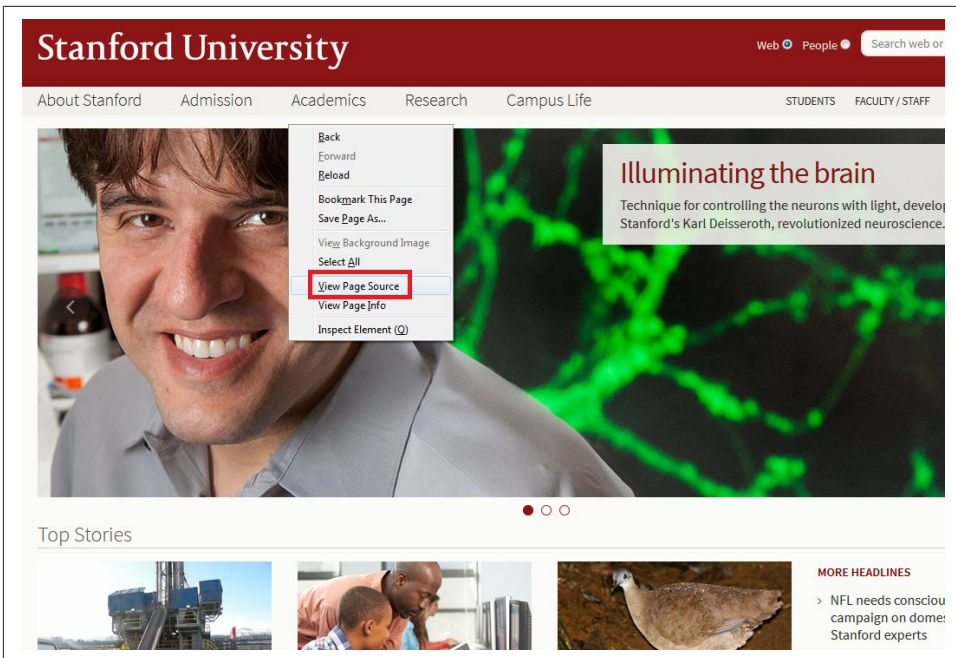


Figure 2-15. Viewing source in Firefox

There are also various in-browser web development tools (add-ons and extensions) that facilitate viewing source code in your browser of choice, as well as detecting web applications and JavaScript libraries. One of the most widely used code analysis tools is **Web Developer**, by Chris Pederick, available for Chrome, Firefox, and Opera. Once you view the source, you will be presented with the exact code for the web page that the web server sent to your browser. This is most of what the search engine crawler sees (the search engine also sees the HTTP headers for the page, which are status codes it receives from the web server where the page is hosted). In Some cases Google will execute JavaScript on the page as well. For more on how they do this, please refer to **Chapter 6**. When trying to analyze the user-visible content on a web page, search engines largely ignore code related to the navigation and display of the page, such as that shown in **Figure 2-16**, as it has nothing to do with the page's content.

```

<body class="home home-index layout-default">
  <div id="root" class="no-overflow">
    <nav class="globalnav-container default">
      <ul class="globalnav">
        <li class="active">
          <a href="http://moz.com" onclick="_gaq.push(['_trackEvent', 'multiproduct_nav', 'button', 'moz_com']>Moz.com</a>
        </li>
        <li class="">
          <a href="http://moz.com/pro" onclick="_gaq.push(['_trackEvent', 'multiproduct_nav', 'button', 'moz_pro']>Moz Pro</a>
        </li>
        <li class="">
          <a href="https://moz.com/local" onclick="_gaq.push(['_trackEvent', 'multiproduct_nav', 'button', 'moz_local']>Moz Local</a>
        </li>
      </ul>
    </nav><header class="background-blue-canvas masthead-container">
      <div class="container masthead">
        <div id="logo" class="span3 center">
          <a href="/">
            <svg width="120" height="35" alt="Moz">
              <path fill="#fff" transform="scale(1)"
                d="M102.3,22.8l120,4.2h83.6c-0.7,0-1.3,0.6-1.3,1.310,3.8c78.9,3.7,72.3,0.64.6,0c-10,0-18.3,6.4-19.8,14.7
                V4.2h-6.3c-1.5,0-2.9,0.7-3.8,1.7L22.4,19.5L10.1,5.9c-1-1-2.3-1.7-3.8-1.7H0v27.2h7.4c0.7,0,1.3-0.6,1.3-1.3h0c0,0,0-0.1,0-0.1
                V17.3l19.7,19.3l13.7-15.3l0,12.8c0,0.7,0.6,1.3,1.3,1.3h7.4v20.3c1.5,8.3,9.8,14.7,19.8,14.7c11.1,0,20.1-7.9,20.1-17.5
                c0-1.6-0.3-3.2-0.8-4.8h16L2.4,31.4h1.8h34.4c0,0,0,0,0,0,0,0,0,0,0h0c0.7,0,1.3-0.6,1.3-1.3h0v-7.3h102.3z M64.6,26.2
                c-5.5,0-10-3.9-10-8.7c0-4.8,4.5-8.7,10-8.7c5.5,0,10,3.9,10,8.7c74.6,22.3,70.1,26.2,64.6,26.2z" />
            </svg>
          </a>
        </div>
      </div>
    </header>
  </div>

```

Figure 2-16. Sample web page source code

The search engine crawler is most interested in the HTML text on the page.

Figure 2-17 is an example of HTML text for the Moz home page.

```

<h2 class="h3 top0 bottom1"><a href="http://moz.com/blog/tips-and-tactics-for-amplifying-your-content-whiteboard-friday"
class="slate">Tips and Tactics for Amplifying Your Content - Whiteboard Friday</a></h2>
<p class="bottom1 small">
  <time datetime="2014-06-20 00:12:00">June 20th, 2014</time>
  - Posted by <a href="http://moz.com/community/users/17229">Ben Lloyd</a> to
  <a href="http://moz.com/blog/category/content">Content</a> and
  <a href="http://moz.com/blog/category/whiteboard-friday">Whiteboard Friday</a>
</p>
</headers>
<p>Content marketing should never be approached with a "set it and forget it" mentality. It needs to be structured and
shared in the right ways, and in today's Whiteboard Friday, the folks from Add3 are here to show you what that means.</p>
</div>
<div class="pull-right">
  <a href="http://moz.com/blog/tips-and-tactics-for-amplifying-your-content-whiteboard-friday">Read Full Entry</a>
  |
  <span class="comment-count">
    <a href="http://moz.com/blog/tips-and-tactics-for-amplifying-your-content-whiteboard-friday#comments">
      <1 class="icon icon-speech"></1>
      32 comments </a>
    </span>
  </div>

```

Figure 2-17. Sample HTML text in the source code showing real content

Although Figure 2-17 still shows some HTML encoding, you can see the “regular” text clearly in the code. This is the unique content that the crawler is looking to find.

In addition, search engines read a few other elements. One of these is the page title. The page title is one of the most important factors in ranking a given web page. It is the text that shows in the browser’s title bar (above the browser menu and the address bar).

Figure 2-18 shows the code that the crawler sees, using Trip Advisor as an example.

The first highlighted area in Figure 2-18 is for the <title> tag. The <title> tag is also often (but not always) used as the title of your listing in search engine results (see Figure 2-19).


```
-----  
<meta http-equiv="imagealt" content="no"/>  
<title>Reviews of Hotels, Flights and Vacation Rentals - TripAdvisor</title>  
<meta http-equiv="pragma" content="no-cache"/>  
<meta http-equiv="cache-control" content="no-cache,must-revalidate"/>  
<meta http-equiv="expires" content="0"/>  
<meta property="og:image" content="http://cl.tacdn.com/img2/postimg.jpg" height="150px"  
width="150px"/>  
<meta name="keywords" content="vacation, vacations, vacation packages, vacation package,  
travel package, travel packages, travel, planning, hotel, hotels, motel, bed and  
breakfast, inn, guidebook, review, reviews, popular, plan, airfare, cheap, discount, map,  
maps, golf, ski, articles, attractions, advice, restaurants"/>  
<meta name="description" content="TripAdvisor - Unbiased hotel reviews, photos and travel  
advice for hotels and vacations - Compare prices with just one click."/>  
<link rel="alternate" hreflang="en" href="http://www.tripadvisor.com/">  
-----
```

Figure 2-18. Meta tags in HTML source

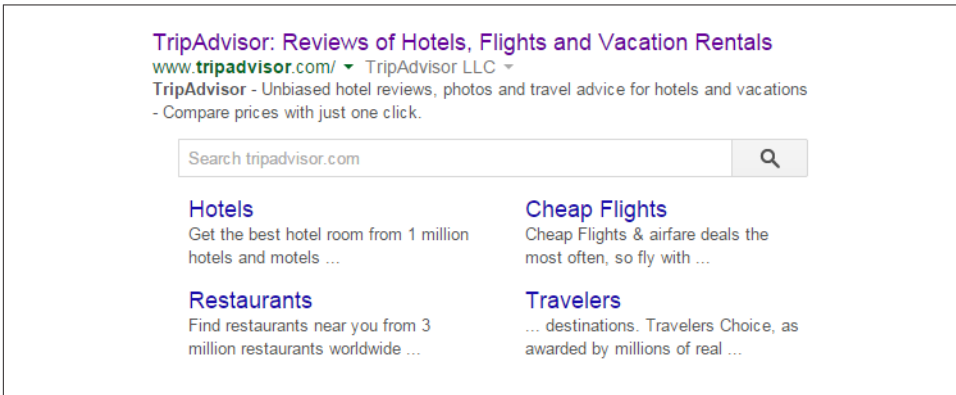


Figure 2-19. Search result showing the title tag

In addition to page titles, search engines previously used the [meta keywords tag](#). This is a list of keywords that you wish to have associated with the page. *Spammers* (people who attempt to manipulate search engine results in violation of the search engine guidelines) ruined the SEO value of this tag many years ago, so its value is now negligible, as search engines don't use it anymore. Spending time on meta keywords is not recommended because of the lack of SEO benefit.

The second highlighted area in [Figure 2-18](#) shows an example of a meta keywords tag.

Search engines also read the [meta description tag](#) (the third highlighted area in the HTML source in [Figure 2-18](#)). However, the content of a meta description tag is not directly used by search engines in their ranking algorithms.²

Nonetheless, the meta description tag plays a key role, as search engines often use it as a part or all of the [description for your page in search results](#). Therefore, a

2 For more information, see Matt McGee, "21 Essential SEO Tips & Techniques," Search Engine Land, June 20, 2011, <http://searchengineland.com/21-essential-seo-tips-techniques-11580>.

well-written meta description can have a significant influence on how many clicks you get on your search listing, and the click-through rate on your search listing can impact your ranking. As a result, time spent on meta descriptions is quite valuable.

Figure 2-20 uses a search on *trip advisor* to show an example of the meta description tag being used as a description in the search results.

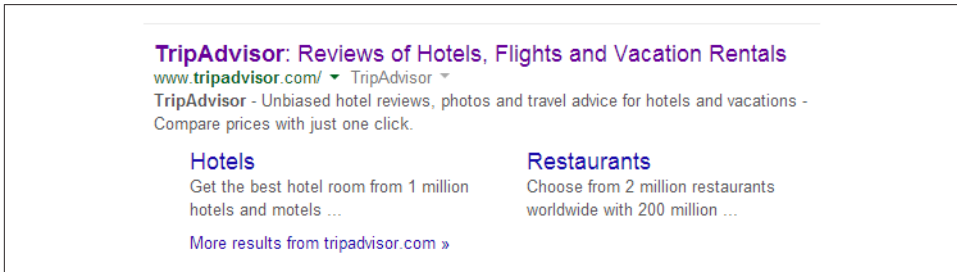


Figure 2-20. Meta description used in search results

NOTE

The user's keywords are typically shown in boldface when they appear in the search results (sometimes close synonyms are shown in boldface as well). As an example, in Figure 2-20, *TripAdvisor* is in boldface at the beginning of the description. This is called keywords in context (KWIC).

A fourth element that search engines read is the [alt attribute for images](#). The alt attribute was originally intended to allow something to be rendered for audiences who cannot view the images, primarily:

- Vision-impaired people who do not have the option of viewing the images.
- People who turn off images for faster surfing. This is generally an issue only for those who do not have a broadband connection.

Support for the vision-impaired remains a major reason for using the alt attribute. You can read more about this by visiting [the W3C's Web Accessibility Initiative page](#).

Search engines also read the text contained in the alt attribute of an image tag (). An image tag is an element that is used to tell a web page to display an image.

Another element that search engines read is the [<noscript> tag](#). Historically, the ability of search engines to read JavaScript was quite limited, but this has been changing over time and [Google says they execute more JavaScript today](#).³ However, [a small percentage of users do not allow JavaScript to run when they load a web page](#) (our experi-

³ Google Webmaster Central Blog, "Understanding Web Pages Better," May 23, 2014, http://bit.ly/web_pages_understanding.

ence is that it is about 2%). For those users, nothing would be shown to them where the JavaScript is on the web page, unless the page contains a `<noscript>` tag.

Here is a very simple JavaScript example that demonstrates this:

```
<script type="text/javascript">
document.write("It is a Small World After All!")
</script>
<noscript>Your browser does not support JavaScript!</noscript>
```

The `<noscript>` portion of this is `Your browser does not support JavaScript!`. In this example, you could also choose to make the `<noscript>` tag contain the text `"It is a Small World After All!"`. The `<noscript>` tag should be used only to represent the content of the JavaScript.

What search engines cannot see

It is also worthwhile to review the types of content that search engines cannot “see” in the human sense.

For instance, although search engines are able to detect that you are displaying an image, they have little idea what the image is a picture of, except for whatever information you provide in the `alt` attribute, as discussed earlier. They can recognize only some very basic types of information within images, such as the presence of a face, or whether images have pornographic content by how much flesh tone they contain. A search engine cannot easily tell whether an image is a picture of Bart Simpson, a boat, a house, or a tornado. In addition, search engines typically don’t recognize any text rendered in the image.

The reality is that the search engines have the technology to handle these types of tasks to some degree. For example, you can take a picture of the Taj Mahal and drag it into the search box in Google image search, and the search engine will recognize it. However, because of the processing power required for image recognition, search engines do not currently try to recognize all of the images they encounter across the Web.

Search engines are also experimenting with technology to use optical character recognition (OCR) to extract text from images, but it is not yet in general use within search. The main problem with applying OCR and image processing technology is that it’s very computationally intensive, and not practical to apply at the scale of the Web.

In addition, conventional SEO wisdom has always held that the search engines cannot read Flash files, but this is a little overstated. Search engines have been extracting some information from Flash for years, as indicated by [this Google announcement in 2008](#). However, the bottom line is that it’s not easy for search engines to determine what is in Flash. One of the big issues is that even when search engines look inside

Flash, they are still looking for textual content, but Flash is a pictorial medium and there is little incentive (other than the search engines) for a designer to implement text inside Flash. All the semantic clues that would be present in HTML text (such as heading tags, boldface text, etc.) are missing too, even when HTML is used in conjunction with Flash.

A third type of content that search engines cannot see is the pictorial aspects of anything contained in Flash, so this aspect of Flash behaves in the same way images do. For example, when text is converted into a vector-based outline (i.e., rendered graphically), the textual information that search engines can read is lost. Chapter 6 discusses methods for optimizing Flash.

Audio and video files are also not easy for search engines to read. As with images, the data is not easy to parse. There are a few exceptions where the search engines can extract some limited data, such as ID3 tags within MP3 files, or enhanced podcasts in AAC format with textual “show notes,” images, and chapter markers embedded. Ultimately, though, search engines cannot distinguish a video of a soccer game from a video of a forest fire.

Search engines also cannot read any content contained within a program. The search engine really needs to find text that is readable by human eyes looking at the source code of a web page, as outlined earlier. It does not help if you can see it when the browser loads a web page—it has to be visible and readable in the source code for that page.

One example of a technology that can present significant human-readable content that search engines cannot see is AJAX. AJAX is a JavaScript-based method for dynamically rendering content on a web page after retrieving the data from a database, without having to refresh the entire page. This is often used in tools where a visitor to a site can provide some input and the AJAX tool then retrieves and renders the correct content.

The problem arises because the content is retrieved by a script running on the client computer (the user’s machine) only after receiving some input from the user. This can result in many potentially different outputs. In addition, until that input is received, the content is not present in the HTML of the page, so the search engines cannot easily see it.

Similar problems arise with other forms of JavaScript that don’t render the content in the HTML until a user action is taken. New forms of JavaScript, such as AngularJS,

make this even more challenging for search engines. For more information on SEO for single-page web applications, please see “[Single-Page Applications](#)” on page 164.⁴

As of HTML 5, a construct known as the **embed tag** (<embed>) was created to allow the incorporation of *plug-ins* into an HTML page. Plug-ins are programs located on the user’s computer, not on the web server of your website. The embed tag is often used to incorporate movies or audio files into a web page; it tells the plug-in where it should look to find the data file to use. Content included through plug-ins may or may not be invisible to search engines.

Frames and iframes are methods for incorporating the content from another web page into your web page. Iframes are more commonly used than frames to incorporate content from another website. You can execute an iframe quite simply with code that looks like this:

```
<iframe src ="http://accounting.careerbuilder.com" width="100%" height="300">
  <p>Your browser does not support iframes.</p>
</iframe>
```

Frames are typically used to subdivide the content of a publisher’s website, but they can be used to bring in content from other websites, as in <http://accounting.careerbuilder.com> on the *Chicago Tribune* website, shown in [Figure 2-21](#).

[Figure 2-21](#) is an example of something that works well to pull in content (provided you have permission to do so) from another site and place it on your own. However, the **search engines recognize an iframe or a frame used to pull in another site’s content for what it is, and therefore may ignore that content**. In other words, they don’t consider content pulled in from another site as part of the unique content of your web page.

⁴ For even further discussion of this topic, see “[How do search engines deal with AngularJS applications?](#)” on [StackOverflow](#).

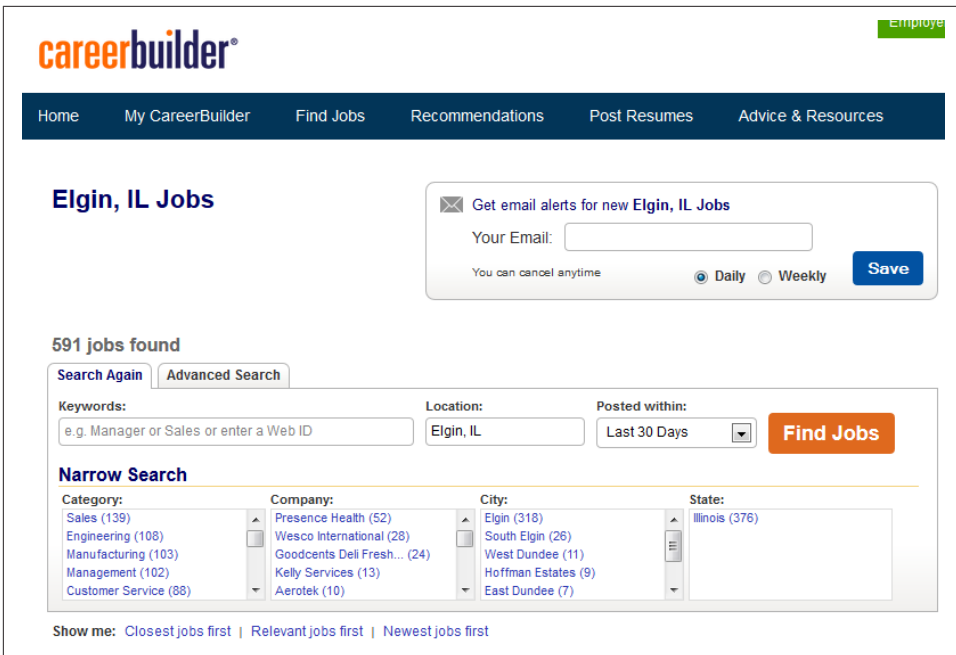


Figure 2-21. Framed page rendered in a browser

Determining Searcher Intent and Delivering Relevant, Fresh Content

Modern commercial search engines rely on the science of *information retrieval* (IR). This science has existed since the middle of the 20th century, when retrieval systems powered computers in libraries, research facilities, and government labs. Early in the development of search systems, IR scientists realized that two critical components comprised the majority of search functionality: relevance and importance (which we defined earlier in this chapter). To measure these factors, search engines perform document analysis (including semantic analysis of concepts across documents) and link (or citation) analysis.

Document Analysis and Semantic Connectivity

In document analysis, search engines look at whether they find the search terms in important areas of the document—the title, the metadata, the heading tags, and the body of the text. They also attempt to automatically measure the quality of the document based on document analysis, as well as many other factors.

Reliance on document analysis alone is not enough for today's search engines, so they also look at semantic connectivity. Semantic connectivity refers to words or phrases that are commonly associated with one another. For example, if you see the word *aloha*,

you associate it with Hawaii, not Florida. Search engines actively build their own thesaurus and dictionary to help them determine how certain terms and topics are related. By simply scanning their massive databases of content on the Web, they can use fuzzy set theory and certain equations to connect terms and start to understand web pages and sites more like a human does.

The professional SEO practitioner does not necessarily need to use semantic connectivity measurement tools to optimize websites, but for those advanced practitioners who seek every advantage, semantic connectivity measurements can help in each of the following sectors:

- Measuring which keyword phrases to target
- Measuring which keyword phrases to include on a page about a certain topic
- Measuring the relationships of text on other high-ranking sites and pages
- Finding pages that provide “relevant” themed links

Although the source for this material is highly technical, SEO specialists need only know the principles to obtain valuable information. It is important to keep in mind that although the world of IR has hundreds of technical and often difficult-to-comprehend terms, these can be broken down and understood even by an SEO novice.

Common types of searches in the IR field include:

Proximity searches

A proximity search uses the order of the search phrase to find related documents. For example, when you search for “*sweet German mustard*” you are specifying only a precise proximity match. If the quotes are removed, the proximity of the search terms still matters to the search engine, but it will now show documents that don’t exactly match the order of the search phrase, such as *Sweet Mustard—German*.

Fuzzy logic

Fuzzy logic technically refers to logic that is not categorically true or false. A common example is whether a day is sunny (i.e., is 50% cloud cover a sunny day?). In search, fuzzy logic is often used for misspellings.

Boolean searches

These are searches that use Boolean terms such as AND, OR, and NOT. This type of logic is used to expand or restrict which documents are returned in a search.

Term weighting

Term weighting refers to the importance of a particular search term to the query. The idea is to weight particular terms more heavily than others to produce supe-

rior search results. For example, the appearance of the word *the* in a query will receive very little weight in selecting the results because it appears in nearly all English language documents. There is nothing unique about it, and it does not help in document selection.

IR models (search engines) use fuzzy set theory (an offshoot of fuzzy logic created by Dr. Lotfi Zadeh in 1969) to discover the semantic connectivity between two words. Rather than using a thesaurus or dictionary to try to reason whether two words are related to each other, an IR system can use its massive database of content to puzzle out the relationships.

Although this process may sound complicated, the foundations are simple. Search engines need to rely on machine logic (true/false, yes/no, etc.). Machine logic has some advantages over humans, but it doesn't have a way of thinking like humans, and concepts that are intuitive to humans can be quite hard for a computer to understand. For example, oranges and bananas are both fruits, but oranges and bananas are not both round. To a human this is intuitive.

For a machine to understand this concept and pick up on others like it, semantic connectivity can be the key. The massive human knowledge on the Web can be captured in the system's index and analyzed to artificially create the relationships humans have made. Thus, a machine knows an orange is round and a banana is not by scanning thousands of occurrences of the words *banana* and *orange* in its index and noting that *round* and *banana* do not have great concurrence, while *orange* and *round* do.

This is how the use of fuzzy logic comes into play, and the use of fuzzy set theory helps the computer to understand how terms are related simply by measuring how often and in what context they are used together.

For example, a search engine would recognize that *trips* to the *zoo* often include *viewing wildlife* and *animals*, possibly as part of a *tour*.

To see this in action, conduct a search on Google for *zoo trips*. Note that the boldface words that are returned match the terms that are italicized in the preceding paragraph. Google is setting "related" terms in boldface and recognizing which terms frequently occur concurrently (together, on the same page, or in close proximity) in their indexes.

Search companies have been investing in these types of technologies for many years. In September 2013, **Google quietly let the world know that it had rewritten its search engine and given it the name "Hummingbird"**. This rewrite was in large part done to enable a whole new set of capabilities for recognizing the relationships between things.

For example, if you use Google's voice search (click on the microphone icon at the right of the search box on Google.com) and ask it "Who is Tom Brady?" it will answer

that question for you with a search result, but then use audio to tell you that he is an “American football quarterback for the New England Patriots of the National Football League.”

This shows that Google understands many aspects of Tom Brady. For example:

- He has an occupation: quarterback, playing American football (as distinct from the way the term *football* is used outside of the United States and Canada).
- He plays on a team: the New England Patriots.
- The New England Patriots belong to a league: the NFL.

This is far more sophisticated than search was in 2012. You can take this much further. For example, if you now use the voice search feature to ask “Who is his wife?” it will answer that question too (see [Figure 2-22](#)).

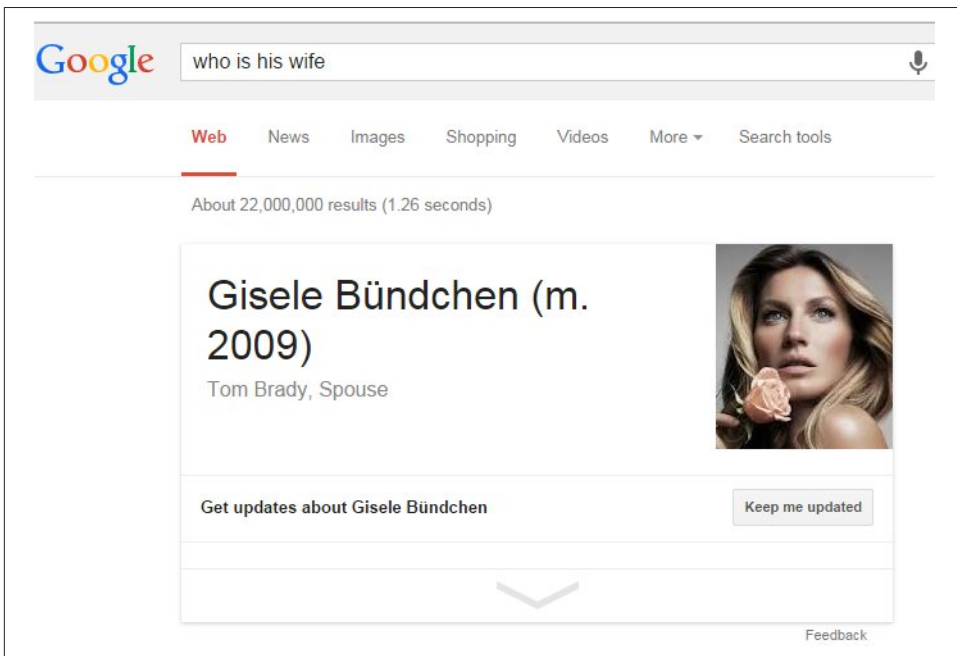


Figure 2-22. *Gisele Bündchen is Tom Brady's wife*

Notice too that in the second query we did not repeat Tom Brady’s name, and that Google remembered the context of the conversation, in that “his” refers to Tom Brady. You can continue with a question such as “Does he have children?” and Google will answer that as well.

For SEO purposes, this usage opens our eyes to realizing how search engines recognize the connections between words, phrases, and ideas on the Web. As semantic connec-

tivity becomes a bigger part of search engine algorithms, you can expect greater emphasis on the theme of pages, sites, and links. It will be important going into the future to realize search engines' ability to pick up on ideas and themes and recognize content, links, and pages that don't fit well into the scheme of a website.

Content Quality and User Engagement

Search engines also attempt to measure the quality and uniqueness of a website's content. One method they may use for doing this is evaluating the document itself. For example, if a web page has lots of spelling and grammatical errors, that can be taken as a sign that little editorial effort was put into that page.⁵

They can also analyze the reading level of the document. One popular formula for doing this is the Flesch-Kincaid Grade Level Readability Formula, which considers factors like the average word length and the words per sentence to determine the level of education needed to be able to understand the sentence. Imagine a scenario where the product being sold on a page is children's toys and the search engine calculates a reading level of a college senior. This could be another indicator of a poor editorial effort.

The other method that search engines can use to evaluate the quality of a web page is measuring actual user interaction. For example, if a large number of users who visit the web page after clicking on a search result immediately return to the search engine and click on the next result, that would be a strong indicator of poor quality.

Engagement with a website began to publicly emerge as a ranking factor with the release of the Panda update by Google on February 23, 2011.⁶ Google has access to a large number of data sources that it can use to measure how visitors interact with your website. Just because Google has access to this data, however, does not mean that it's definitely using the data as a ranking factor. That noted, some of those sources include:

Interaction with web search results

For example, if a user clicks through on a SERP listing and comes to your site, clicks the back button, and then clicks on another result in the same set of search results, that could be seen as a negative ranking signal. Or if the results below you in the SERPs are getting clicked on more than you are, that could be seen as a negative ranking signal for you and a positive ranking signal for them. Whether search engines use this signal or not, or how much weight they might put on it, is not known.

5 For more information, see Matt McGee, "Google: Low PageRank & Bad Spelling May Go Hand-In-Hand; Panda, Too?", October 5, 2011, http://bit.ly/pagerank_spelling.

6 See Danny Sullivan, "Google Forecloses On Content Farms With "Panda" Algorithm Update," February 24, 2011, http://bit.ly/panda_update.

Google Analytics

It is hard to get a firm handle on just what percentage of websites run Google Analytics. A 2008 survey of websites by Immeria.net showed their share at 59%,⁷ and the Metric Mail Blog checked the top 1 million sites in Alexa and found that about 50% of those had Google Analytics.⁸ Suffice it to say that Google is able to collect detailed data about what is taking place on a large percentage of the world's websites.

This provides Google with a rich array of data on that site, including:

Bounce rate

The percentage of visitors who visit only one page on your website.

Time on site

The time spent by the user on the site. Note that Google Analytics receives information only when each page is loaded, so if you view only one page it does not know how much time you spent on that page. More precisely, then, this metric tells you the average time between the loading of the first page and the loading of the last page, but does not take into account how long visitors spent on the last page loaded.

Page views per visitor

The average number of pages viewed per visitor on your site.

Google Toolbar

It is not known how many users out there use the Google Toolbar, but we believe that it numbers in the millions. For these users, Google can track their entire web surfing behavior. Unlike Google Analytics, the Google Toolbar can measure the time from when a user first arrives on a site to the time when she loads a page from a different website. It can also get measurements of bounce rate and page views per visitor.

Google +1 button

This enables users to vote for a page on the page itself. There is currently no evidence that Google uses this as a ranking factor, but in theory, it could. You can see a lot more about this in **Chapter 8**.

Chrome Personal Blocklist Extension

Google offers a Chrome add-on called **the Personal Blocklist Extension**. This enables users of the Chrome browser to indicate a search result they don't like. This

7 Stéphane Hamel, "Web Analytics vendors market shares," immeria - S.Hamel's blog, January 4, 2008, <http://blog.immeria.net/2008/01/web-analytics-vendors-market-shares.html>.

8 Metric Mail, "Google Analytics Market Share," The Metric Mail Blog, August 4, 2010, http://bit.ly/analytics_mkt_share.

was first used by Google as a part of its Panda algorithm, which attempts to measure the quality of a piece of content. You can read more about this algorithm in [Chapter 9](#).

Goo.gl

Google has its own URL shortener. This tool allows Google to see what content is being shared, and which content is being clicked on, even in closed environments where Google web crawlers are not allowed to go.

What matters most is how your site compares to that of your competition. If your site has better engagement metrics, this is likely to be seen as an indication of quality and can potentially boost your rankings with respect to your competitors. Little has been made public about the way search engines use these types of signals, so the preceding comments are our speculation on what Google may be doing in this area. Social and user engagement ranking factors are discussed in more detail in [Chapter 8](#).

Link Analysis

In link analysis, search engines measure who is linking to a site or page and what they are saying about that site/page. They also have a good grasp on who is affiliated with whom (through historical link data, the site's registration records, and other sources), who is worthy of being trusted based on the authority of sites linking to them, and contextual data about the site on which the page is hosted (who links to that site, what they say about the site, etc.).

Link analysis goes much deeper than counting the number of links a web page or website has, as all links are not created equal (one link can be worth 10 million times more than another one). Links from a highly authoritative page on a highly authoritative site will count more than other links of lesser authority. A search engine can determine a website or page to be authoritative by combining an analysis of the linking patterns and semantic analysis.

For example, perhaps you are interested in sites about dog grooming. Search engines can use semantic analysis to identify the collection of web pages that focus on the topic of dog grooming. The search engines can then determine which of these pages about dog grooming have the most links from the set of websites relevant to the topic of dog grooming. These pages are most likely more authoritative on the topic than the others.

The actual analysis is a bit more complicated than that. For example, imagine that there are five pages about dog grooming with a lot of links from pages across the Web on the topic, as follows:

- Page A has 213 topically related links.
- Page B has 192 topically related links.

- Page C has 203 topically related links.
- Page D has 113 topically related links.
- Page E has 122 topically related links.

Further, it may be that Pages A, B, D, and E all link to one another, but none of them links to Page C. In fact, Page C appears to have the great majority of its relevant links from other pages that are topically relevant but have few links to them. In this scenario, Page C may not be considered authoritative because it is not linked to by the right sites.

The concept of grouping sites based on who links to them, and whom they link to, is referred to as grouping sites by *link neighborhood*. The neighborhood you are in says something about the subject matter of your site, and the number and quality of the links you get from sites in that neighborhood say something about how important your site is to that topic.

The degree to which search engines rely on evaluating link neighborhoods is not clear, and links from irrelevant pages can still help the rankings of the target pages. Nonetheless, the basic idea remains that a link from a relevant page or site should be more valuable than a link from a nonrelevant page or site.

Another factor in determining the value of a link is the way the link is implemented and where it is placed. For example, the text used in the link itself (i.e., the actual text that will go to your web page when the user clicks on it) is also a strong signal to the search engines.

This is referred to as *anchor text*, and if that text is keyword-rich (with keywords relevant to your targeted search terms), it can potentially do more for your rankings in the search engines than if the link is not keyword-rich. For example, anchor text of “Dog Grooming Salon” may bring more value to a dog grooming salon’s website than anchor text of “Click here.” However, take care. If you get 10,000 links using the anchor text “Dog Grooming Salon” and you have few other links to your site, this definitely does not look natural and could lead to a penalty.

The semantic analysis of a link’s value goes deeper than just the anchor text. For example, if you have that “Dog Grooming Salon” anchor text on a web page that is not really about dog grooming at all, the value of the link is lower than if the page is about dog grooming. Search engines also look at the content on the page immediately surrounding the link, as well as the overall context and authority of the website that is providing the link.

All of these factors are components of link analysis, which we will discuss in greater detail in [Chapter 7](#).

Evaluating Social Media Signals

Sites such as [Facebook](#), [Twitter](#), and [Google+](#) have created whole new ways for users to share content or indicate that they value it. This has led many to speculate that search engines could be using these signals as a ranking factor. Fueling that speculation, in August 2013, [Moz released the data from its latest correlation study](#), and it showed a very strong correlation between +1s and ranking in Google.

Figure 2-23 shows the top 10 results in that data, and Google +1s had the second strongest correlation with rankings.

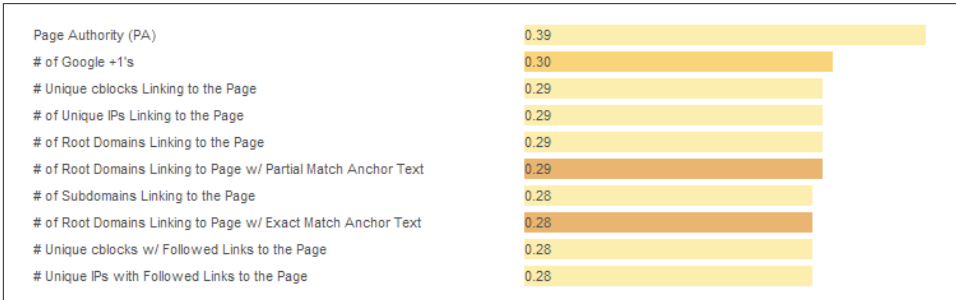


Figure 2-23. Top 10 results in Moz's 2013 correlation study

However, the fact that there is a correlation in no way means that +1s are used by Google as a ranking signal, or that they cause higher rankings. It can simply mean that good content that receives lots of links (which is known to be a signal that causes higher rankings) also happens to get many +1s.

In fact, Stone Temple Consulting did a different study targeted at measuring whether or not Google+ activity was used as a ranking factor by Google.⁹ This study showed that there was no material evidence that Google+ shares or +1s had any material impact on rankings. The potential for social signals as a ranking factor is discussed in depth in [Chapter 8](#).

Problem Words, Disambiguation, and Diversity

On the opposite side of the coin are words that present an ongoing challenge for the search engines. One of the greatest challenges comes in the form of disambiguation. For example, when someone types in *boxers*, does he mean the prize fighter, the breed of dog, or the type of underwear? Another example is *jaguar*, which is at once a jungle

⁹ Eric Enge, "Direct Measurement of Google Plus Impact on Search Rankings," Stone Temple Consulting, September 17, 2013, <https://www.stonetemple.com/measuring-google-plus-impact-on-search-rankings/>.

cat, a car, a football team, an operating system, and a guitar. Which does the user mean?

Search engines deal with these types of ambiguous queries all the time. The two examples offered here have inherent problems built into them, but the problem is much bigger than that. For example, if someone types in a query such as *cars*, does he:

- Want to read reviews?
- Want to go to a car show?
- Want to buy one?
- Want to read about new car technologies?

The query *cars* is so general that there is no real way to get to the bottom of the searcher's intent based on this query alone. One way that search engines deal with this is by looking at prior queries by the same searcher to find additional clues to his intent. We discuss this a bit more in “Adaptive Search” on page 48.

Another solution they use is to offer diverse results. As an example, Figure 2-24 shows a generic search, this time for *GDP*.

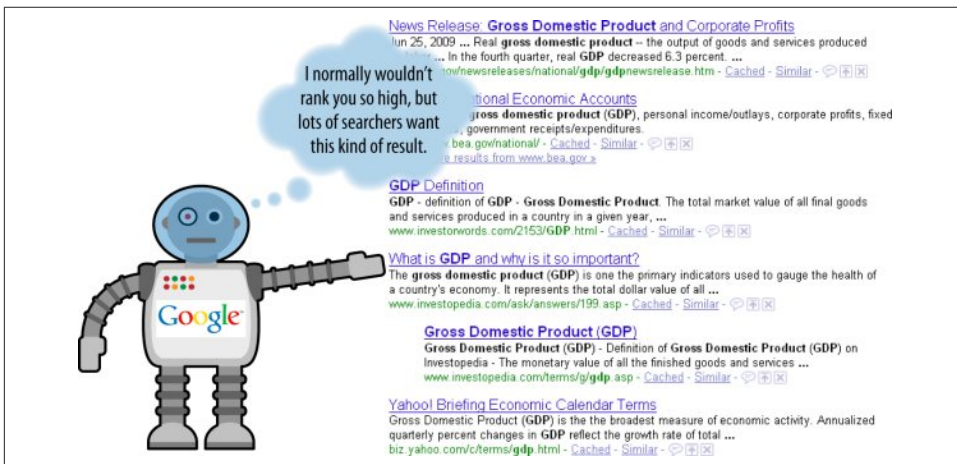


Figure 2-24. Diverse results example

This brings up an important ranking concept. It is possible that a strict analysis of the relevance and link-driven importance scores in Figure 2-24 would not have resulted by itself in the Investopedia.com result being on the first page, *but* the need for diversity elevated the page's ranking. This concept of altering the results in this manner is sometimes referred to as *query deserves diversity* (QDD).

A strict relevance- and importance-based ranking system might have shown a variety of additional government pages discussing the GDP of the United States. However, a

large percentage of users will likely be satisfied by the government pages already shown, but for those users who are not, showing more of the same types of pages is not likely to raise their level of satisfaction with the results.

Introducing a bit of variety allows Google to also provide a satisfactory answer to those who are looking for something different from the government pages. Google's testing has shown that this diversity-based approach has resulted in a higher level of satisfaction among its users.

For example, the testing data for the nondiversified results may have shown lower click-through rates in the SERPs, greater numbers of query refinements, and even a high percentage of related searches performed subsequently.

The idea to deliberately introduce diversity into the result algorithm makes sense and can enhance searcher satisfaction for queries such as:

- Company names (where searchers might want to get positive and negative press, as well as official company domains)
- Product searches (where ecommerce-style results might ordinarily fill up the SERPs, but Google tries to provide some reviews and noncommercial, relevant content)
- News and political searches (where it might be prudent to display “all sides” of an issue, rather than just the left- or right-wing blogs that did the best job of obtaining links)

Search engines also personalize results for users based on their search history or past patterns of behavior. For example, if a searcher has a history of searching on card games, and then does a search for *dominion*, the search engine may choose to push some of the results related to the *Dominion* card game higher in the results, instead of emphasizing the power company.

Where freshness matters

Much of the time, it makes sense for the search engines to deliver results from older sources that have stood the test of time. However, other times the response should be from newer sources of information.

For example, when there is breaking news, such as an earthquake, the search engines begin to receive queries within seconds, and the first articles begin to appear on the Web within 15 minutes.

In these types of scenarios, there is a need to discover and index new information in near real time. Google refers to this concept as *query deserves freshness* (QDF). According to the *New York Times*, QDF takes several factors into account,¹⁰ such as:

- Search volume
- News coverage
- Blog coverage

QDF applies to up-to-the-minute news coverage, but also to other scenarios such as hot, new discount deals or new product releases that get strong search volume and media coverage. There has also been speculation that Google will apply QDF more to sites that have higher PageRank.¹¹

Why These Algorithms Sometimes Fail

As we've outlined in this chapter, search engines do some amazing stuff. Nonetheless, there are times when the process does not work as well as you would like to think. Part of this is because users often type in search phrases that provide very little information about their intent (e.g., if they search on *car*, do they want to buy one, read reviews, learn how to drive one, learn how to design one, or something else?). Another reason is that some words have multiple meanings, such as the *jaguar* example we used previously in this section.

For more information on why search algorithms sometimes fail, you can read Hamlet Batista's Moz article, "[7 Reasons Why Search Engines Don't Return Relevant Results 100% of the Time](#)".

The Knowledge Graph

Traditional search results are derived by search engines crawling and analyzing web pages and then presenting that information in the search results. However, **Google's mission "is to organize the world's information and make it universally accessible and useful"**. Google is actively pursuing initiatives to build databases of information that go far beyond traditional web-based search.

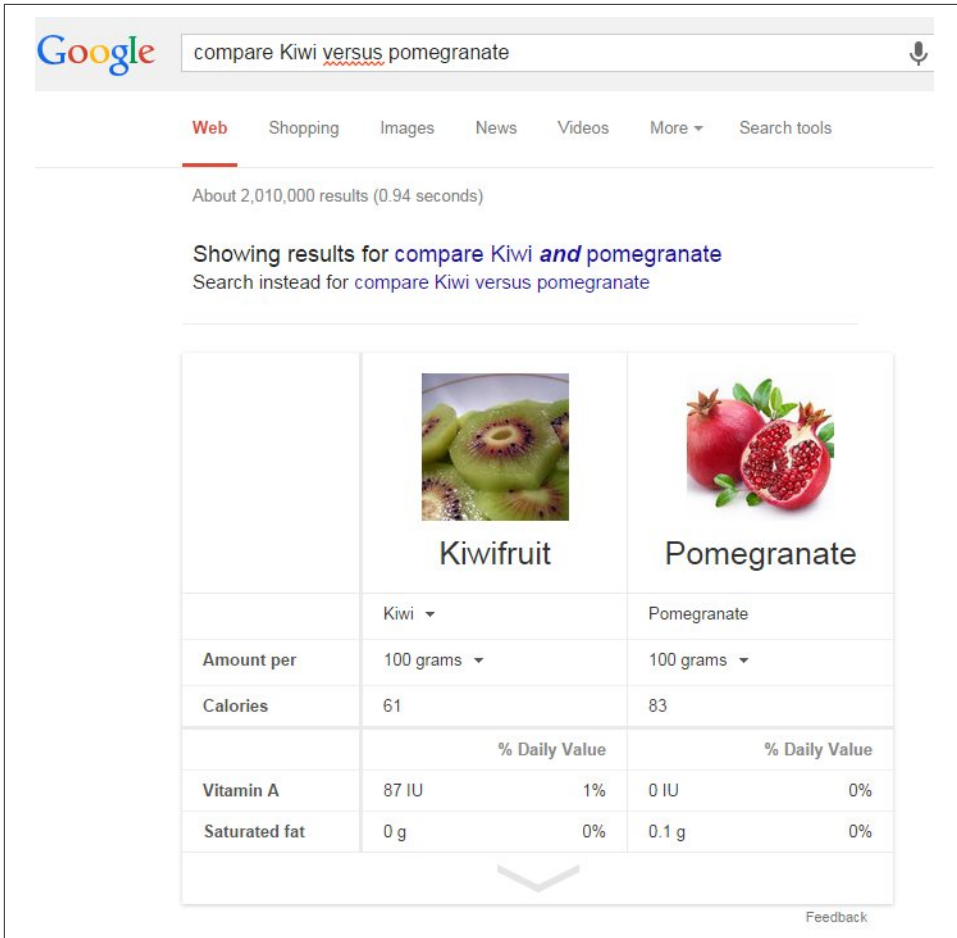
Note that earlier in this chapter we wrote about vertical search. Vertical search relates to breaking search into different categories, such as a search for images, videos, or local business information. The Knowledge Graph is more about providing richer answers

¹⁰ Saul Hansell, "Google Keeps Tweaking Its Search Engine," *New York Times*, June 3, 2007, http://www.nytimes.com/2007/06/03/business/yourmoney/03google.html?pagewanted=3&_r=0.

¹¹ Sean Jackson, "The Surprising Effect of Freshness and Authority on Search Results By," *Copyblogger*, February 21, 2013, <http://www.copyblogger.com/query-deserves-freshness/>.

directly in the search results, often answering the user's question directly without her having to click through to a website.

In May 2012, **Google announced the Knowledge Graph**. Initially, this was a set of structured databases of information that allows Google to access information without deriving it from the Web. You can see an example of the type of data that Google might extract from its Knowledge Graph database in **Figure 2-25**.



The screenshot shows a Google search interface with the query "compare Kiwi versus pomegranate". The search results are filtered to show a comparison between Kiwifruit and Pomegranate. The comparison table is as follows:

	Kiwi	Pomegranate
Amount per	100 grams	100 grams
Calories	61	83
	% Daily Value	
Vitamin A	87 IU (1%)	0 IU (0%)
Saturated fat	0 g (0%)	0.1 g (0%)

Figure 2-25. *Kiwi versus pomegranate search result*

Google initially built the Knowledge Graph using data from Freebase, Wikipedia, and the CIA Fact Book. This allowed Google to answer many questions, but really only satisfied a very small number of search queries. For that reason, Google is constantly working on expanding the information in the Knowledge Graph.

In addition, Google is investing in ways to more reliably extract information from other sources, including websites, to present as direct answers in search. Google refers to these as “featured snippets.” [Figure 2-26](#) shows the search result for buying a car.

In this result, Google provides a set of step-by-step instructions extracted from the CNN Money website. Note that two steps are omitted, so to get the complete procedure or additional details on each step, the user must click through to the CNN Money website.

In some cases, Google does provide the complete instructions in the search results, but most of the time it does not. A study performed by Stone Temple Consulting examined 276 examples of step-by-step instructions, and found that 217 of these (79%) did not provide the complete instructions.

A related concept is *semantic search*, which overlaps the Knowledge Graph to some degree, but also takes into account many other factors to personalize results for the searcher. You can see a depiction of some of these factors in [Figure 2-27](#).

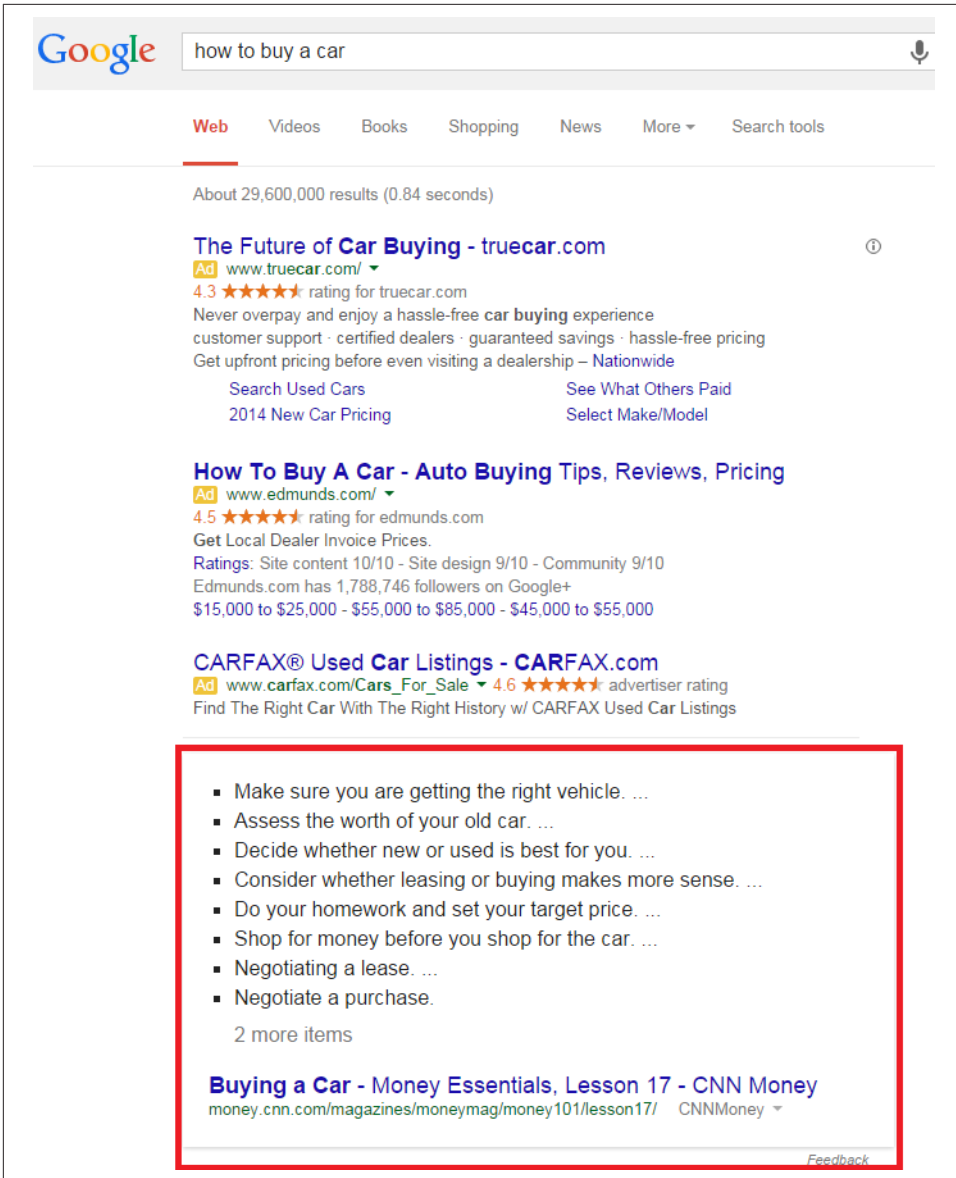


Figure 2-26. Step-by-step instructions for buying a car

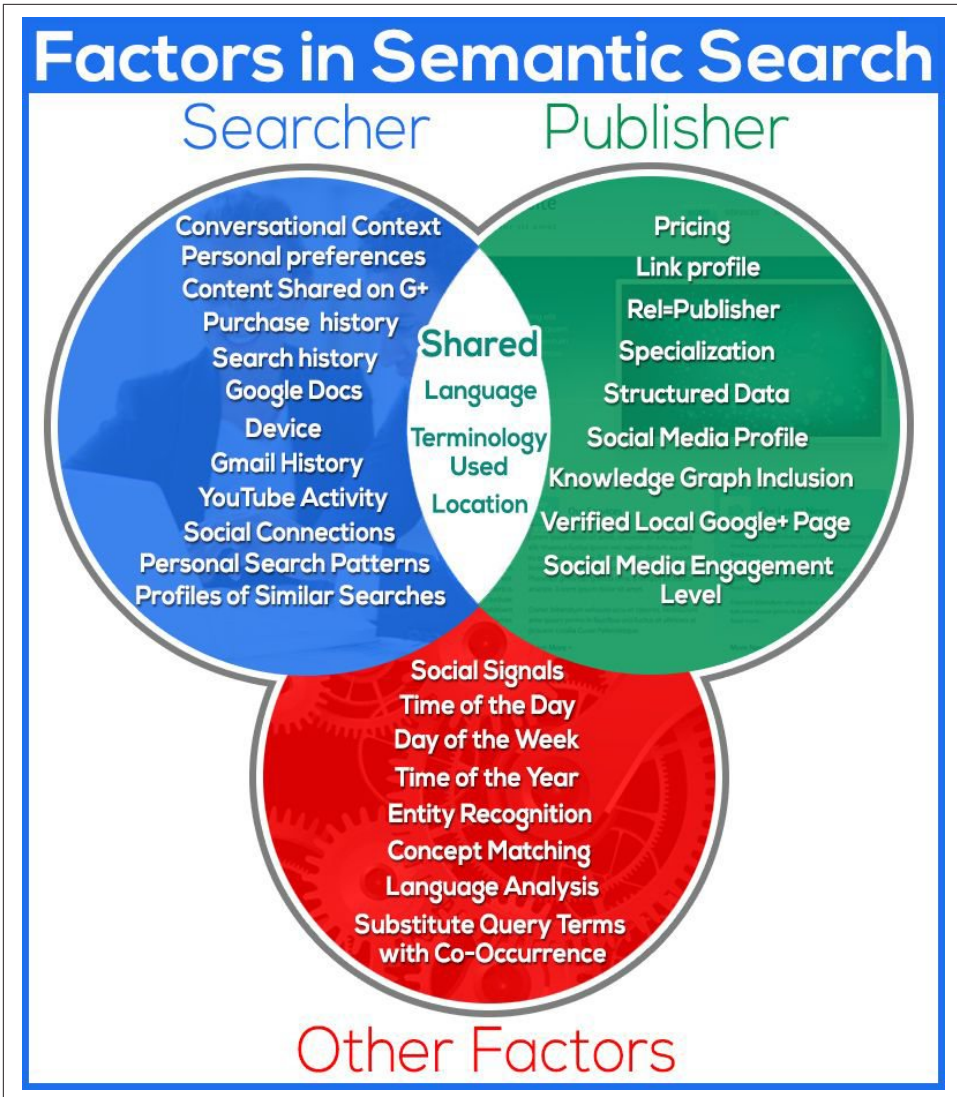


Figure 2-27. Factors involved in semantic search

The Knowledge Graph and semantic search are discussed in more detail in [Chapter 6](#).

Analyzing Ranking Factors

Moz periodically conducts surveys of leading SEOs to determine what they think are the most important ranking factors.¹² Here is a high-level summary of the top nine results, in priority order (as suggested by the referenced study):

- [Domain-level link authority features](#)
- [Page-level link metrics](#)
- [Page-level keywords and content](#)
- [Page-level keyword-agnostic features](#)
- [Domain-level brand metrics](#)
- [Usage and traffic/query data](#)
- [Page-level social metrics](#)
- [Domain-level keyword usage](#)
- [Domain-level keyword-agnostic features](#)

Here is a brief look at each of these:

Domain-level link authority features

Domain-level link authority is based on a cumulative link analysis of all the links to the domain. This includes factors such as the number of different domains linking to the site, the trust/authority of those domains, the rate at which new inbound links are added, the relevance of the linking domains, and more.

Page-level link metrics

This refers to the links as related to the specific page, such as the number of links, the relevance of the links, and the trust and authority of the links received by the page.

Page-level keywords and content

This describes the use of the keyword term/phrase in particular parts of the HTML code on the page (<title> tag, <h1>, alt attributes, etc.).

Page-level features other than keywords

Factors included here are page elements such as the number of links on the page, number of internal links, number of followed links, number of “nofollow” links, and other similar factors.

¹² For more information, see “2013 Search Engine Ranking Factors,” <https://moz.com/search-ranking-factors>.

Domain-level brand metrics

This factor includes search volume on the website's brand name, mentions, whether it has a presence in social media, and other brand-related metrics.

Page-level traffic/query data

Elements of this factor are click-through rate to the page in the search results, bounce rate of visitors to the page, and other similar measurements.

Page-level social metrics

Social metrics considered include mentions, links, shares, likes, and other social media site-based metrics. It should be emphasized that many SEO practitioners believe that this is a ranking factor even though studies have since shown otherwise, and representatives from Google clearly state that social signals are not part of their algorithm.

Domain-level keyword usage

This refers to how keywords are used in the root or subdomain name, and how impactful that might be on search engine rankings.

Domain-level keyword-agnostic features

Major elements of this factor in the survey include the number of hyphens in the domain name, number of characters in the domain name, and domain name length.

Negative Ranking Factors

It's also possible to have negative ranking factors. For example, if a site has a large number of low-quality inbound links that appear to be the result of artificial efforts by the publisher to influence search rankings, the site's rankings can be lowered. This is, in fact, exactly what Google's *Penguin* algorithm does. This algorithm is discussed more in [Chapter 9](#). Some other potential negative ranking factors include:

Malware being hosted on the site

The search engines will act rapidly to penalize sites that contain viruses or Trojans.

Cloaking

Search engines want publishers to show the same content to the search engine as is shown to users.

Pages on the sites with links for sale

Google has **a strong policy against paid links**, and sites that sell links may be penalized.

Content that advertises paid links on the site

As an extension of the prior negative ranking factor, promoting the sale of paid links may be a negative ranking factor.

Page speed

Back in 2010, Google's Matt Cutts announced that Google was making page speed a ranking factor. In general, it is believed that this is a negative factor for pages that are exceptionally slow.

Other Ranking Factors

The ranking factors we've discussed so far are really just the basics. Search engines potentially factor in many more signals. Some of these include:

Rate of acquisition of links

If, over time, your site has acquired an average of 5 links per day, and then the links suddenly start to come in at a rate of 10 per day, that could be seen as a positive ranking signal. On the other hand, if the rate of new links drops to 2 per day, that could be a signal that your site has become less relevant.

However, it gets more complicated than that. If your site suddenly starts to get 300 new links per day, you have either become a lot more relevant or started to acquire links in a spammy way. The devil is in the details here, with one of the most important details being the origin of those new links. The concept of considering temporal factors in link analysis is documented in a U.S. patent held by Google, which you can look up by [searching for patent number 20050071741](#).

User data

Personalization is one of the most talked-about frontiers in search. There are a few ways in which personalization can take place. For one, a search engine can perform a geolocation lookup to figure out where a user is approximately located. Based on this, the search engine can show results tailored to a user's current location. This is very helpful, for example, if the user is looking for a local restaurant.

Another way a search engine can get some data on a user is if he creates a profile with the search engine and voluntarily provides some information. A simple example would be a language preference. If the user indicates he prefers Portuguese, the search engine can tailor the results to that preference.

Search engines can also look at the search history for a given user. Basically, the search engine maintains a log of all the searches the user has performed when he is logged in. Based on this, it can see that he has been checking out luxury cars recently, and can use that knowledge to tweak the results he sees after he searches on *jaguar*. This is sometimes referred to as *adaptive search*.

To reduce the level of personalization, users can log out of their Google account. However, this does not disable *all* personalization, as Google may still tie some history to the person's computer. A user can disable all personalization by using Google's Chrome browser in *Incognito* mode. This will allow her to see Google results

that are not personalized based on search history. However, the results will still be personalized to her location.

A user can also depersonalize search results by performing her search query, and then appending *?pws=0* to the end of the search page URL and reloading the page. Note, this works only if she has turned off Google Instant (Google's feature of showing results instantly as the user types). Or, the user can choose the option "Disable customizations based on web history" under "webhistory" under the gear icon in the SERPs.

Using Advanced Search Techniques

One of the basic tools of the trade for an SEO practitioner is the search engines themselves. They provide a rich array of search operators that can be used to perform advanced research, diagnosis, and competitive analysis. The following are some of the more basic operators:

-keyword

Excludes the keyword from the search results. For example, *loans -student* shows results for all types of loans *except* student loans.

"key phrase"

Shows search results for the exact phrase—for example, *"seo company"*. You can also use *" "* to force the inclusion of a specific word. This is particularly useful for including *stopwords* (keywords that are normally stripped from a search query because they usually do not add value, such as the word *the*) in a query, or if your keyword is getting converted into multiple keywords through automatic stemming. For example, if you mean to search for the TV show *The Office*, you would want the word *The* to be part of the query. As another example, if you are looking for Patrick Powers, who was from Ireland, you would search for *"patrick powers" Ireland* to avoid irrelevant results for Patrick Powers.

keyword1 OR keyword2

Shows results for *at least one* of the keywords—for example, *google OR Yahoo!*.

These are the basics, but for those who want more information, what follows is an outline of the more advanced search operators available from the search engines.

Advanced Google Search Operators

Google supports a number of **advanced search operators** that you can use to help diagnose SEO issues. **Table 2-1** gives a brief overview of the queries, how you can use them for SEO purposes, and examples of usage.

Table 2-1. *Google's advanced search operators*

Operator	Short description	SEO application	Examples
<i>site:</i>	Domain-restricted search; narrows a search to one or more specific domains or directories.	Shows approximately how many URLs are indexed by Google.	For a website: <i>site:www.google.com</i> From a directory: <i>site.mit.edu/research/</i> Including all subdomains: <i>site:google.com</i> From a specific top-level domain (TLD): <i>site.org</i>
		<ul style="list-style-type: none"> • From a directory. 	
		<ul style="list-style-type: none"> • Including all subdomains. 	
		<ul style="list-style-type: none"> • From a specific top-level domain (TLD). 	
<i>inurl:/allinurl:</i>	URL keyword restricted search; narrows the results to documents containing one or more search terms in the URLs.	Find web pages having your keyword in a filepath.	<i>inurl:seo inurl:company</i> = <i>allinurl:seo company</i>

Operator	Short description	SEO application	Examples
<i>intitle:;/</i> <i>(allintitle:</i>	Title keyword restricted search; restricts the results to documents containing one or more search terms in a page title.	Find web pages using your keyword in a page title.	<i>intitle:seo intitle:company =</i> <i>allintitle:seo company</i>
<i>inanchor:;/</i> <i>allinanchor:</i>	Anchor text keyword restricted search; restricts the results to documents containing one or more search terms in the anchor text of backlinks pointing to a page.	Find pages having the most backlinks/the most powerful backlinks with the keyword in the anchor text.	<i>inanchor:seo inanchor:company =</i> <i>allinanchor:seo company</i>
<i>intext:</i>	Body text keyword restricted search; restricts the results to documents containing one or more search terms in the body text of a page.	Find pages containing the most relevant/most optimized body text.	<i>intext:seo</i>

Operator	Short description	SEO application	Examples
<i>ext:/filetype:</i>	File type restricted search; narrows search results to the pages that end in a particular file extension.	A few possible extensions/file types: <ul style="list-style-type: none"> • <i>.pdf</i> (Adobe Portable Document Format) • <i>.html</i> or <i>.htm</i> (Hypertext Markup Language) • <i>.xlsx</i> (Microsoft Excel) • <i>.pptx</i> (Microsoft PowerPoint) • <i>.docx</i> (Microsoft Word) 	<i>filetype:pdf ext:pdf</i>
<i>related:</i>	Similar URLs search; shows <i>related pages</i> by finding pages linking to the site and looking at what else they tend to link to (i.e., “co-citation”); usually 25–31 results are shown.	Evaluate how relevant the site’s “neighbors” are.	Compare: <i>related:www.searchengineland.com</i> and <i>related:www.alchemistmedia.com</i>

Operator	Short description	SEO application	Examples
<i>info:</i>	Information about a URL search; gives information about the given page.	Learn whether the page has been indexed by Google; provides links for further URL information; this search can also alert you to possible site issues (duplicate content or possible DNS problems).	<i>info:scienceofseo.com</i> will show you the page title and description, and invite you to view its related pages, incoming links, and page cached version.
<i>cache:</i>	What the page looked like when Google crawled it; shows Google's saved copy of the page.	Google's <i>text</i> version of the page works the same way as SEO browser.	<i>cache:www.stonetemple.com</i>

NOTE

When you use the *site:* operator, some indexed URLs might not be displayed (even if you use the “repeat the search with omitted results included” link to see the full list). The *site:* query is notoriously inaccurate. You can obtain a more accurate count of the pages of your site indexed by Google by appending *&start=990&filter=0* to the URL of a Google set for a search using the *site:* operator.

This tells Google to start with result 990, which is the last page Google will show you, as it limits the results to 1,000. This must take place in two steps. First, enter a basic *site:<yourdomain.com>* search, and then get the results. Then go up to the address bar and append the *&start=990&filter=0* parameters to the end of the URL. Once you've done this, you can look at the total pages returned to get a more accurate count. Note that this works only if Google Instant is turned off.

To see more results, you can also use the following search patterns:

- `site:<yourdomain.com>/<subdirectory1> + site:<yourdomain.com>/<subdirectory2>` + etc. (the “deeper” you dig, the more/more accurate results you get)
- `site:<yourdomain.com> inurl:<keyword1> + site:<yourdomain.com> inurl:<keyword2>` + etc. (for subdirectory-specific keywords)
- `site:<yourdomain.com> intitle:<keyword1> + site:<yourdomain.com> intitle:<keyword2>` + etc. (for pages using the keywords in the page title)

To learn more about Google advanced search operators, check out Stephan Spencer’s book *Google Power Search* (O’Reilly).

Combined Google queries

To get more information from Google advanced search, it helps to learn how to effectively combine search operators. [Table 2-2](#) illustrates which search patterns you can apply to make the most of some important SEO research tasks.

Table 2-2. Combined Google search options

What for	Description	Format	Example
Competitive analysis	Find recent mentions of your competitor on other sites; use the date range option under the search tools in the SERPs; the following brand-specific search terms can be used: <code><domainname.com></code> , <code><domain name></code> , <code><domainname></code> , <code><site owner name></code> , and more.	<code><domainname.com></code> - <code>site:<domainname.com></code> To select one day, pick "Search tools" → "Any time" → "Past 24 hours".	<code>moz-site:moz.com</code> <i>during past 24 hours</i>
Keyword research	Evaluate the given keyword competition (sites that apply proper SEO to target the term).	<code>inanchor:<keyword></code> <code>intitle:<keyword></code>	<code>inanchor:seo</code> <code>intitle:seo</code>
SEO site auditing	Find more keyword phrases. Learn whether the site has canonicalization problems. Find the site's most powerful pages.	<code>key * phrase</code> <code>site:<domain.com></code> - <code>inurl:www</code> <code>site:<domain.com></code> - <code>inurl:www</code> <code>inurl:<domainsite></code> - <code><domain.com></code>	<code>free * tools</code> <code>site:stephanspencer.com</code> <code>-inurl:www</code> <code>www</code> <code>site:alchemistmedia.com</code> <code>inurl:stonetemple</code> <code>site:stonetemple.com</code>
Link building	Find the site's most powerful page related to the keyword. Find authority sites offering a backlink opportunity. Search for relevant forums and discussion boards to participate in discussions and probably link back to your site.	<code><domain></code> <code>site:<domain.com></code> <code>site:<domain.com></code> <code><keyword></code> <code>site:<domain.com></code> <code>intitle:<keyword></code> <code>site:<domain></code> <code>inanchor:<keyword></code> <code>site:org</code> <code>bookmarks/links/"favorite sites"/</code> <code>site:gov</code> <code>bookmarks/links/"favorite sites"/</code> <code>[site:edu</code> <code>bookmarks/links/"favorite sites"/</code> <code>inurl:forum</code> <code>OR</code> <code>inurl:forums</code> <code><keyword></code>	<code>alchemistmedia</code> <code>site:alchemistmedia.com</code> <code>site:scienceofseo.com</code> <code>seo</code> <code>site:stonetemple.com</code> <code>intitle:seo</code> <code>site:moz.com</code> <code>inanchor:seo</code> <code>site:org</code> <code>donors</code> <code>inurl:forum</code> <code>OR</code> <code>inurl:forums</code> <code>seo</code>

Firefox plug-ins for quicker access to Google advanced search queries

You can use a number of plug-ins with Firefox to make accessing these advanced queries easier:

- **Advanced Dork**, for quick access to *intitle:*, *inurl:*, *site:*, and *ext:* operators for a highlighted word on a page, as shown in [Figure 2-28](#)
- **SearchStatus**, for quick access to a *site:* operator to explore a currently active domain, as shown in [Figure 2-29](#)

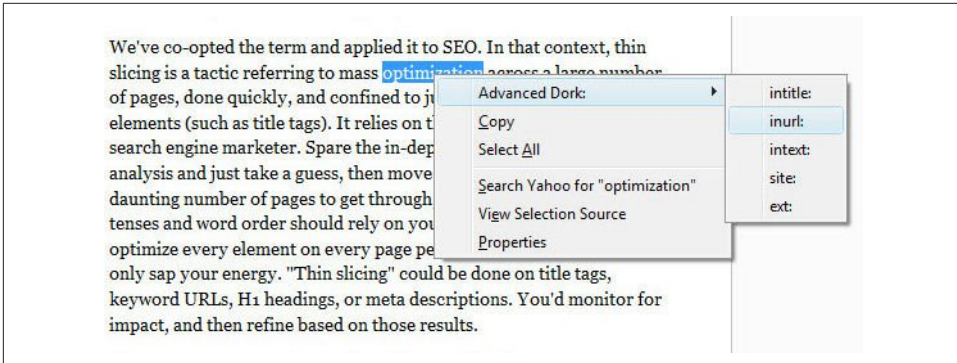


Figure 2-28. Advanced Dork plug-in for Firefox

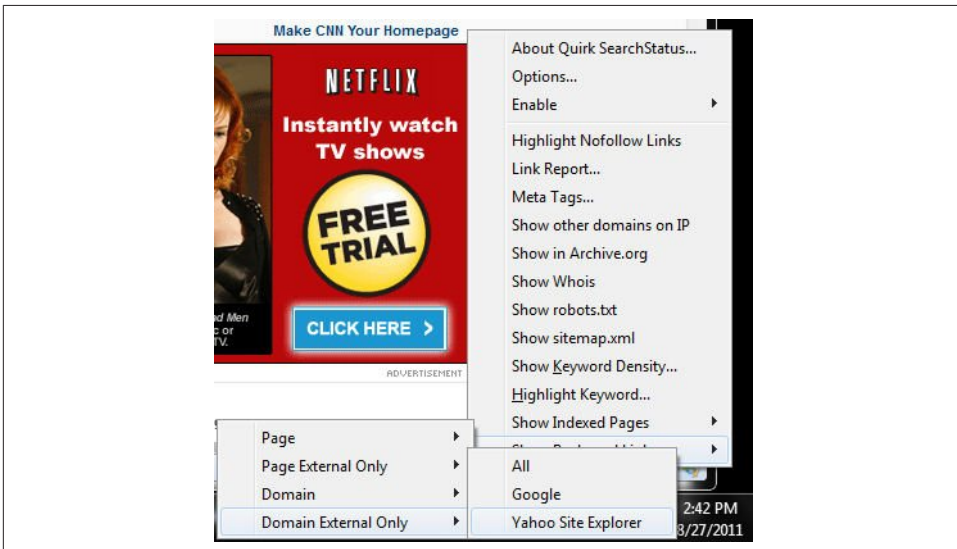


Figure 2-29. SearchStatus plug-in for Firefox

Bing Advanced Search Operators

Bing also offers several unique search operators worth looking into, as shown in [Table 2-3](#).

Table 2-3. *Bing advanced operators*

Operator	Short description	SEO application	Examples
<i>linkfromdomain:</i>	Domain outbound links restricted search; finds all pages the given domain links out to.	Find most relevant sites your competitor links out to.	<i>linkfromdomain:moz.com seo</i>
<i>contains:</i>	File type links restricted search; file type: is also supported; narrows search results to pages linking to a document of the specified file type.	Find pages linking to a specific document type containing relevant information.	<i>contains:wma seo</i>
<i>ip:</i>	IP address restricted search; shows sites sharing one IP address.	<i>ip:xxx.xxx.xxx.xxx</i>	<i>ip:207.182.138.245</i>
<i>inbody:</i>	Body text keyword restricted search; restricts the results to documents containing query word(s) in the body text of a page.	Find pages containing the most relevant/best optimized body text.	<i>inbody:seo</i> (equivalent to Google's <i>intext:</i>)

Operator	Short description	SEO application	Examples
<i>location:loc:</i>	Location-specific search; narrows search results to a specified location (multiple location options can be found under Bing's advanced search).	Find geospecific documents using your keyword.	<i>seo loc:AU</i>
<i>feed:</i>	Feed keyword restricted search; narrows search results to terms contained in RSS feeds.	Find relevant feeds.	<i>feed:seo</i>
<i>hasfeed:</i>	Feed keyword restricted search; narrows search results to pages linking to feeds that contain the specified keywords.	Find pages linking to relevant feeds.	<i>hasfeed:seo site:cnn.com</i>

More Advanced Search Operator Techniques

You can also use more advanced SEO techniques to extract more information.

Determining keyword difficulty

When you are building a web page, it can be useful to know how competitive the keyword is that you are going after, yet this information can be difficult to obtain. However, there are steps you can take to get some idea of how difficult it is to rank for a keyword. For example, the *intitle:* operator shows pages that are more focused on your search term than the pages returned without that operator (e.g., *intitle:"dress boots"*).

You can use different ratios to give you a sense of how competitive a keyword market is (higher results mean that it is more competitive). For example:

- *dress boots* (108,000,000) versus *"dress boots"* (2,020,000) versus *intitle:"dress boots"* (375,000)
- Ratio: $108,000/375 = 290:1$
- Exact phrase ratio: $2,020/37 = 5.4:1$

Another significant parameter you can look at is the *inanchor:* operator—for example, *inanchor:"dress boots"*. You can use this operator in the preceding equation instead of the *intitle:* operator.

Using number ranges

The number range operator can help restrict the results set to a set of model numbers, product numbers, price ranges, and so forth. For example:

site:stevespanglerscience.com "product/1700..1750"

Unfortunately, the number range combined with *inurl:* is not supported, so the product number must be on the page. The number range operator is also great for copyright year searches (to find abandoned sites to acquire). Combine it with the *intext:* operator to improve the signal-to-noise ratio—for example, *intext:"copyright 1993..2011"-2014 blog*.

Using advanced doc type search

The *filetype:* operator is useful for looking for needles in haystacks. Here are a couple of examples:

confidential business plan -template filetype:doc
forrester research grapevine filetype:pdf

Determining listing age

You can label results with dates that give a quick sense of how old (and thus trusted) each listing is; for example, by appending the *as_qdr=m199* parameter to the end of a Google SERP URL, you can restrict results to those within the past 199 months.

Uncovering subscriber-only or deleted content

You can sometimes get to subscriber-only or deleted content from the Cached link in the listing in the SERPs (found under the down arrow after the URL in the search listing) or by using the *cache:* operator. Don't want to leave a footprint? Add *strip=1* to the end of the Google cached URL. Images on the page won't load.

If no Cached link is available, use Google Translate to take your English document and translate it from Spanish to English (this will reveal the content even though no Cached link is available):

<http://translate.google.com/translate?prev=hl=en&u=<URL-GOES-HERE>&sl=es&tl=en>

Identifying neighborhoods

The *related:* operator will look at the sites linking (the *linking sites*) to the specified site, and then see which other sites are commonly linked to by the linking sites. These are commonly referred to as *neighborhoods*, as there is clearly a strong relationship between sites that share similar link graphs.

Finding Creative Commons (CC) licensed content

Use the *as_rights* parameter in the URL to find Creative Commons licensed content. Here are some example scenarios to find CC-licensed material on the Web:

Permit commercial use

```
http://google.com/search?as_rights=(cc_publicdomain\cc_attribute\cc_sharealike\cc_nonderived).-(cc_noncommercial)&q=<KEYWORDS>
```

Permit derivative works

```
http://google.com/search?as_rights=(cc_publicdomain\cc_attribute\cc_sharealike\cc_nonderived).-(cc_nonderived)&q=<KEYWORDS>
```

Permit commercial and derivative use

```
http://google.com/search?as_rights=(cc_publicdomain\cc_attribute\cc_sharealike).-(cc_noncommercial\cc_nonderived)&q=<KEYWORDS>
```

Make sure you replace <KEYWORDS> with the keywords that will help you find content that is relevant to your site. The value of this to SEO is indirect. Creative Commons content can potentially be a good source of content for a website. An easier option if you don't need this same freedom in your Creative Commons searches is to use Google's Advanced Search page, where you can specify your Creative Commons license type.

Vertical Search Engines

Vertical search is a term sometimes used for specialty or niche search engines that focus on a limited data set. Examples of vertical search solutions provided by the major search engines are image, video, news, and blog searches. These may be standard offerings from these vendors, but they are distinct from the engines' general web search functions.

Vertical search engines sometimes come in the form of specialty websites, such as travel sites (such as [TripAdvisor](#)), and local business listing sites (such as [YellowPages.com](#)). Any site that focuses on vertically oriented niche markets could be considered a vertical search engine.

Vertical search results can provide significant opportunities for the SEO practitioner. High placement in these vertical search results can equate to high placement in the

web search results, often above the traditional 10 blue links presented by the search engines.

Vertical Search from the Major Search Engines

The big three search engines offer a wide variety of vertical search products. Here is a partial list:

Google

[Google Maps](#), [Google Images](#), [Google Shopping](#), [Google Blog Search](#), [Google Video](#), [Google News](#), [Google Custom Search Engine](#), [Google Book Search](#)

Yahoo!

[Yahoo! News](#), [Yahoo! Local](#), [Yahoo! Images](#), [Yahoo! Video](#), [Yahoo! Shopping](#), [Yahoo! Autos](#)

Bing

[Bing Images](#), [Bing Videos](#), [Bing News](#), [Bing Maps](#)

Image search

All three search engines offer image search capability. Basically, image search engines limit the data that they crawl, search, and return in results to images. This means files that are in GIF, TIF, JPG, PNG, and other similar formats. **Figure 2-30** shows the image search engine from Bing.

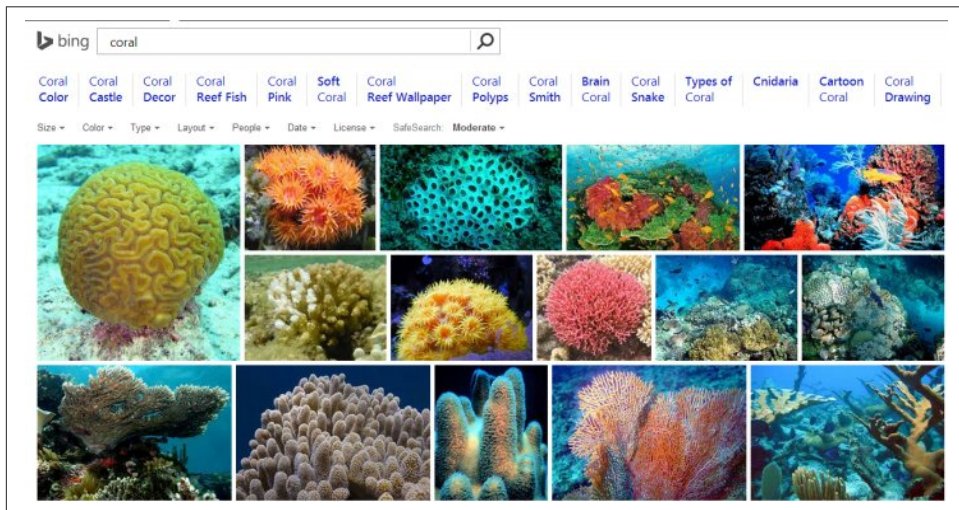


Figure 2-30. Image search results from Bing

Image search engines get a surprisingly large number of searches performed on them. Unfortunately, market data on these volumes is not often published, but according to

comScore, more than 1 billion image searches are performed on [Google Images](#) per month. However, it is likely that at least that many image-related search queries occur within Google web search in the same timeframe. However, because an image is a binary file, it cannot be readily interpreted by a search engine crawler.

[Search engines](#) have had to historically [rely on text surrounding the image](#), the `alt` attribute within the `` tag, and the image filename. However, Google now offers a [search by image feature](#) that allows users to drag an image file into the Google Images search box and it will attempt to identify the subject matter of the image and show relevant results. Optimizing for image search is its own science, and we will discuss it in more detail in [“Optimizing for Image Search” on page 663](#).

Video search

As with image search, video search engines focus on searching specific types of files on the Web—in this case, video files such as MPEG, AVI, and others. [Figure 2-31](#) shows a quick peek at video search results from YouTube.

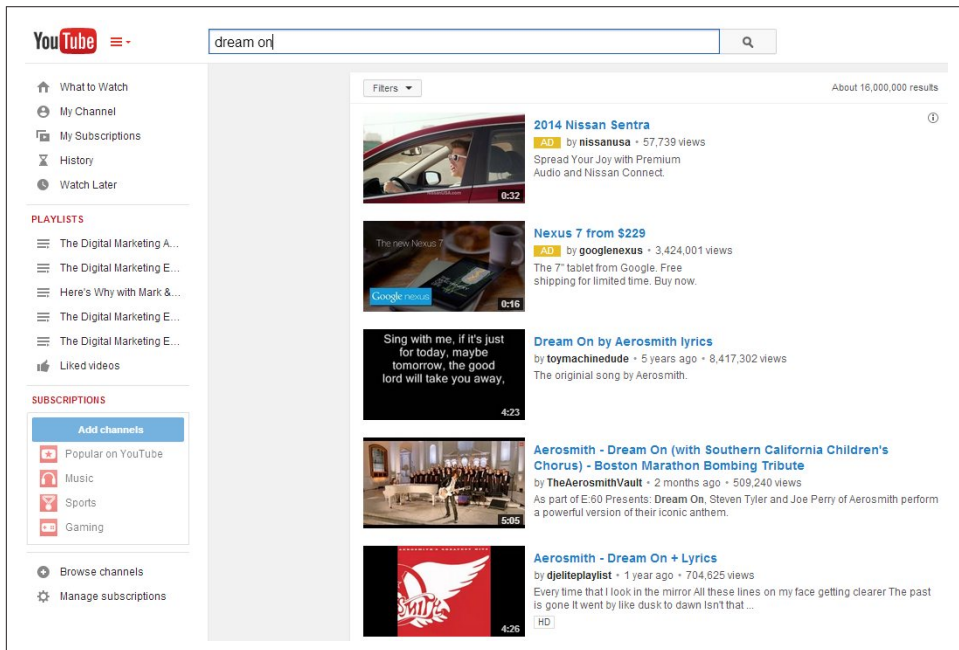


Figure 2-31. Video search results from YouTube

A very large number of searches are also performed in video search engines. [YouTube is the dominant video search engine](#). Current data on total monthly searches is not readily available, but in June 2011, over 3.8 billion searches were performed on YouTube. This suggests that YouTube is the third largest search engine on the Web (Bing is

larger when you consider the cumulative search volume of Bing and Yahoo!). As with image search, many video searches are also performed directly within Google web search.

You can gain significant traffic by optimizing for video search engines and participating in them. Once again, these are binary files and the search engine cannot easily tell what is inside them.

This means optimization is constrained to data in the header of the video and on the surrounding web page. We will discuss video search optimization in more detail in “Others: Mobile, Video/Multimedia Search” in “**Optimizing for Video/Multimedia Search**” on page 694.

However, each search engine is investing in technology to analyze images and videos to extract as much information as possible. For example, the search engines are experimenting with OCR technology to look for text within images, and transcription and other advanced technologies are being used to analyze video content. In addition, flesh-tone analysis is being used to detect pornography or recognize facial features. The application of these technologies is in its infancy, and is likely to evolve rapidly over time.

News search

News search is also unique. News search results operate on a different time schedule; they have to be very, very timely. Few people want to read the baseball scores from a week ago when several other games have been played since then.

News search engines must be able to retrieve information in real time and provide near-instantaneous responses. Modern consumers tend to want their news information now. **Figure 2-32** is a quick look at the results from a visit to Yahoo! News.

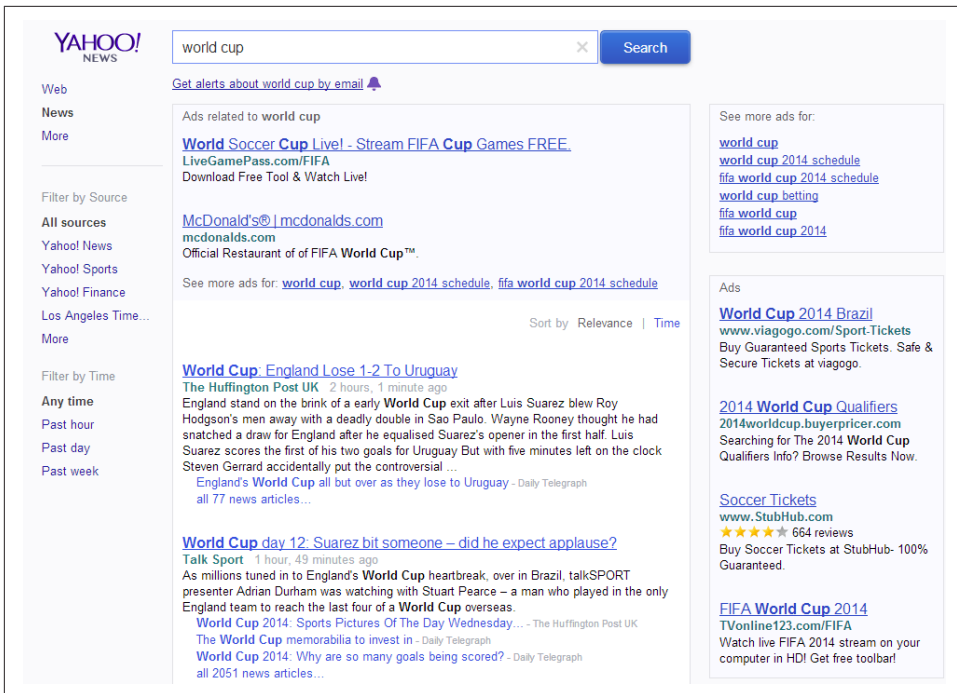


Figure 2-32. News search results from Yahoo!

As with the other major verticals, there is a lot of search volume here as well. To have a chance of receiving this volume, you will need to become a news source. This means timely, topical news stories generated on a regular basis. These and other requirements are discussed further in *“Optimizing for News Search: Google News”* on page 677.

Local search/maps

Next up in our hit parade of major search verticals is local search (a.k.a. map search). Local search results are now heavily integrated into the traditional web search results, so a presence in local search can have a large impact on organizations that have one or more brick-and-mortar locations. Local search engines search through databases of locally oriented information, such as the name, phone number, and location of local businesses around the world, or just provide a service, such as offering directions from one location to another. Figure 2-33 shows Google Maps local search results.

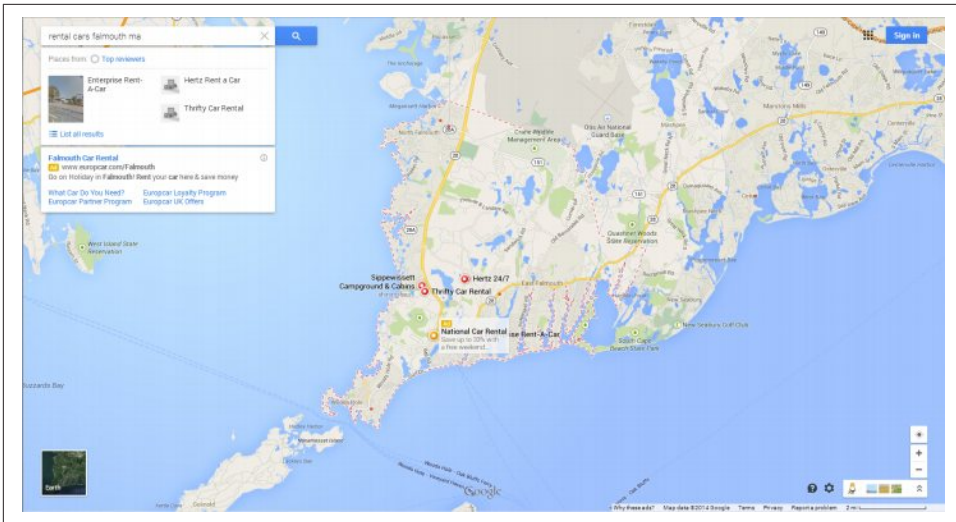


Figure 2-33. Local search results from Google Maps

The integration of local search results into regular web search results has dramatically increased the potential traffic that can be obtained through local search. We will cover local search optimization in detail in [“Optimizing for Local Search”](#) on page 648.

Blog search

Google has implemented a search engine focused just on blog search called Google Blog Search (misnamed because it is an RSS feed engine, not a blog engine). This search engine will respond to queries, but only searches blogs (more accurately, feeds) to determine the results. [Figure 2-34](#) is an example search result for the search phrase *barack obama*.

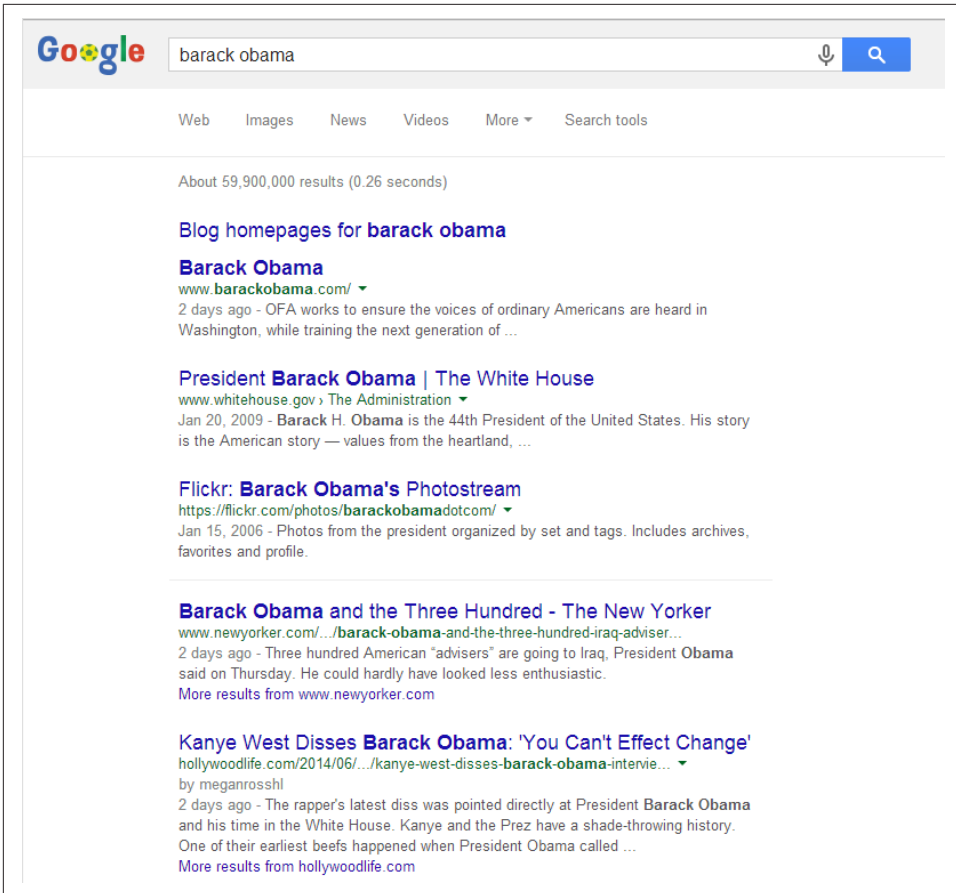


Figure 2-34. Results from Google Blog Search

We explore the subject of optimizing for Google Blog Search in “[Optimizing for Blog Search](#)” on page 673.

Book search

The major search engines also offer a number of specialized offerings. One highly vertical search engine is Google Books search, which specifically searches only content found within books, as shown in [Figure 2-35](#).

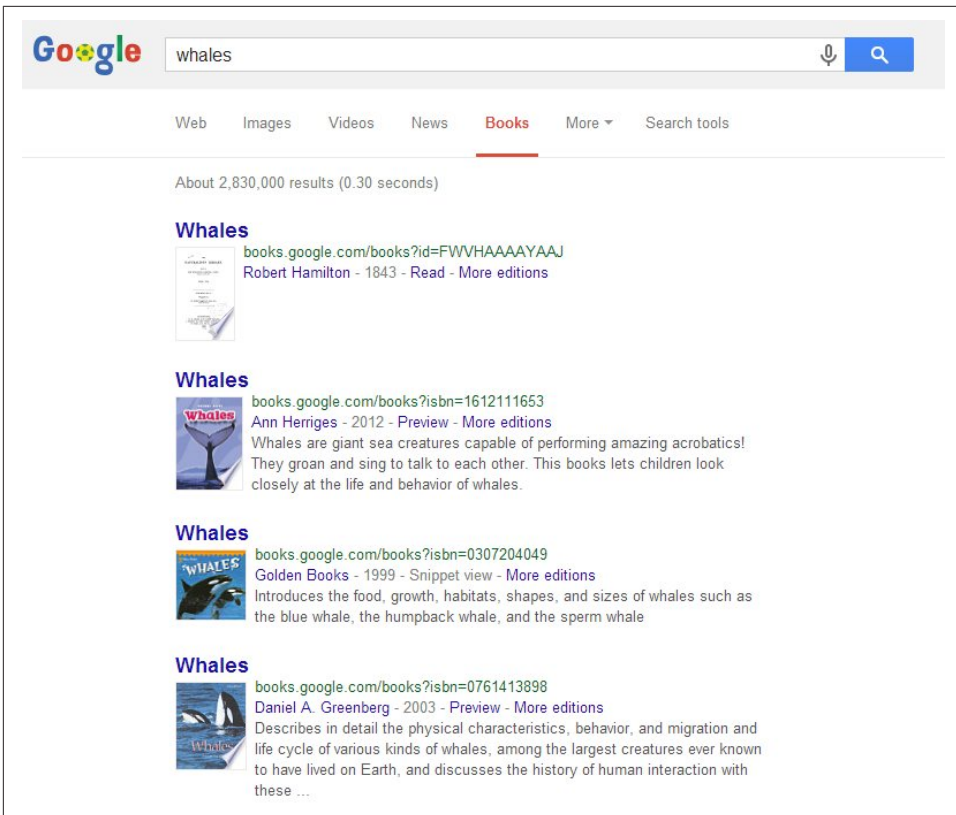


Figure 2-35. Google Books search

Product search

Bing also has some unique vertical search features. One of the more interesting ones is its product search solution. Instead of having a separate shopping search engine, Bing has integrated the product results into the main body of its search results, as shown on the right side of [Figure 2-36](#).

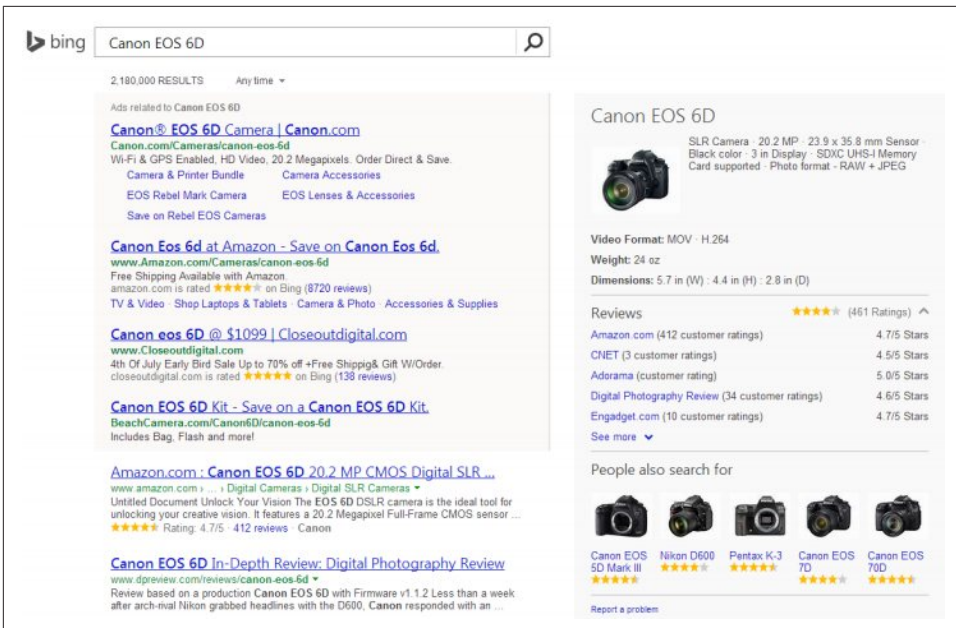


Figure 2-36. Bing product search

Universal Search/Blended Search

Google made a big splash in 2007 when it announced Universal Search, the notion of integrating images, videos, and results from other vertical search properties directly into the main web search results. Prior to this announcement, all the search engines showed their search results in separate vertical search engines. You have already seen an example of this in Figure 2-36, which shows Bing’s way of integrating product search features directly into the main search results.

After Google’s announcement, both Bing and Yahoo! quickly followed with their own implementations. Each type of result you see on a search results page offers different opportunities for obtaining traffic from search engines. People now refer to this general concept as *blended search* (because Universal Search is specifically associated with Google).

Figure 2-37 shows an example of blended search results from a Google search.

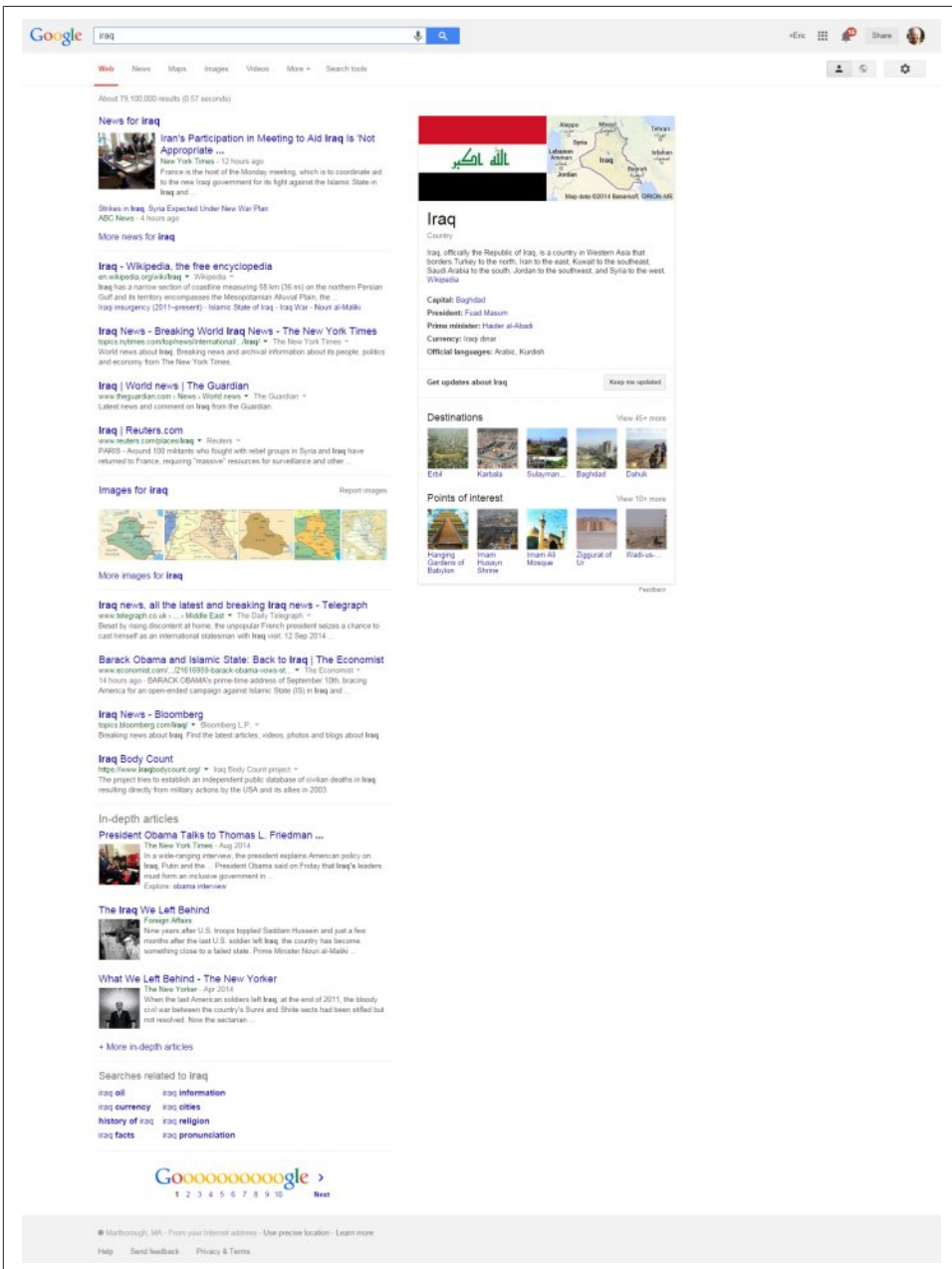


Figure 2-37. Google Universal Search results

More specialized vertical search engines

Vertical search can also come from third parties. Here are some examples:

- [Comparison shopping engines \(e.g., PriceGrabber, Shopzilla, and Nextag\)](#)
- [Travel search engines \(e.g., Expedia, Travelocity, and Kayak\)](#)
- [Real estate search engines \(e.g., Trulia and Zillow\)](#)
- [Job search engines \(e.g., Indeed, CareerBuilder, and SimplyHired\)](#)
- [Music search engines \(e.g., iTunes Music Store\)](#)
- [B2B search engines \(e.g., KnowledgeStorm and ThomasNet\)](#)

There is an enormous array of different vertical search offerings from the major search engines, and from other companies as well. We can expect that this explosion of different vertical search properties will continue.

Effective search functionality on the Web is riddled with complexity and challenging problems. Being able to constrain the data types (to a specific type of file, a specific area of interest, a specific geography, etc.) can significantly improve the quality of the results for users.

Country-Specific Search Engines

At this stage, search is truly global in its reach. [Google is the dominant search engine in many countries, but not all of them.](#) [How you optimize your website depends heavily on the target market for that site,](#) and the one or more search engines that are the most important in that market.

According to comScore, Google receives 54.3% of all searches performed worldwide as of April 2014. In many countries, that market share is 80% or more.

Here is some data on **countries** where other **search engines** are major players:

China

China Internet Watch reported in September 2014 that [Baidu](#) had about 70% market share. This is significant because China has the largest Internet usage in the world, with 618 million users in 2010 according to China Internet Network Information Center.

Russia

According to figures reported by Yandex, the company's market share in Russia comprised about 62% of all searches in April 2014.¹³

¹³ Amy Gesenhues, "Yandex Reports 62% Share Of Russian Search Market With Q1 2014 Revenue Up 36%," Search Engine Land, April 24, 2014, http://bit.ly/russian_search_market.

South Korea

[Naver](#) was estimated to have about 70% market share in South Korea in March 2014.¹⁴

Czech Republic

In January 2014, the Startup Yard blog reported that [Seznam](#) had more than 60% market share in the Czech Republic.¹⁵

Optimizing for Specific Countries

One of the problems international businesses continuously need to address with search engines is identifying themselves as “local” in the eyes of the search engines. In other words, if a search engine user is located in France and wants to see where the wine shops are in Lyon, how does the search engine know which results to show?

Here are a few of the top factors that contribute to international ranking success:

- Owning the proper domain extension (e.g., [.com.au](#), [.uk](#), [.fr](#), [.de](#), [.nl](#)) for the country that your business is targeting
- Hosting your website in the country you are targeting (with a country-specific IP address)
- Registering with local search engines:
 - [Google My Business](#)
 - [Yahoo! Small Business](#)
 - [Bing Places](#)
- Having other sites from the same country link to you
- Using the native language on the site (an absolute requirement for usability)
- Helping Google serve the correct language or regional URL in the search results by adding the hrefLang attribute (<https://support.google.com/webmasters/answer/189077?hl=en>)
- Placing your relevant local address data on major pages of the site
- Setting your geographic target in Google Search Console (you can read more about this at http://bit.ly/country_targeting); note that Google does not really need

14 kmc, “Should Korean Search Engine Naver Worry About Google?,” <http://www.korea-marketing.com/should-naver-worry-about-google/>.

15 Lloyd Waldo, “Meet the Only Company in Europe that is Beating Google Seznam.cz,” Startup Yard Blog, January 3, 2014, http://bit.ly/seznam_europe.

you to do this if your site is on a country code top-level domain (ccTLD), such as *.de* or *.co.uk*, as the preferred regional target is assumed

All of these factors act as strong signals to the search engines regarding the country you are targeting, and will make them more likely to show your site as a relevant local result.

The complexity increases when you are targeting multiple countries. We will discuss this in more depth in [“Best Practices for Multilanguage/Country Targeting” on page 375](#).

Conclusion

Understanding how search engines work is an important component of SEO. The search engines are constantly tuning their algorithms. For that reason, the successful SEO professional is constantly studying search engine behavior and learning how search engines work.