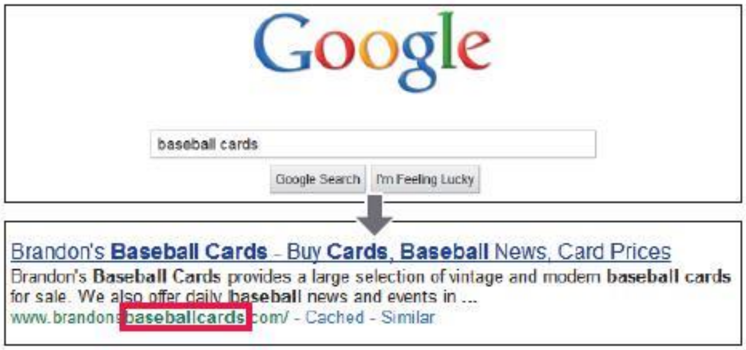
How Search Engines Work



Brandon's **Baseball Cards** - Buy **Cards**, **Baseball** News, Card Prices
Brandon's **Baseball Cards** provides a large selection of vintage and modern **baseball cards** for sale. We also offer daily **baseball** news and events in ...
www.brandons**baseballcards**.com/ - Cached - Similar

(3) A user performs the query [baseball cards]. Our homepage appears as a result, with the URL listed under the title and snippet.

How Search Works - GOOGLE

These processes lay the foundation — they're how we gather and organize information on the web so we can return the most useful results to you. Our index is well over 100,000,000 gigabytes, and we've spent over one million computing hours to build it. Learn more about the basics in this short video.

Finding information by crawling

We use software known as "web crawlers" to discover publicly available webpages. The most well-known crawler is called "Googlebot." Crawlers look at webpages and follow links on those pages, much like you would if you were browsing content on the web. They go from link to link and bring data about those webpages back to Google's servers.

The crawl process begins with a list of web addresses from past crawls and sitemaps provided by website owners. As our crawlers visit these websites, they look for links for other pages to visit. The software pays special attention to new sites, changes to existing sites and dead links.

Computer programs determine which sites to crawl, how often, and how many pages to fetch from each site. Google doesn't accept payment to crawl a site more frequently for our web search results. We care more about having the best possible results because in the long run that's what's best for users and, therefore, our business.

Choice for website owners

Most websites don't need to set up restrictions for crawling, indexing or serving, so their pages are eligible to appear in search results without having to do any extra work. That said, site owners have many choices about how Google crawls and indexes their sites through Webmaster Tools and a file called "robots.txt". With the robots.txt file, site owners can choose not to be crawled by Googlebot, or they can provide more specific instructions about how to process pages on their sites.

Site owners have granular choices and can choose how content is indexed on a page-by-page basis. For example, they can opt to have their pages appear without a snippet (the summary of the page shown below the title in search results) or a cached version (an alternate version stored on Google's servers in case the live page is unavailable). Webmasters can also choose to integrate search into their own pages with Custom Search.

Organizing information by indexing

The web is like an ever-growing public library with billions of books and no central filing system. Google essentially gathers the pages during the crawl process and then creates an index, so we know exactly how to look things up. Much like the index in the back of a book, the Google index includes information about words and their locations. When you search, at the most basic level, our algorithms look up your search terms in the index to find the appropriate pages.

The search process gets much more complex from there. When you search for "dogs" you don't want a page with the word "dogs" on it hundreds of times. You probably want pictures, videos or a list of breeds. Google's indexing systems note many different aspects of pages, such as when they were published, whether they contain pictures and videos, and much more. With the Knowledge Graph, we're continuing to go beyond keyword matching to better understand the people, places and things you care about.

Webmaster Tools

To learn about the tools and resources available to site owners, visit Webmaster Central.

How Search Works handout

Check out a graphic illustrating the various phases of the search process, from before you search, to ranking, to serving results.

Algorithms

You want the answer, not trillions of webpages. Algorithms are computer programs that look for clues to give you back exactly what you want.

For a typical query, there are thousands, if not millions, of webpages with helpful information. Algorithms are the computer processes and formulas that take your questions and turn them into answers. Today Google's algorithms rely on more than 200 unique signals or "clues" that make it possible to guess what you might really be looking for. These signals include things like the terms on websites, the freshness of content, your region and PageRank.

Search Projects

There are many components to the search process and the results page, and we're constantly updating our technologies and systems to deliver better results. Many of these changes involve exciting new innovations, such as the Knowledge Graph or Google Instant. There are other important systems that we constantly tune and refine. This list of projects provides a glimpse into the many different aspects of search.

Answers
Displays immediate answers and information for things such as the weather, sports scores and quick facts.

Autocomplete
Predicts what you might be searching for. This includes understanding terms with more than one meaning.

Books
Finds results out of millions of books, including previews and text, from libraries and publishers worldwide.

Freshness
Shows the latest news and information. This includes gathering timely results when you're searching specific dates.

Google Instant
Displays immediate results as you type.

Images
Shows you image-based results with thumbnails so you can decide which page to visit from just a glance.

Indexing
Uses systems for collecting and storing documents on the web.

Knowledge Graph
Provides results based on a database of real world people, places, things, and the connections between them.

Mobile
Includes improvements designed specifically for mobile devices, such as tablets and smartphones.

News
Includes results from online newspapers and blogs from around the world.

Query Understanding

Gets to the deeper meaning of the words you type.

Refinements
Provides features like "Advanced Search," related searches, and other search tools, all of which help you fine-tune your search.

SafeSearch
Reduces the amount of adult web pages, images, and videos in your results.

Search Methods
Creates new ways to search, including "search by image" and "voice search."

Site & Page Quality
Uses a set of signals to determine how trustworthy, reputable, or authoritative a source is. (One of these signals is PageRank, one of Google's first algorithms, which looks at links between pages to determine their relevance.)

Snippets
Shows small previews of information, such as a page's title and short descriptive text, about each search result.

Spelling
Identifies and corrects possible spelling errors and provides alternatives.

Synonyms
Recognizes words with similar meanings.

Translation and Internationalization
Tailors results based on your language and country.

Universal Search
Blends relevant content, such as images, news, maps, videos, and your personal content, into a single unified search results page.

User Context
Provides more relevant results based on geographic region, Web History, and other factors.

Videos
Shows video-based results with thumbnails so you can quickly decide which video to watch.

The Evolution of Search

Our goal is to get you to the answer you're looking for faster, creating a nearly seamless connection between you and the knowledge you seek. If you're looking to deepen your understanding of how search has evolved, this video highlights some important features like universal results and quick answers.

Experiments: From Idea to Launch
A typical algorithmic change begins as an idea from one of our engineers about how to improve search. We take a data-driven approach and all proposed algorithm changes undergo extensive quality evaluation before release. Engineers typically start by running a series of experiments, tweaking small variables and getting feedback from colleagues until they are satisfied and ready to release the experiment to a larger audience.

Search Quality Rating Guidelines
This document is a version of our Search Quality Rater Guidelines, which gives evaluators examples and guidelines for appropriate ratings. The document focuses on a type of rating task called "URL rating." In this kind of task, the evaluator looks at a search query and a result that could be returned. They rate the relevance of the result for that

query on a scale described within the document. Sounds simple, right? As you can see, there are many tricky cases to think through.
Download Now (English only)

Precision Evaluations
118,812
The first phase is to get feedback from evaluators, people who evaluate search quality based on our guidelines. We show evaluators search results and ask them to rate the usefulness of the results for a given search. *Note: These ratings don't directly impact ranking.*

Side-by-Side Experiments
10,391
In a side-by-side experiment, we show evaluators two different sets of search results: one from the old algorithm and one from the new, and we ask them for details about which results they prefer.

Live Traffic Experiments
7,018
If the evaluators' feedback looks good, we move forward with a "live traffic experiment." In these experiments, we change search for a small percentage of real Google users and see how it changes the way they interact with the results. We carefully analyze the results to understand whether the change is an improvement to the search results. For example, do searchers click the new first result more often? If so, that's generally a good sign.

Launches
665
Finally, our most experienced search engineers carefully review the data from all the different experiments and decide if the change is approved to launch. It sounds like a lot, but the process is well refined, so an engineer can go from idea to live on Google for a percentage of users in 24 hours. Based on all of this experimentation, evaluation and analysis, we launched 665 improvements to search in 2012.

Fighting Spam
Every day, millions of useless spam pages are created. We fight spam through a combination of computer algorithms and manual review.

Spam sites attempt to game their way to the top of search results through techniques like repeating keywords over and over, buying links that pass PageRank or putting invisible text on the screen. This is bad for search because relevant websites get buried, and it's bad for legitimate website owners because their sites become harder to find. The good news is that Google's algorithms can detect the vast majority of spam and demote it automatically. For the rest, we have teams who manually review sites.

Identifying Spam
Spam sites come in all shapes and sizes. Some sites are automatically-generated gibberish that no human could make sense of. Of course, we also see sites using subtler spam techniques. Check out these examples of "pure spam," which are sites using the most aggressive spam techniques. This is a stream of live spam screenshots that we've manually identified and recently removed from appearing in search results.

*We've removed some pornographic content and malware from this demo, but otherwise this is an unfiltered stream of fresh English examples of "pure spam" removals.

Types of Spam
In addition to spam shown above, here are some other types of spam that we detect and take action on.

Cloaking and/or sneaky redirects
Site appears to be cloaking (displaying different content to human users than is shown to search engines) or redirecting users to a different page than Google saw.

Hacked site
Some pages on this site may have been hacked by a third party to display spammy content or links. Website owners should take immediate action to clean their sites and fix any security vulnerabilities.

Hidden text and/or keyword stuffing
Some of the pages may contain hidden text and/or keyword stuffing.

Parked domains
Parked domains are placeholder sites with little unique content, so Google doesn't typically include them in search results.

Pure spam
Site appears to use aggressive spam techniques such as automatically generated gibberish, cloaking, scraping content from other websites, and/or repeated or egregious violations of Google's Webmaster Guidelines.

Spammy free hosts and dynamic DNS providers
Site is hosted by a free hosting service or dynamic DNS provider that has a significant fraction of spammy content.

Thin content with little or no added value
Site appears to consist of low-quality or shallow pages which do not provide users with much added value (such as thin affiliate pages, doorway pages, cookie-cutter sites, automatically generated content, or copied content).

Unnatural links *from* a site
Google detected a pattern of unnatural, artificial, deceptive or manipulative outbound links on this site. This may be the result of selling links that pass PageRank or participating in link schemes.

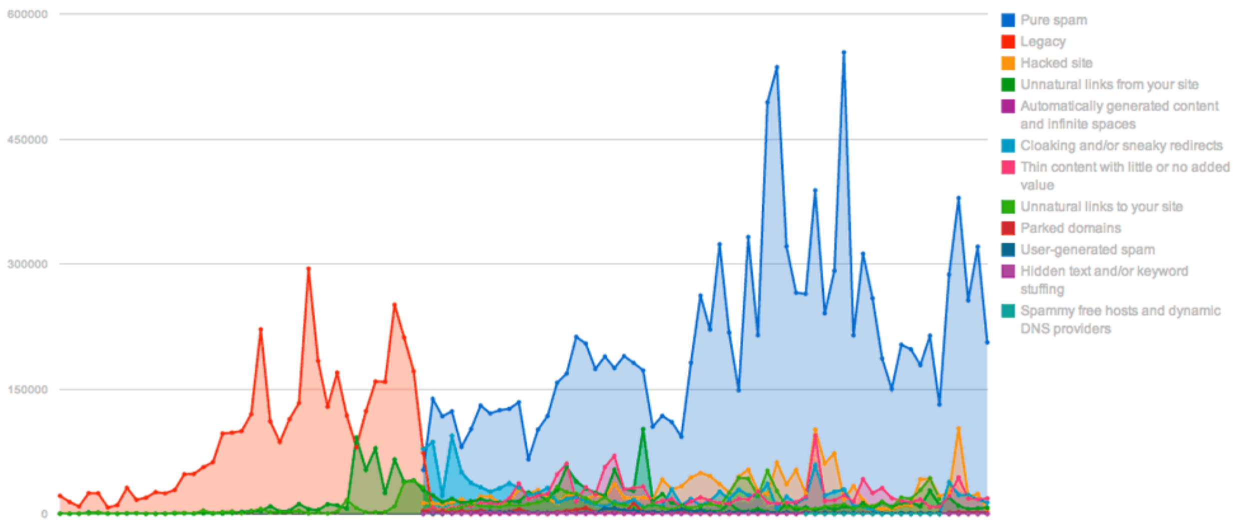
Unnatural links *to* a site
Google has detected a pattern of unnatural artificial, deceptive or manipulative links pointing to the site. These may be the result of buying links that pass PageRank or participating in link schemes.

User-generated spam
Site appears to contain spammy user-generated content. The problematic content may appear on forum pages, guestbook pages, or user profiles.

Taking Action

While our algorithms address the vast majority of spam, we address other spam manually to prevent it from affecting the quality of your results. This graph shows the number of domains that have been affected by a manual action over time and is broken down by the different spam types. The numbers may look large out of context, but the web is a really big place. A recent snapshot of our index showed that about 0.22% of domains had been manually marked for removal.

Manual Action by Month



Milestones for manual spam fighting

February 2005
We expanded our manual spam-fighting team to Hyderabad, India.

March 2005
We expanded our manual spam-fighting team to Dublin, Ireland.

April 2006
We expanded our manual spam-fighting team to Tokyo, Japan.

June 2006
We expanded our manual spam-fighting team to Beijing, China.

October 2007 - Legacy
In the fall of 2007, we changed our classification system to keep data in a more structured format based on the type of webspam violation (which allowed us to create this chart). Actions that couldn't be categorized appropriately into the new system are in the “legacy” category. We were still taking action on spam types like thin affiliates and cloaking prior to this time, but the breakdown by spam type isn't readily available for the older data.

October 2009 - Unnatural links from your site
Improvements in our systems allowed us to reduce the number of actions taken on sites with unnatural outbound links.

November 2009 - Hacked sites
We noticed an increase in hacked sites and increased our efforts to prevent them from affecting search results.

February 2011 - Spammy free hosts and dynamic DNS providers
We increased enforcement of a policy to take action on free hosting services and dynamic DNS providers when a large fraction of their sites or pages violate our Webmaster Guidelines. This allows us to protect our users from seeing spam, when taking action on the individual spammy accounts would be impractical.

October 2011 - Cloaking and/or sneaky redirects

We made a change to our classification system so that the majority of cloaking and sneaky redirect actions were labeled as "Pure spam." Actions related to less egregious violations continue to be labeled separately.

October 2011 - Parked domains

We reduced our efforts to manually identify parked domains due to improvements in our algorithmic detection of these sites.

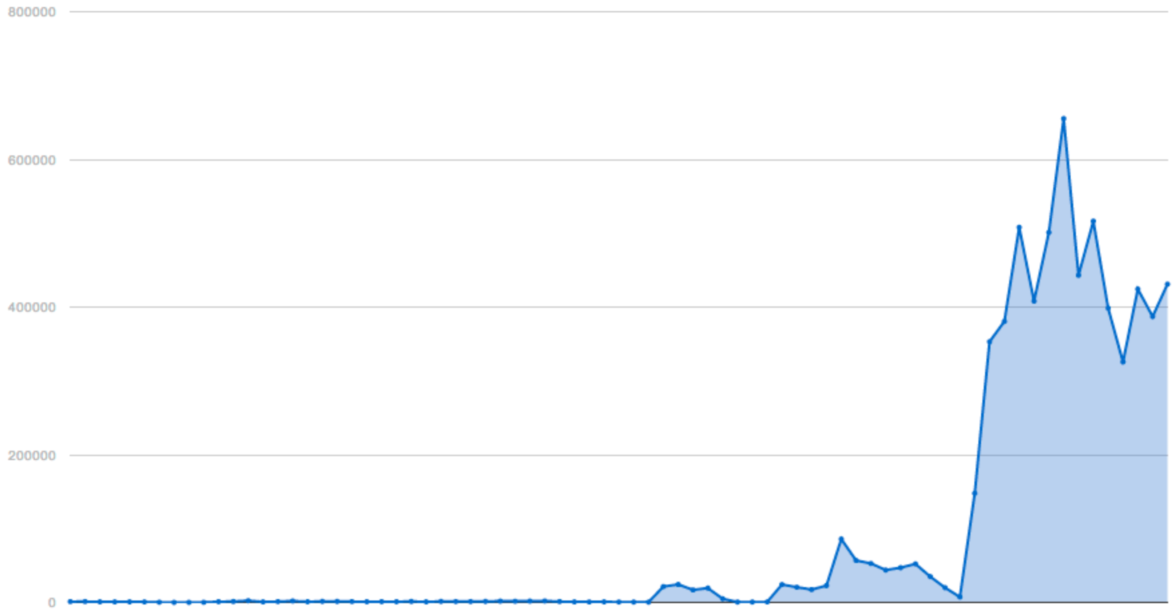
April 2012

We launched an algorithmic update codenamed "Penguin" which decreases the rankings of sites that are using webspam tactics.

Notifying Website Owners

When we take manual action on a website, we try to alert the site's owner to help him or her address issues. We want website owners to have the information they need to get their sites in shape. That's why, over time, we've invested substantial resources in webmaster communication and outreach. The following graph shows the number of spam notifications sent to site owners through Webmaster Tools.

Messages by Month



History of webmaster communication

May 2007

We used to send notifications only via email, and in 2007 webmasters reported receiving fake notifications of Webmaster Guidelines violations. We temporarily paused our notifications in response to this incident while we worked on a new notification system.

July 2007

With the launch of the Message Center feature in Webmaster Tools, we resumed sending notifications in July 2007 after pausing the notifications in May due to email spoofing.

March 2010
We began using a new notification system which enabled us to more easily send messages to the Message Center of Webmaster Tools when we found spam. The first category of spam to use this new system was hacked sites.

July 2010
A bug in our hacked sites notification system reduced the number of messages that we sent to hacked sites.

November 2010
We upgraded our notification system. With this update, we fixed the hacked sites notification bug and began experimenting with sending messages for additional categories of spam such as unnatural links from a site.

February and March 2011
We expanded notifications to cover additional types of unnatural links to a site.

June 2011
We expanded the number of languages we send many of our messages in.

September 2011
We made a change to our classification system for spam. Messages for some categories of spam were not sent, while we created and translated new messages to fit the new categories.

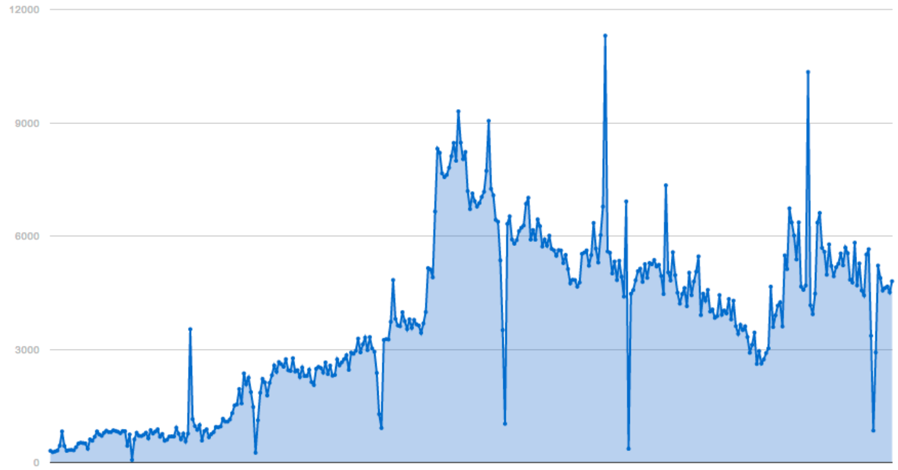
November 2011
A bug in our hacked sites notification system reduced the number of messages that we sent to hacked sites.

December 2011
We expanded the categories of spam that we send notifications for to include pure spam and thin content.

February 2012
The bug affecting our hacked sites notifications was fixed.

Listening for Feedback
Manual actions don't last forever. Once a website owner cleans up her site to remove spammy content, she can ask us to review the site again by filing a reconsideration request. We process *all* of the reconsideration requests we receive and communicate along the way to let site owners know how it's going.
Historically, most sites that have submitted reconsideration requests are not actually affected by any manual spam action. Often these sites are simply experiencing the natural ebb and flow of online traffic, an algorithmic change, or perhaps a technical problem preventing Google from accessing site content. This chart shows the weekly volume of reconsideration requests since 2006.

Reconsideration Requests by Week



Notable moments for reconsideration requests

December 2006
A bug prevented us from properly storing reconsideration requests for about a week. On December 25th (Christmas), we submitted requests on behalf of sites affected by the bug, creating a small spike at the end of the year.

May/June 2007
Many webmasters received fake notifications of Webmaster Guidelines violations, leading an unusual number to file reconsideration requests.

December 2007
Every year webmasters submit fewer reconsideration requests during the late December holidays.

April 2009
We released a video with tips for reconsideration requests.

June 2009
We started sending responses to reconsideration requests to let webmasters know that their requests have been processed.

October 2010
We upgraded our notification system and starting sending out more messages.

April 2011
We rolled out the Panda algorithm internationally. In the past, sites have often filed reconsideration requests when they see traffic changes that aren't actually due to manual action.

April - Sept 2011
We started sending reconsideration responses with more information about the outcomes of reconsideration requests.

June 2012
We began sending messages for a wider variety of webspam issues. We now send notifications for all manual actions by the webspam team which may directly affect a site's ranking in web search results.

-END-

About the Curator:

Lisa Chapman helps company leaders define, plan and achieve their goals, both online and offline. After 25+ years as an entrepreneur, she is now a business and marketing consultant, business planning consultant and social media consultant. Online, she works with clients to establish and enhance their online brand, attract their target market, engage them in meaningful social media conversations, and convert online traffic into revenues. Email: Lisa (at) LisaChapman (dot) com.

Her book, ***The WebPowered Entrepreneur - A Step-by-Step Guide*** is available at:

- Amazon.com: http://bit.ly/AmazonTheWebPoweredEntrepreneur
- Barnes & Noble: http://bit.ly/BNTheWebPoweredEntrepreneur