

CHECK OUT THE NEW PORTENT EBOOK PLUG & PLAY CONTENT STRATEGY: EXERCISES TO PLAN KILLER CONTENT  
([HTTP://ESSENTIALS.PORTENT.COM/PLUG-AND-PLAY-CONTENT-STRATEGY/?  
UTM\\_SOURCE=WEBSITE&UTM\\_MEDIUM=SUPERHEADER&UTM\\_CAMPAIGN=PLUG-AND-PLAY-CONTENT-  
STRATEGY&UTM\\_TERM=CONTENT%20STRATEGY](http://essentials.portent.com/plug-and-play-content-strategy/?utm_source=website&utm_medium=superheader&utm_campaign=plug-and-play-content-strategy&utm_term=content%20strategy))

SERVICES ([HTTPS://WWW.PORTENT.COM/SERVICES](https://www.portent.com/services))

---

PHILOSOPHY ([HTTPS://WWW.PORTENT.COM/ADVANTAGE/PHILOSOPHY](https://www.portent.com/advantage/philosophy))

---

# 3 Google Algorithms We Know About & 200 We Don't

PEOPLE ([HTTPS://WWW.PORTENT.COM/ADVANTAGE/OUR-PEOPLE](https://www.portent.com/advantage/our-people))

PORTENT STAFF // MAY 15, 2013  
WORK ([HTTPS://WWW.PORTENT.COM/PORTFOLIO](https://www.portent.com/portfolio))

---

When I meet with clients or present at conferences, I am always asked: “How do I rank high in Google (keyword-phrases-du-jour)?” I give the standard answer: “Only the search engineers and Google can tell you and they aren’t talking.”

Inevitably, the questioner looks dejected, mutters a slur on my credentials, and walks away. I scream silently in my head: “Don’t kill the messenger because we are all hapless Wile E. Coyotes chasing the Larry and Sergey Road Runner with zero chance of catching them, no matter what we order from ACME!”

Thirteen years ago, before the Cone of Silence dropped on Google’s method of operation, we got a glimpse of the method behind their madness. This, combined with the common knowledge of the foundational tenets of all search engines, gives us some idea of what’s going on behind that not-so-simple box on the white page.

In this post, I am going to explore the 3 algorithms that we know for sure Google is using to produce search results, and speculate about the 200+ other algorithms that we suspect they are using based on patent filings, reverse engineering, and the Ouija board.

**What is an algorithm (you might ask)?**  
SHARES

There are many definitions of algorithm. The National Institute of Standards and Technology defines an algorithm as “a computable set of steps to achieve a desired result.” Ask a developer and they will tell you that an algorithm is “a set of instructions (procedures or functions) that is used to accomplish a certain task.” My favorite definition, and the one that I’m going with, comes from MIT’s Kevin Slavin’s TED Talk “How Algorithms Shape Our World” ([http://www.ted.com/talks/kevin\\_slavin\\_how\\_algorithms\\_shape\\_our\\_world.html](http://www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world.html)): algorithms are “math that computers use to decide stuff.”

## 3 Google algorithms we know about

### PageRank

The most famous Google algorithm is PageRank, a pre-query value that has no relationship to the search query. In its infancy, the PageRank algorithm used links pointing to the page as an indication of its importance. Larry Page, after whom the algorithm is named, used the academic citation model where the papers citing another were endorsements of its authority. Strangely enough, they do not have citation rings or citation buying schemes as with web links. Warning, scary, eye-bleeding computational math ahead.

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

Initial PageRank algorithm

To combat spam, a **Random Surfer algorithm** was added was added to PageRank. This algorithm “imagined” a Random Surfer that traveled the Web and would follow the links on each page. However, sometimes, the Random Surfer would arbitrarily, much like us thought-processing bipeds, not return to the original page and keep going or would stop following links and “jump” to another page. The algorithm steps are:

1. *At any time  $t$ , surfer is on some page  $P$*
2. *At time  $t+1$ , the surfer follows an outlink at random*
3. *Surfer ends up on some page  $Q$  (from page  $P$ )*
4. *The process repeats indefinitely*

That's the benefit of algorithms, no overtime and they never get tired or bored.

## **Hilltop Algorithm**

Surf's up Dude algorithm worked for about 10 minutes before the SEO community found the hole in its wet suit to manipulate rankings. In the early 2000s, processors caught up to computational mathematics and Google was able to deploy the Hilltop Algorithm (around 2001). This algorithm was the first introduction of semantic influence on search results inasmuch as a machine can be trained to understand semantics (<http://en.wikipedia.org/wiki/Semantics>).

Hilltop is like a linguistic Ponzi scheme that attributes a quality to links based on the authority of the document pointing the link to the page. One of Hilltop's algorithms segments the web into a corpus of broad topics. If certain documents in a topic area have lots of links from unaffiliated experts within the same topic area, that document must be an authority. Links from authority documents carry more weight. Authority documents tend to link to other authorities on the same subject and to Hubs, pages that have lots of links to documents on the same subject.

## **Topic-Sensitive PageRank**

The Topic-Sensitive PageRank algorithm is a set of algorithms that take the semantic reasoning a few steps further. Ostensibly the algorithm uses the Open Directory ontology ([dmoz.org](http://dmoz.org)) to sort documents by topic.

Another algorithm calculates a score for context sensitive relevance rank based on a set of "vectors". These vectors represent the context of term use in a document, the context of the term used in the history of queries, and the context of previous use by the user as contained in the user profile.

So, I know what you're thinking. How can they do that for the whole web? They don't. They use predictive modeling algorithms to perform these operations on a representational subset of the web, collect the vectors, and apply the findings to all of the "nearest neighbors."

**D'oh!**

[Added May 16, 2013]

There are a lot of algorithms for indexing, processing and clustering documents that I left out because including them would have many of you face-first-in-your cereal-from-boredom. However, it is NOT OK to leave out the mother of all information retrieval algorithms, TF-IDF, known affectionately to search geeks as Term Frequency-Inverse Document Frequency.

Introduced in the 1970s, this primary ranking algorithm uses the presence, number of occurrences, and locations of occurrence to produce a statistical weight on the importance of a particular term in the document. It includes a normalization feature to prevent long boring documents from taking up residence in search results due to the sheer nature of their girth. This is my favorite algorithm because it supports Woody Allen's maxim that 80% of success is showing up.

## **The 200+ we don't know about**

All of the search engines closely guard their complete algorithm structure for ranking documents. However, we live in a wonderful country that has patent protection for ideas. These patents provide insight into Google's thinking and you can usually pinpoint which ones are deployed.

Panda, the most famous update is an evolving set of algorithms that are combined to determine the quality of the content and user experience on a particular website. There are algorithms that apply decision trees to large data sets of user behavior.

These decision trees look at if this/then that:

- If the page has crossed a threshold a certain ratio of images to text, then it is a poor user experience.
- If a significant portion of searchers do not engage with anything on the page (links, navigation, interaction points), then the page is not interesting for searchers using that particular query.
- If the searchers do not select the page from the results set, then it is not relevant to the query.
- If the searcher returns to the search engine results to select another result or refine the search, then the content was not relevant and not a good user experience.

Complementing the decisions trees could be any one of a number of page layout algorithms that determine the number and placement of images on a page in relation to the amount of content in relation to a searcher's focus of attention.

Following on the heels of Panda are the Penguin algorithms. These algorithms are specifically targeted at detecting and removing web spam. They use Google's vast data resources to evaluate the quality of links pointing to a site, measure the rate of link acquisition, the link source relationship to the page subject, shared domain ownership of the linking sites, and relationships between the linking sites.

Once a site passes an established threshold, another algorithm likely flags the site for additional review by a human evaluator or automatically re-ranks the page so that it drops in search results.

## **Let's stop, guess, and go with what we know**

As with the formula for Coca-Cola or the recipe for Colonel Sanders' Kentucky Fried Chicken, specifics on what Google uses to decide who gets where in the search results set are a closely guarded secret. Instead of speculating on what we might know, let's focus on what we do know:

- In order to rank for a term, that term must be present in the document. Sometimes, a contextual or semantic match for a term will get you into the SERP swimsuit competition for placement. Don't count on that though.
- Being picky and realistic about what you want to rank for is the best start.

- Text on the page drives inclusion in the results for a searcher's query. Be about *some* thing instead of many things.
- Quality content is focused, fresh and engaging.
- Custom titles that describe the page using contextually relevant keywords are THE low hanging fruit. Pick it for your most important pages.
- Compelling description text in search results will draw clicks. Meeting the searcher's expectations with rich content will keep them on the page.
- Pictures, images, and ads are most effective when used in moderation.
- Links are important, but only the right kind. For Google, the "right" kinds are links from pages about the same subject and place high in contextually-related searches.

Are there any major algorithms we missed? Let us know in the comments.

Tags

:

DIGITAL  
MARKETING ([HTTPS://WWW.PORTENT.COM/TAG/INTERNET-MARKETING-2](https://www.portent.com/tag/internet-marketing-2))

GOOGLE ([HTTPS://WWW.PORTENT.COM/TAG/GOOGLE](https://www.portent.com/tag/google))

SEARCH  
MARKETING ([HTTPS://WWW.PORTENT.COM/TAG/SEARCH-MARKETING](https://www.portent.com/tag/search-marketing))

SEO ([HTTPS://WWW.PORTENT.COM/TAG/SEO](https://www.portent.com/tag/seo))



**Portent Staff**  
**([Https://Www.Portent.Com/Author/Portent-Staff](https://www.portent.com/author/portent-staff))**

Read More (<https://www.portent.com/author/portent-staff>)

**20 Comments**  
SHARES



MATTHEW MONTGOMERY ([HTTP://WWW.ONETAKEMEDIA.NET](http://www.onetakemedia.net))

MAY 15 2013, 09:54:48

This is an excellent post, there were somethings in here I knew instinctively and some I had never considered. "Being picky and realistic about what you want to rank for is the best start." I keep trying to pound this into my client's heads and I hope that one day it will stick! :-)



MARIANNE SWEENY

MAY 16 2013, 09:53:43

Hello Matthew

Yes picky is the way to go when looking at keywords to target. There is no value in ranking high in the results if the visitor does not stay on the site, learn about what you have to offer and ultimately convert. The search engines own the rankings and they do not play fair with the many mysterious adjustments, tuning and new additions that change the level of the playing field. Instead of a moment in time rank, let's focus on conversions and profitability for search success. Did organic traffic go up? Did the amount of \$\$ you made from organic traffic go up? In the end, this is what the clients want. We just have to convince them that it position plus site experience that delivers.



JOLY ([HTTP://WWW.RIGHTEOUS-MIND.COM/WHAT-WE-DO/SMO-SERVICES.HTM](http://www.righteous-mind.com/what-we-do/smo-services.htm))

MAY 15 2013, 10:18:16

Marianne According to me pagerank just a single concept to show any particular rank to any website. pagerank will not a big role the main factors which search engine like are website content, site navigation and best related images . you article is really helpful for freshers as well as for experts



MARIANNE SWEENY

MAY 16 2013, 09:38:25

Hello Joly

PageRank is an assigned value that is intended to represent the quality of a Web page. It has no relationship whatsoever to the search query. In the beginning, it was calculated on a per-page basis and updated monthly with the now fondly remembered Google Dance. Now there are other factors that

SHARES

contribute to ranking, factors that are harder to “game” by website owners such as Panda-related content quality calculations, user behavior metrics and more.

Good luck with your site.



CAMI ([HTTP://WWW.WEBCAMI.COM](http://www.webcami.com))

MAY 15 2013, 18:46:20

Great read, Marianne! Can't wait to share on my Facebook page!



MARIANNE SWEENY

MAY 16 2013, 09:54:38

Thanks Cami. In the end we'll all be able to do less running around after the search engines and more for our customers with a great user experience based on strong content.



DAN ([HTTP://WWW.BLINDSPOLESANDTRACKS.CO.UK](http://www.blindspolesandtracks.co.uk))

MAY 16 2013, 00:49:12

Thanks Marianne for a well written recap on this subject. I was with you right until the very last two sentences. I am trounced by my competitors in the SERPS for one discernible reason – they have tens of thousand of crappy paid-for links and I don't. It's extremely frustrating. All the good people in SEO (such as Portent) warn me not to go down that route (so I don't). But it's tempting, and the accepted 'truth' about poor quality links just doesn't seem to be reflected by reality.



MARIANNE SWEENY

MAY 16 2013, 09:32:05

Hello Dan

“Down these mean streets a man must go who is not himself mean, who is neither tarnished, nor afraid...” [Raymond Chandler]

It's hard to do the right thing when others are zooming ahead doing the wrong thing and are they really?

SHAME They may be at the top of the search results but do the folks who click through to their page do



anything, buy what they have, convert? That is and should be the goal, not a specific position in search results. Poor quality links may put you at the top of the results for a bit. However, the folks that click on a site that does not meet their expectations of relevance won't be there for long. Trust me on this. Machines are not that smart but they are relentless and good practices will win in the end.



JANET ([HTTP://CYBERTURFSTRATEGIC.COM](http://cyberturfstrategic.com))

MAY 17 2013, 17:12:30

Didn't Matt Cutts issue a video this week about projects that Google management approved – including clamping down on paid links? Hang in there, because it sounds as though help is on the way.



MARIANNE SWEENEY

MAY 20 2013, 12:00:13

Hello Janet, Matt's video was a “bugga bugga” about the latest Penguin update, or so it seemed to me. Penguin seems to be iterations of what is oftentimes referred to as Adversarial IR or spam detection. Google has been very proactive on the spam link detection case since the dual media hit in 2011 of the JCPenney.com and Overstock.co cases.



STEFAN DE BRUIJN ([HTTP://WWW.NUBILOSOFTE.COM](http://www.nubilosoftware.com))

MAY 16 2013, 03:48:08

Hi Marianne,

Great post. Allow me to contribute a few additional things we know Google to use. Most of these are based on IR research; publications can be found on conferences like DIR, ECIR, SIGIR and CLEF.

1. We know Google to have multiple search corpora that it mixes to produce the final search results.
2. We know that Google uses past search queries to do auto completion. It's very likely this is also based on a region (country / language), and the past few days (otherwise trending topics would disappear).
3. We know Google to use 'Learning to Rank', which basically means that the relevance algorithms are not static but change over time. While the details about how this works exactly are quite complex, in a nutshell the effect is that you get a 'personal profile' with some feature vectors that are used to balance the different relevance algorithms. The links you click on update the feature vectors. It is known that these profiles are personal (individual) – however, it's unknown if a non-personal profile is also used.
4. Another 'learning to rank' applies to picking the right algorithm for merging search results from different collections — but is most likely not individual. The concept is that if more people are interested in news, the

news corpus is more likely to produce search results. (and the same for images, blogs, etc)

5. Some features included for computing relevant search results are probably based on a derivative of Tf.IDF and language models (such as BM25 and DFR). Afaik all search engines do this.

6. Google uses NLP to process text. Language recognition is obvious, but also entity tagging and normalization are there. To illustrate, try Googling for brands that are a stopword in your language (in dutch 'DE' is the most frequent stopword and a coffee brand :-)) — and see how it 'magically' produces amazing search results.

7. Google probably uses word distance to as a feature in relevance ranking. Because the whole engine runs in-memory, this is an option — search engines that use disk to store the data usually become too slow and use a simpler measure of relevance.

I think I can go on for quite a while here; there are quite a few academic papers around that give insights (both on the scoring and the data structures)... and if you've build a few search engines for yourself (like I have) there's much you can derive from between the lines.

-Stefan.



MARIANNE SWEENY

MAY 16 2013, 09:24:01

Hello Stefan and great to meet another IR fan.

I agree wholeheartedly that Google's index is vast. Over the years, they have migrated away from MapReduce as a method to populate BigTable, the multidimensional index, and towards the incremental indexing that we find with Caffeine.

Spot-on about Google's use of user profiles, either deep if a Google account holder, or geo-centric if logged in anonymously. Google uses both profiles to tune results based on user behavior. However, this is done on a limited basis and usually for more common queries. The Web is so vast that it would be impossible for Google to process personalization refinements on global basis. My speculation is that they take a representational subset of the Web, process that and then use a Bayesian predictive model to apply to the rest of us. As Bill Slawski from SEO By The Sea (<http://www.seobythesea.com>) said in a post, "We're all Google's lab rats."

Here, I depart slightly in agreement. Google does not change its algorithms so much as add functionality with additional algorithms. My understanding is that an algorithm has a specific role and outcome.

I disagree with you on the "learning to rank" from different collections. The Web search engines treat all documents the same. They may pull from "the Deep Web" but only the unstructured content they find in the Shallow Web. The federation that you reference is more likely found in enterprise (e.g. within the firewall) search systems that have to search across many applications.

Yes to TF-IDF. I just added that to the post. It is the foundation ranking algorithm for most search engines. The language models you cite are two of many.

If Google used word distance it was early on. Now the search engines seem to be using a form of phrase indexing to present results for common term phrases.

SHARES

The struggle with this post was deciding between putting in too much or too little. For the takeaway message I wanted to leave behind, I hope I erred on just right.

I sometimes teach Introduction to Information Retrieval at the University of Washington Information School. Perhaps I'll see you in class Spring quarter 2014?



BRIAN ALLAN ([HTTP://WWW.NOVELSINORDER.COM](http://www.novelsinorder.com))

MAY 16 2013, 04:06:04

Hi Marianne,

Great post, very informative for me as a novice.

I'm a little unclear on the difference between Hilltop and Topic Sensitive algorithms – is the difference just the degree of refinement in terms of topic area, or does it mean a greater level of scrutiny of the onpage context of the link instead of just the broad subject?

Also, in terms of shared ownership links – I'm doing a little of this in terms of “my other sites”, etc. but it seems I should be treating this with caution – to what extent will Penguin punish this?

Brian



MARIANNE SWEENY

MAY 16 2013, 08:59:27

Hello Brian and many thanks for the kind words. Some see Topic Sensitive (and its cousin Hypertext Induced Topic Selection or HITS) as semantic cousins of Hilltop. The Hilltop algorithm introduced the capability of using links and contextual mapping between linked pages to determine whether or not a page is an authority. The search engine takes a subset of documents and, based on PageRank determines that certain documents are “experts” on certain subjects. If enough “experts” point to a single source, that source must be an Authority on the subject. Topic Sensitive added the nuance of subject matter so that documents did not just become authorities on all matters but matters of a certain subject, e.g. NASA likely ranks very high on space-related topics due to its authoritative nature on that subject. As for fashion, likely you will not find a NASA url in the top 1000 SERP.

Ian is the subject of shared ownership links. If the domains that share the links also share context with regard to subject matter and if the shared links do not trip the too-many-to-be-believed threshold, this should not be a problem. Google is looking for blatant examples of abuse, thousands of links over a short period of time, too many links from too few domains owned by the same entity or links from sites like <http://www.bestseodirectory.com> (<http://www.bestseodirectory.com>) when your site is about restoring water sheds.

SHARES



GEORGE ANDERSON ([HTTP://WWW.HIGHERLEAP.CO.UK](http://www.higherleap.co.uk))

MAY 20 2013, 04:02:47

Hi Marianne,

Thanks for posting this good advice. I wonder how many SEO practitioners completely ignore the PR scoring. I know that a lot of beginners put way too much emphasis on it.

Thanks again

George

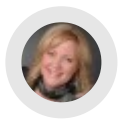


MARIANNE SWEENY

MAY 20 2013, 11:56:26

Hello George,

It is hard to put an exact number on it and, to be fair, the SEO industry has been slow to come around to the development of more meaningful benchmarks than PR and rank in search results. These are “easy measures” although mean little in the long run.



SANDY GERBER ([HTTP://NEXTFORYOURBUSINESS.COM](http://nextforyourbusiness.com))

MAY 29 2013, 21:45:36

Thanks for this! This was a very detailed explanation! I am curious about one thing (actually many things) but one in particular. I notice that keyword search tools such as Market Samurai and others will give you an indicator of how many .edu or .gov backlinks are coming in to a page. I guess this has to do with the credibility of a third party source linking to you, and that source being credible. However, in Canada and other countries, we don't use .edu or .gov – do you know if Google's algorithm is deployed in the same way in countries outside the U.S.? For example some countries don't have the Google Places rankings – is this mostly for political reasons or due to resources within Google trying to test the best methods in one area first before deploying the strategy in the entire world?



MARIANNE SWEENY

SHARES

JUN 6 2013, 10:04:15

Hello Sandy,

Curiosity is a very good thing! In the early days of PageRank, the search engines favored .GOV and .EDU domains because they were seen as “less commercial” and therefore “more credible.” In the world of unintended consequences, the search engines did not factor in cultural differences were all educational or government agencies might use .com domains. I am confident that their response would be that they fixed this issue with the enhanced semantic capabilities that search engines now employ along with other factors such as Authority designation through links from expert documents on the same topic, user preference as indicated by those signals and social signals.

You bring up very good points about the political and cultural bias that influences US search engine performance. Of course this is a concern for us all as the US search engine dominate the landscape in terms of market share and coverage. Please stay tuned for a deep dive blog post on this issue.



JOYCE GRACE ([HTTP://WWW.JOYCEGRACE.CA](http://www.joycegrace.ca))

MAY 29 2013, 21:48:23

Do you know more about the “learning machine” google set up (I heard) that tries to think more like a human? How would that play a factor in the algorithm? For example on the design and usability side of things, the information might be useful to some humans, but not to others. So if a user clicks and bounces off your site, then does that penalize your ranking due to the idea that must not have been a good user experience? I would think UI has a lot more to it than that and can be relative and based on individual context – so wouldn’t that make the Google algorithm “unfair” to some legitimate sites?



MARIANNE SWEENEY

JUN 6 2013, 08:25:03

Hello Joyce

I am unaware of Google’s “learning machine.” Frankly, I see the entire Google search engine entity as one giant learning machine. :) Google has its Knowledge Graph that is an attempt to associate conceptual data around a single item, e.g. searches on Abraham Lincoln pull up the knowledge graph that would contain links to the Gettysburg Address and other distantly related concepts to the main concept of our 16th president. Your point about information is well taken. I am now reading a paper by Jens-Erik Mai at the University of Toronto that addresses the concept of what is information and how do we determine the quality. For search engines, information is the end result. For humans, information is a component that they internally transform to make sense. Presently, the search engines are using a variety of factors to determine relevance and this is a fluid judgement that changes over time. A user’s bounce from a page is related to the query terms, the subject of the page, how other users have

SHARES

engaged with the page (e.g. if it is contact us and they are just noting the contact information), the quality of pages linking to the page, page layout, position in site structure and other variables that the search engine may be trying out.

Individual context is what we humans use to make sense of information and determine its relevance. Quantitative data (clicks, bounces, links) and predictive modeling are what the search engines use to initially determine relevance with user behavior to fine tune. Search engine algorithms are unfair by default. There is only so many resources for crawling indexing and storing information. So, the search engine “decide” who get how much. Search engines also discriminate in their indexing of international sites. Finally, and the press has been all over this, the search engines manually adjust results for reasons known and unknown. Unfair, yes. Something that we can do something about, not really. We CAN understand how this works and incorporate the knowledge into what we can control, the Web pages that the search engines rely on to build their indexes and serve our customers relevant search results.

Thanks for the great questions.

---

Comments are closed.

---

## Portent, Inc.


+1 206 575 3740 (tel:+12065753740)

info@portent.com (mailto:info@portent.com)

 (<https://twitter.com/portent>)

 (<https://www.facebook.com/portent.marketing>)

 (<http://www.linkedin.com/company/portent-inc>)

 (<https://plus.google.com/113999984820418066734/posts>)

 (<http://feeds.feedburner.com/conversationmarketing/mrji>)

307 3rd Ave. South Suite 400

Seattle, WA 98104-2687

 Directions (<http://www.portent.com/contact#map>)

**Careers** (<https://www.portent.com/advantage/careers>)   **Training** (<https://www.portent.com/training>)

© 1995-2018, Portent, Inc. All Rights Reserved. *Privacy Policy* (<http://www.portent.com/privacy>)

SHARES