



**On-Page**

# **How Google Search REALLY Works**

Complete Research

**By Eric Lancheres**

SEO Researcher and Founder of On-Page.ai

# 1. The Massive Google Leak

## Exposes Ranking Secrets

The biggest leak in Google's history just occurred, exposing 14,014 secret internal API references used within its search engine.

While many people underestimate the impact of this latest Google leak, **I believe it allows us to confirm or debunk theories about Google's inner workings**, revealing new insights that can revolutionize how we approach SEO.

**It allows us to dodge many of the spam traps that Google has set in place and shows us the exact path forward to rank on Google.** I've spent hundreds of hours analyzing the leak and cross-referencing it with my 12 years of SEO notes, including insights from ranking, examining search patents, and conducting SEO tests.

I'm excited to share all my discoveries with you.

And I won't stop at just sharing discoveries—I'll also explain how you can implement everything to achieve better SEO results that can significantly impact your bottom line.

1. If you've been affected by a recent Google update or are looking to rank your site in record time, this is for you.
2. To all search engine professionals, SEO agencies and sole-entrepreneurs owners looking to rank your clients' website rankings, this is for you.



In March 2024, both Erfan Azimi & Dan Petrovic discovered an exposed Google repository containing many of the API references used internally by Google.

Erfan shared the leak with Rand Fishkin which then relayed the information with Mike King while Dan (who had already discovered the leak independently) was in the process of disclosing the leak to Google. After a bit of persisting, Google finally acknowledge there was a leak and resolved while still leaving the data indexed. **To Google's credit, they have left it indexed as they are holding themselves to the same standard as everyone else.**

[https://hexdocs.pm/google\\_api\\_content\\_warehouse/0.4.0/api-reference.html](https://hexdocs.pm/google_api_content_warehouse/0.4.0/api-reference.html)

## GoogleApi.ContentWarehouse.V1.Model.AnchorsAnchor

### Attributes

- `creationDate` (type: `integer()`, default: `nil`) - used for history - the first and last time we have seen this anchor. `creation_date` also used for Freshdocs Twitter indexing, a retweet is an anchor of the original tweet. This field records the time when a retweet is created.
- `origText` (type: `String.t`, default: `nil`) - Original text, including capitalization and punctuation. Runs of whitespace are collapsed into a single space.
- `context2` (type: `integer()`, default: `nil`) - This is a hash of terms near the anchor. (This is a second-generation hash replacing the value stored in the 'context' field.)
- `fontSize` (type: `integer()`, default: `nil`) -
- `experimental` (type: `boolean()`, default: `nil`) - If true, the anchor is for experimental purposes and should not be used in serving.
- `fragment` (type: `String.t`, default: `nil`) - The URL fragment for this anchor (the foo in `http://www.google.com#foo`)
- `sourceType` (type: `integer()`, default: `nil`) - is to record the quality of the anchor's source page and is correlated with but not identical to the index tier of the source page. In the docjoins built by the indexing pipeline (Alexandria), - Anchors marked `TYPE_HIGH_QUALITY` are from base documents. - Anchors marked `TYPE_MEDIUM_QUALITY` are from documents of medium quality (roughly but not exactly supplemental tier documents). - Anchors marked `TYPE_LOW_QUALITY` are from documents of low quality (roughly but not exactly blackhole documents). Note that the `source_type` can also be used as an importance indicator of an anchor (a lower `source_type` value indicates a more important anchor), so it is important to enforce that `TYPE_HIGH_QUALITY < TYPE_MEDIUM_QUALITY < TYPE_LOW_QUALITY`. To add a new source type in future, please maintain the proper relationship among the types as well. `TYPE_FRESHDOCS`, only available in freshdocs indexing, is a special case and is considered the same type as `TYPE_HIGH_QUALITY` for the purpose of anchor importance in duplicate anchor removal.
- `pagerankWeight` (type: `number()`, default: `nil`) - Weight to be stored in linkmaps for pageranker
- `isLocal` (type: `boolean()`, default: `nil`) - The bit ~roughly~ indicates whether an anchor's source and target pages are on the same domain. Note: this plays no role in determining whether an anchor

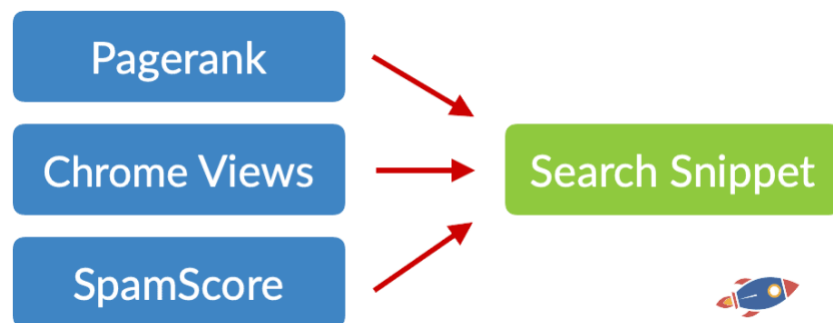
*Example of an API reference*

## 2. The Building Blocks Of Search (Google's Internal API Documentation)

The data warehouse leak contains documentation on the APIs that Google uses to build its algorithms.

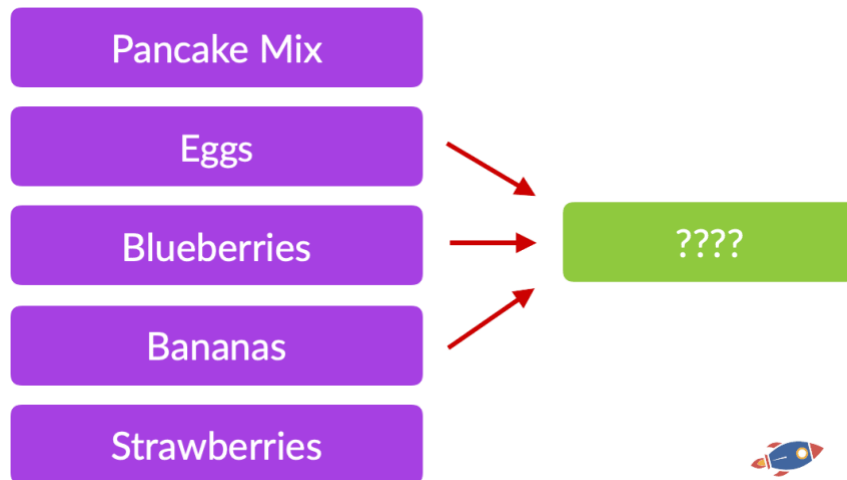
These are the building blocks of the search engine.

For example, if one morning a search engineer wakes up and decides he wants to create a new algorithm that only show search snippets from websites with a **PR5+** that have **more than 20,000 Chrome views** and have a **Spamscore of 10 or lower**. Then he can retrieve the data using these APIs.



*Hypothetical example of how an engineer can use API information create new algorithms. Please note that this is NOT how the search snippet actually works. (I'll explain that later)*

This is similar to walking into your favorite restaurant's kitchen while the chef is away. **Imagine discovering all his fresh ingredients laying on the counter...**



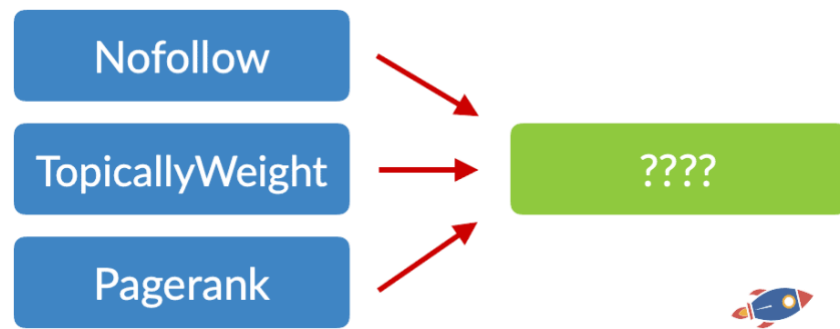
After a few minutes in the kitchen... you can ask:

1. *"What is he making here?"*
2. *"What kind of toppings is he using when they make pancakes?"*

If we know the ingredients, and we know that he's making pancakes... **we can easily deduce that the preferred toppings are going to be: blueberries, bananas and strawberries.**

This is very similar to ranking factors...

So if we see something like this:



Even though we don't know what the final algorithm is (represented in green), we know that:

1. **NoFollow** is exclusively related to links
2. We know that **Pagerank** is **ALSO** related to links
3. Therefore, the green block must be covering a link related algorithm.
4. Thus, **we can reasonably assume that **topicality** is also ranking factor when it comes to links.** (A measure of how related to the topic the link is.)

**Plus, we have the added benefit that Google employees have often added elaborate descriptions explaining how these APIs fit into the algorithm so sometimes we don't even have to make assumptions, we can just read it!**

## Freshness & Accuracy Of The Leak

### GoogleApi.ContentWarehouse.V1.Model.RepositoryWebrefWebrefMustangAttachment

*THIS ATTACHMENT IS DEPRECATED, SEE [go/udr/migrate-wma](#). . We still allow legacy use case to exist (no forced migration), but we will not accept any new usage of WMA, incl. from existing clients. UDR has the same features and can be used similarly: - To consume the topical entities (+properties, incl. hitcat, browsy, ...) [go/udr/migrate-wma](#) provides a migration with minimal changes. - To consume IQL, please consult [go/udr/superroot#access](#) and [go/pianno](#) team. The top-level proto used to store WebRef entities and IQL expressions in Mustang/TG. The proto uses packed repeated fields and variable-length integers in order to be as compact as possible. See [http://b/5802389](#) and [b/7473898](#) for details on other approaches that were considered and space/readability/extensibility trade-offs made. Note: It is not recommended to read this proto directly. Clients of the attachment should use the decoder instead: [repository/webref/tools/kc/indexing/webref-attachment-decoder.h](#) Next available tag: 25*

*If you choose to take action on any of the data presented here, do so at your own risk. You are responsible for any changes you make to your website.*

When it comes to the freshness and accuracy of these leaks, even though I was unable to find exact dates, **they seem to have been updated in 2024, giving us a very recent snapshot of Google's inner workings.**

Furthermore, it's apparent that many of the APIs documented are still actively in use, backed by **carefully maintained reference documentation.** This careful upkeep highlights their ongoing relevance.

APIs no longer in use are clearly marked for deprecation. This is evident from numerous comments that detail items currently being phased out. Finally, **Google has been building its collection of APIs for over a decade so even though there are always updates, the core algorithm remains stable, ensuring that much of the foundational knowledge and techniques remain applicable over time.**

I will NOT be reproducing the leaks here; however, I will include partial mentions below for context and educational purposes.

## 3. Building Better Links

### Insights From Leak

It should come as no surprise that links still play a major role in rankings, contrary to what Google or some SEOs might claim. They are one of the major ranking factors that help Google understand context, authority and importance.

However...

With regards to links, I was intrigued to see such a **large emphasis on link anchor text**.

Here's what I discovered:

### Anchors

`context2` - This is a hash of terms near the anchor

*Please note that these are partial references to the leak as I do not want to republish the entire document.*

According to the documentation, **the full description of "context2" confirms that the words before and after your anchor text affect the anchor text**. While this has long been suspected within the SEO community, it's nice to finally see it in the flesh.



## How You Can Use This To Build Better Links

For example, if we assumed that the 5 words before and after your anchor text can influence the anchor text

*For the best fishing poles [click here](#)*

Google knows that "click here" is related to fishing poles because of the context. So if you're building links and have little control over the anchor text itself (due to site moderation), then at least try to include related entities NEAR the anchor text.

**However, if you do have complete control over the anchor text, then an ideal link would include both a relevant anchor text AND relevant surrounding content.**

For example:

*catch fish with these [fishing poles](#)*

In this example, both the anchor text is relevant AND the surrounding words are relevant.

**sourceType** - is to record the quality of the anchor's source page

The next element, sourceType, explains how **high quality anchor texts come from "Base documents"**. What this likely means is that links that come from content ranking in the same keyword bucket, aka highly related content), will carry a heavier weight in terms of anchor text.

**Medium quality anchors, according to Google, comes from not very related content.**

**Finally, low quality anchors come from, \*drum roll\*, low quality content.** While Google doesn't tell us what they consider a low quality document, we can assume that it is content that isn't related to the topic and likely has a host of other metrics that classify it as low quality. (More on this later)

However, the big takeaway is that not all anchors are worth the same!

**Ultimately, The BEST links you can get are going to be from OTHER pages currently ranking for the same term.**

*(Because pages that rank for the same term will be classified in the same bucket)*

## Links

**isLocal** - indicates whether an anchor's source and target pages are on the same domain

**expired** - true iff exp domain

**deletionDate**

**locality** - For ranking purposes, the quality of an anchor is measured by its "locality" and "bucket"

**parallelLinks** - The number of additional links from the same source page to the same target domain

One big "ah ha" moment for me was realizing that **internal and external links are VERY similar, only a few parameters separate the two**. While I knew that both were important, *I previously (incorrectly) assumed there might have been a drastically different algorithm that handled both...*

However, according to the documentation, **it appears as if internal links and external links are more closely related than originally anticipated**. Further reinforcing the importance of good internal linking practices.

Within the links section, we also see that **Google has a specific flag for when links come from a domain flagged as an expired domain**. (*This might be something they use for link penalties. If you accumulate too many links from expired domains, it might land you in some trouble*).

In addition, **they have a record of deleted links**. This explains why we can often see the "ghost link" phenomenon in which webpages will continue to rank **EVEN** after links are removed. It's entirely possible that Google continues to count some of the links within their algorithm for some time even after the deletionDate.

*(I speculate this is used to improve search engine result stability as link can sometimes migrate from the homepage, to inner pages and then move deeper into a site as content is shuffled around.)*

Locality is interesting because they **LITERALLY say that they are looking for links within the same BUCKET**. This is as close as it gets to saying that those links are going to be worth considerably more.

And finally, parallel links indicates that **additional links from the same domain might not count for as much**.

## How You Can Use This To Build Better Links

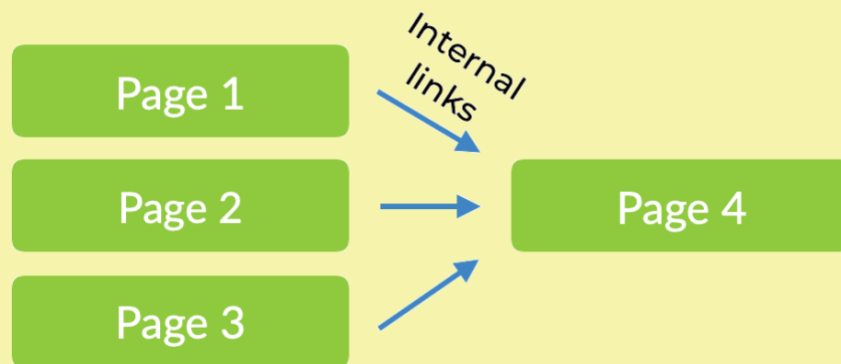
We already know that internal links are important...

**In fact, internal links can count for almost as much as external links as they are treated in a very similar fashion.**

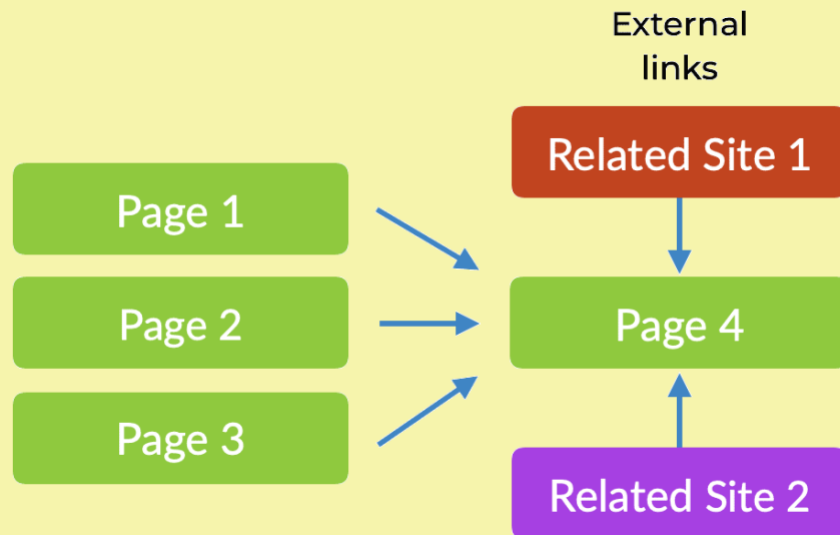
And... they are EASY to create because you control the domain.

However this means when you're creating internal links, you STILL want to:

- 1- Include internal links on relevant pages.
- 2- Vary your internal link anchor text



In terms of saving time, this leak confirms that they are counting parallel links (and therefore, we can assume that there are diminishing returns for multiple links from the same domain) so the **most effective link building will focus on getting links from multiple different sites.**



That said, I want to make clear that multiple links from a single domain still helps...

However, when link building, your time will be better spent getting links from a different domains.

**To locate a list of the domains within your keyword "bucket":**

1. Search for your keyword on Google
2. The top 100 results for your keyword are URLs within the "bucket"

**Obtaining a link from a page that is already ranking for your target term will be one of the most relevant links you can acquire.**

*(For highly competitive terms, one of my personal favorite techniques is to scrape the top 100 results for my keyword. I then hire a worker to manually extract the contact information for each URL and send them a personalized message seeking a link, sparing no expense. While slightly tedious (and sometimes expensive), I will typically acquire a handful of links this way and this can go a long way to securing the #1 spot for highly competitive keywords.)*

## Link Evaluation

Here's more information on the topic of links providing specific insights into how Google evaluates links.

bucket

setiPagerankWeight - TEMPORARY

Once again, we have a mention of the keyword bucket. This means that Google notices the set of pages / entities is the link associated with. **We can assume that links within relevant content will be worth more.**

In addition, we have "setiPagerankWeight" which means that Google likely has a temporary value for Pagerank when Pagerank isn't yet calculated. *(In order for the algorithm to work, they probably need a default PageRank value otherwise the algorithm might break. This is why brand new pages with no apparent links can still provide some temporary benefit as all pages have a default, minimum value.)*

isNofollow - Whether this is a nofollow link

topicalityWeight - The topicality\_weight for each link with this target URL

We start off with the "isNoFollow" tag which clearly indicates we are in the link section of the API. And to my surprise, the description reveals a small and subtle secret:

**If a page has multiple links and just ONE of them is 'follow' then ALL of them will be considered follow.**

*(Note: There is no 'do-follow' tag. A link to be considered as follow by the absence of a no-follow tag.)*

The next available API is TopicalWeight which, once again, **indicates they are measuring link relevance!** This is very likely a numerical value measuring how relevant a link is to the target URL, likely based on the content.

## How You Can Use This To Build Better Links



*We already knew that links coming from related content were worth more...*

However it's important to note that there's a **WEIGHT** associated to it. This means that there's a degree of 'relatedness' that is calculated and **the MORE it's related, the better it is.**

Looking at the example above, if we have a link coming from a general "pet" article... it's good. However a link originating from another document on "dogs" will carry even more weight.

**Therefore, in order to maximize our link building, I would try to get links from highly related documents.**

*(My personal favorite technique when it comes to maximizing relevance is to determine the NLP category of my content using Google's own list of NLP categories.)*



## Step 1) I click on "Determine Google Category"

The screenshot shows the On-Page SEO tool interface. The main content area displays the article title "How to Revive Faded Hardwood Floors in 5 Simple Steps" and its text. The left sidebar shows a "Quick Score" of 100, a "Word Count" of 2627/2306, and "Metric (Yours/Avg)" for H2, H3, and H4. The "NLP Category" is identified as "/Home & Garden/Home Improvement/Flooring". The right sidebar shows "Recommended Words" and "Highly Related Words". A red arrow points to the "Determine Google Category" button in the top right corner.

## Step 2) The text is processed with Google's NLP and returns: /Home & Garden/Home Improvement/Flooring

This uses Google's categorization engine and it provides me with the best idea of what Google really thinks about my text. I built this feature so I can maximize the relevance when link building. The idea behind it is that if I know that both the source and the target documents are in the same category, then the link will be more relevant. )

## Link Attributes

Another section of the leak describes the various attributes a link can have. This provides us with additional insights and confirms some existing theories.

`additionalInfo` - Additional information related to the source, such as news hub info.

From the description of `additionalInfo`, **we can see that links from news hubs are noted.** In other words: "Is a topic trending in the news" and receiving news links. I personally believe that Google treats websites that are trending in the news slightly differently than sites that have no news links.

### My Secret "Trending Website" Trick

*I personally believe that if your site is determined to be in the news and you receive a large influx of links, Google will understand that you're going viral and that you're trending. In contrast, if you receive a large quantity of links and you aren't in the news, it might be seen as spam. While not confirmed, this idea originates from reading Google patents.*

*One of my personal link building tricks involves publishing a press release about my website shortly before going on a link building campaign. I'll write a general press release so the site is trending in the news.*

*After approximately 1 week, I'll then proceed to building links to different inner pages.*

`cluster` - anchor++ cluster id

`homePageInfo` - Information about if the source page is a home page

Then we have another indication of clusters, which serves to identify the cluster that the link belongs to.

Finally, Google pays special attention to homepage links because historically, those were highly abused. In the description they elaborate that **if they find a homepage link, they will verify to see if they trust the homepage.**

`pageTags` - Page tags are used in anchors to identify properties of the linking page

`pagerank` - uint16 scale

`pagerankNs` - unit16 scale

`PagerankNS` - Pagerank-NearestSeeds is a pagerank score for the doc, calculated using NearestSeeds method

`spamrank` - uint16 scale

`spamscore1` - deprecated

`spamscore2` - 0-127 scale

Of course, we have indications that Google uses pagetags to identify page and relate them to the link. Nothing new here.

To my surprise, we also see that **Google STILL uses Pagerank within it's algorithm.** While this was widely acknowledged earlier on in Google's history, they have since mentioned that they no longer use the original Pagerank.

I suppose that's *technically true* as they appear to use a new version of it called PagerankNS.

**The NS part of the PageRank stands for "Nearest Seed" which is incredibly important.** This will significantly impact how we evaluate sites.

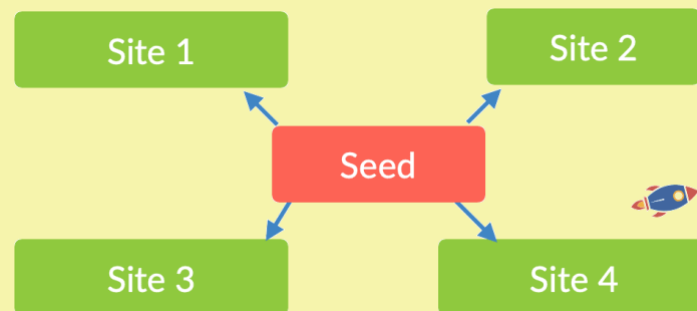
Finally, as expected, links have a "Spamscore" associated with it which is likely derived by looking at the link text and surrounding text.

## How You Can Use This To Build Better Links

Knowing that Google has updated the Pagerank algorithm to PagerankNS is absolutely critical for our link building.

**First, it's notable because Google denied using seed sites in the past**

However now that we know they do... this can help us build more effective links.



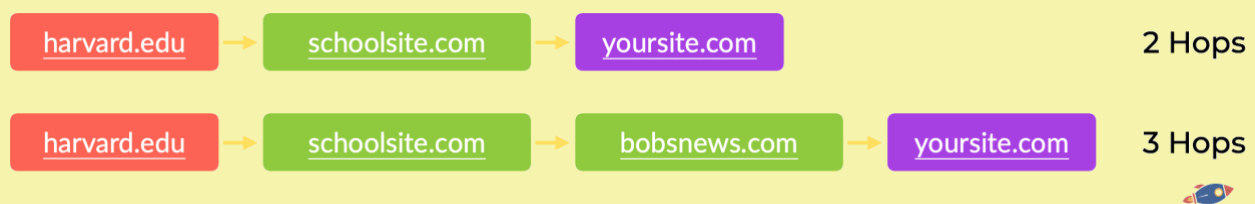
**Seed based Pagerank measures for how far from the main seed the link is.**

Hypothetically, Google can have a pre-determined a list of 1000 trusted seed sites.

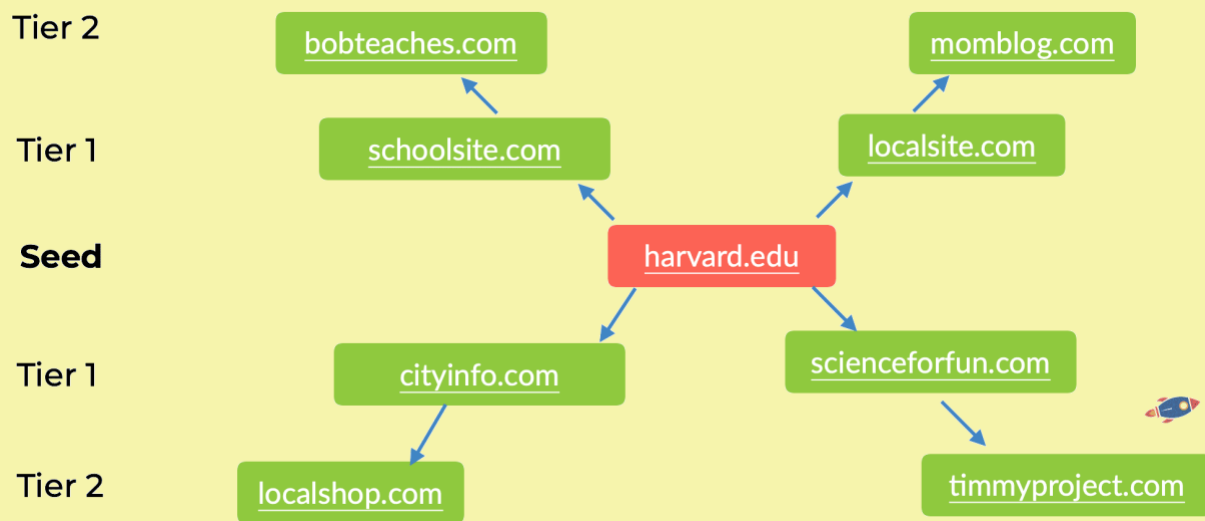
If you get a link from a seed site, you get the most power. If you get a link from a site that has a link from a seed site, then you are getting a tier 1 link.

**The further away from the central seeds, the less the link is worth.**

In a practical sense, if Google were to set Harvard as a seed site.



The closer your link is to the "seed" site, the better it is.



In the opposite scenario, if you were to get a link from "tier 45 website" (a site that is very far away from a seed site) ... it would be almost worthless. This is why getting links from very small websites can sometimes have little to no effect, *even if the link is coming from related content*.

So now the burning question...

**What are the seed sites used by Google?**

We don't know the exact list (and of course, they won't tell us).

**However, we do know that seed sites are typically highly moderated sites that exclusively link to trusted properties.**

Among them, you would expect to find some (but not all):

- .edu Trusted education institutions
- .gov Government websites
- .mil Military website

For example, it might be reasonable to assume that these *might* be seed sites:

harvard.edu	nih.gov	army.mil
mit.edu	cdc.gov	navy.mil
stanford.edu	privacyshield.gov	af.mil
cornell.edu	ca.gov	uscg.mil
berkeley.edu	dataprivacyframework.gov	osd.mil
academia.edu	irs.gov	dtic.mil
yale.edu	ftc.gov	militaryonesource.mil
columbia.edu	fda.gov	darpa.mil
umich.edu	epa.gov	marines.mil
upenn.edu	usda.gov	tricare.mil
washington.edu	nasa.gov	health.mil
psu.edu	hhs.gov	disa.mil
umn.edu	sec.gov	dla.mil
jhu.edu	hud.gov	nga.mil
si.edu	whitehouse.gov	dod.mil
princeton.edu	texas.gov	whs.mil
uchicago.edu	state.gov	dodlive.mil
wisc.edu	ny.gov	defenselink.mil
ucla.edu	noaa.gov	spaceforce.mil
cmu.edu	nps.gov	dfas.mil
nyu.edu	loc.gov	cyber.mil
utexas.edu	ed.gov	esgr.mil
usc.edu	ssa.gov	usmc.mil
purdue.edu	census.gov	dcsa.mil
northwestern.edu	bls.gov	arlingtoncemetery.mil
uci.edu	nist.gov	dodig.mil
unc.edu	va.gov	ng.mil
illinois.edu	cms.gov	jcs.mil
ufl.edu	sba.gov	nationalguard.mil
ucdavis.edu	copyright.gov	dcoe.mil
msu.edu	house.gov	centcom.mil
ucsd.edu	energy.gov	dsca.mil
brookings.edu	dol.gov	dia.mil
umd.edu	justice.gov	doded.mil
duke.edu	medlineplus.gov	dtra.mil
hbs.edu	wa.gov	socom.mil
osu.edu	usa.gov	pentagon.mil
tamu.edu	congress.gov	dpaa.mil
rutgers.edu	senate.gov	pacom.mil
asu.edu	dot.gov	dau.mil
arizona.edu	fcc.gov	sigar.mil
ncsu.edu	osha.gov	mda.mil
bu.edu	treasury.gov	dren.mil
georgetown.edu	archives.gov	dma.mil
colorado.edu	usgs.gov	norad.mil
virginia.edu	weather.gov	africom.mil
utah.edu	dhs.gov	dss.mil
tsinghua.edu.cn	fema.gov	southcom.mil
unl.edu	nyc.gov	stratcom.mil

*In my experience, I discovered that with a bit of outreach, it's possible to get links from local cities and from universities. For local cities, I sometimes offer specific discounts to residents and for universities, I'll try to tie into research and/or contact active graduate students that might have access to a part of the university site. I make sure to remain on the exact top level domain (I don't want a link from a sub-domain).*

*Finally, I previously used to use scholarships as a link building tactic HOWEVER Google addressed this technique directly (possibly by flagging the giant scholarship pages so they don't pass any power) so this is no longer a viable technique. I still believe in supporting students but beware this will no longer be as beneficial from a link building perspective.*

## Link: NSR

**nsr** - This NSR value has range [0,1000] and is the original value [0.0,1.0] multiplied by 1000 rounded to an integer.

This part is incredibly important and I believe has been overlooked by a large portion of the SEO community. This is the first mention of NSR and has a huge multiplier attached to it to amplify its effect within the algorithm.

While Google doesn't explicitly disclose what NSR stands for in the documentation, after analyzing hints from dozens of descriptions, brainstorming dozens of different possibilities and weighing the most likely answer, **we can conclude that NSR likely stands for Normalized Site Rank.**

### How You Can Use This To Build Better Links

Normalized Site Rank is incredibly important.

**It is very likely a comparison of your site's click / engagement performance versus other sites.**

For example,

*Rank #1 Wikipedia*

*Rank #1000 BigSite.com*

*Rank #300000 smallnichesite.com*

Normalized, from 0 to 1 in terms of how prominent a site is compared to others... and then they multiple this by 1000 to get a huge weight that influences rankings.

This ultimately means...

**When you're link building, you want to get links from the biggest players in the industry. This is because the sites that are ranking for everything in your industry will likely have the highest NSR.**



Get a guest post there, a news article about you, something... and that link will be worth considerably more than a bunch of smaller links from non-authoritative sites.

*I will share tricks on how you can potentially increase the NSR for your site within the site quality section.*

*When link building, I personally sign up as a contributor to all the major sites within my industry. It's a slow process, but I start by contributing a handful of articles without any commercial links. Once I've established a solid reputation on a site, I subtly link back to my own site. It can sometimes take 6 months to get a single link... however I believe it's because I'm willing to do the things that other web owners aren't willing to do that I can surpass them.*



## Link Relevance

Another link section includes an interesting API called "score".

**score** - Score in  $[0, \infty)$  that represents how relatively likely it is to see that entity cooccurring with the main entity (in the entity join)

With regards to the score, I believe they might be referring to the anchor text itself (as in, **how likely is the anchor text to appear inside the text that it is linking to**).

Essentially, this could prevent misalignment with regards to anchor text. So for example, if you were to create anchor text links for "*cell phone cases*" but you pointed them to a page about "*skin care*", then the score would be extremely low as it is very unlikely to appear naturally within the content.

Bottom line: **We want links from relevant content, with relevant anchor text, pointing back to our documents.**

## Link Location

**boostSourceBlocker** - Defined as a source-blocker, a result which can be a boost target but should itself not be boosted (e.g. roboted documents)

**inbodyTargetLink**  
**outlinksTargetLink**

One interesting API reference is the "boostSourceBlocker" which essentially means that **a page can potentially be a valuable source for links while not ranking itself.**

Historically, SEO professionals (myself included) might have been wary of receiving links from those pages that weren't ranking... however, according to this API reference, in certain cases, we might be missing out.

It would make sense that links from directories, category pages, rss feeds and historical records could potentially be valuable even if they don't show up in search.

With regards to the other two API references, **I believe that Google identifies the location of the link, making a distinction between links within the main body of a document (Main content) and links outside the main content (header, sidebar, footer, etc).**

If this is the case, then it might be reasonable to assume that that links from main content are worth more.

## Local Mentions

Because "mentions" might be play a role in SEO, I thought it might be appropriate to cover it here.

`annotationConfidence` - Confidence score for business mention annotations

`confidence` - Probability that this is the authority page of the business

The first API reference is: "AnnotationConfidence". I believe this might be the confidence in the mentions / annotations (citations/business name mentions) of the business. We've known for a while that **citations can play a role with local SEO** however **I haven't seen as much testing with just mentioning the business name.**

This might play a role and would be interesting to test.

In addition, I never knew that **Google checks the relevance of the website linked within a Google listing.** This makes sense to prevent spam as we wouldn't want a local daycare Google listing to link to a supplement website.

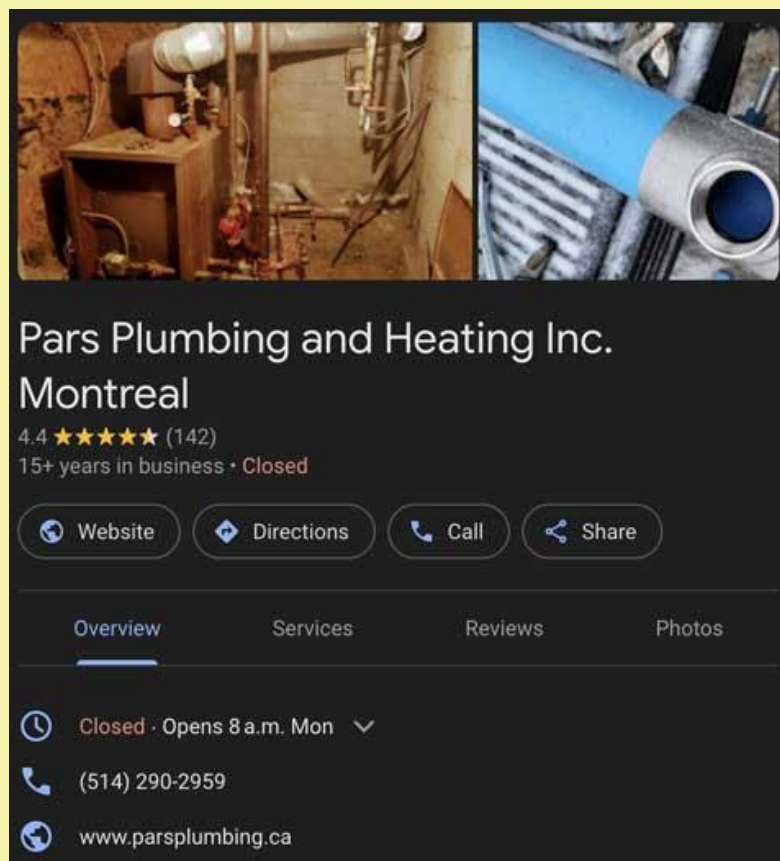
## Better Local Ranking

It is important to **fill out ALL** the possible options in your local Google listing in order to **maximize ranking**.

In addition, we want the **entities** used on the local Google listing to **MATCH** the entities **on the website**.

So for example, if your local business description includes: "*plumber*", then make sure it says "*plumber*" on the homepage of the site.

*(This goes without saying however for local rankings, I personally do everything I can to get links from the official city & from State/Province websites. These tend to have incredible power and getting just a single link from the official city site can sometimes make my listing skyrocket. Don't tell anyone!)*



## Anchors For Rank

I wasn't expecting this much documentation on anchor texts however it appears as if Google uses them extensively when evaluating pages and entire websites.

`SimplifiedAnchor` from the anchor data of the docjoins, by specifying the option `separate_onsite_anchors` to `SimplifiedAnchorsBuilder`, we can also separate the onsite anchors from the other (offdomain) anchors

`anchorText` - The anchor text. Note that the normalized text is not populated  
`count` - The number of times we see this anchor text  
`countFromOffdomain` - Count, score, normalized score, and volume of offdomain anchors  
`countFromOnsite` - Count, score, normalized score, and volume of onsite anchors  
`normalizedScore` - The normalized score, which is computed from the score and the total\_volume  
`normalizedScoreFromOffdomain`  
`normalizedScoreFromOnsite`  
`score` - The sum/aggregate of the anchor scores that have the same text  
`scoreFromFragment` - The sum/aggregate of the anchor scores that direct to a fragment and have the same text

The first notable discovery is that **Google combines external and internal text to determine anchor relevance.**

In addition, they create a score out of many different metrics:

- How many **different anchor texts** a page has.
- How **frequently** do we see the **same anchor text**,
- How many **external anchor texts** are there,
- How many **internal anchor texts**

And finally, it puts it all together to create a final anchor text score.

```
scoreFromOffdomain
scoreFromOffdomainFragment
scoreFromOnsite
scoreFromOnsiteFragment
scoreFromRedirect - The sum/aggregate of the anchor scores that direct to a different wiki title and have
the same text.
totalVolume - The total score volume used for normalization
totalVolumeFromOffdomain
totalVolumeFromOnsite
```

We also see that fragmented anchor texts (partial anchor text) still seems to count towards the main anchor text score... even if it's not an exact match.

## Better Link Building

This highlights the **importance of having MANY internal links** every time you want something to rank.

*(While it takes quite a lot of effort to acquire external links, you can easily create a handful of internal links within a few minutes.)*

**Each internal anchor text should be highly relevant AND should vary to avoid too high of a count.** (They'll still be counted by the fragmented part)

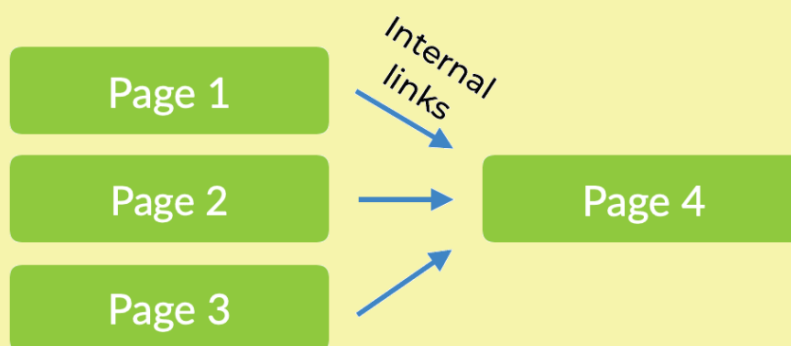
For example:

"fishing poles"

"great fishing poles"

"poles for fishing"

"fishing"



*(I personally believe it's crucial to have relevant internal links between all your content. To build topical clusters and enhance topical relevance, your articles need to be properly linked together. This strategy can potentially boost the rankings of your entire website as it gains topical authority. I will expand on this further in the topical authority section.)*

## Validating Anchor Text

This section highlights some of the more clever tricks that Google uses to rank a wide range pages online.

```
matchedScore - Difference in KL-divergence from spam and non-spam anchors
matchedScoreInfo - Detailed debug information about computation of trusted anchors match
phrasesScore - Count of anchors classified as spam using anchor text
site - Site name from anchor.source().site()
text - Tokenized text of all anchors from the site
```

All these scores, from `matchedScore`, `phrasesScore` seem to be a calculation of the total anchor texts pointing to a site in order to determine the site's relevance. This re-emphasises the importance of anchor texts when ranking.

What's really interesting is the solution that Google devised simultaneously avoid spam and while ranking queries for people that are searching for sensitive topics. (*I'll describe how it works down below.*)



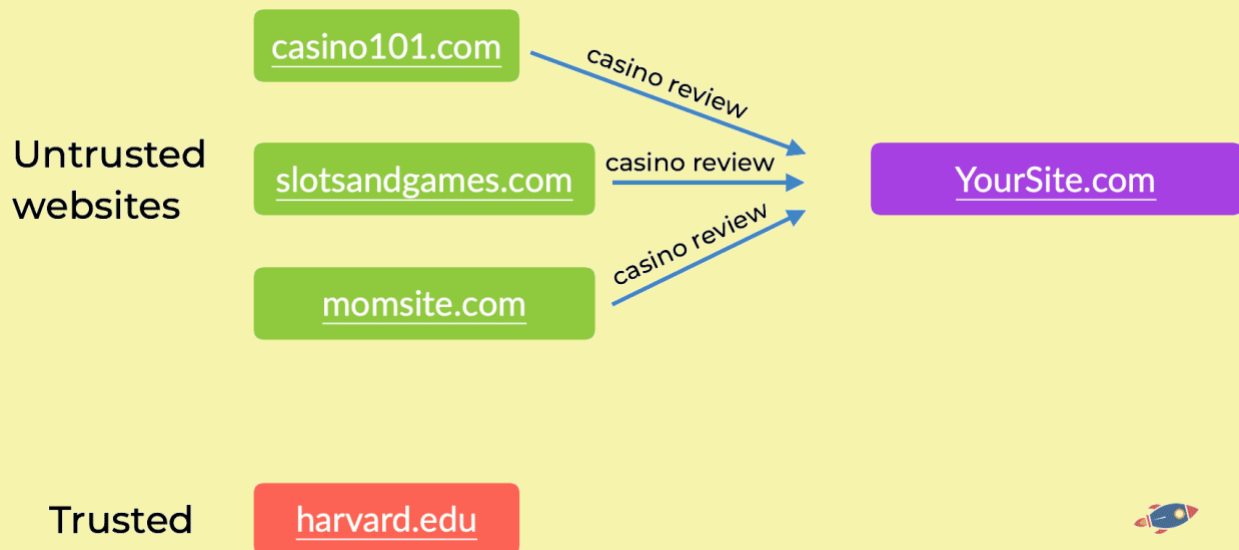
## Better Link Building

This is incredibly smart... if you're in a business such as casinos, CBD or any "high risk" industry.

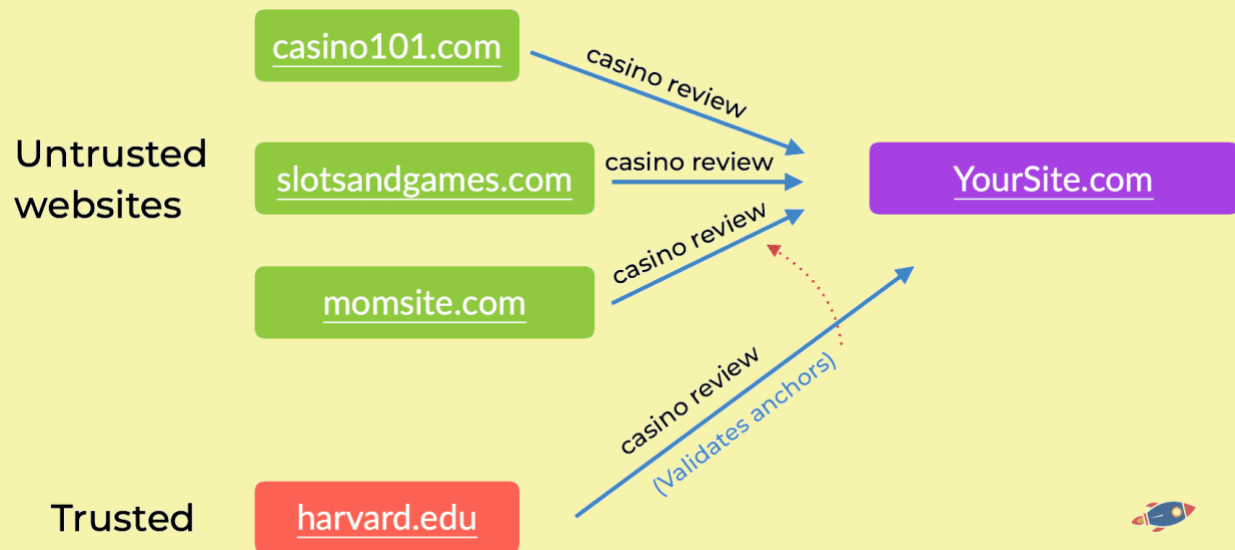
**Then Google will look at anchors from trusted sites in order to VALIDATE the links/anchors from less trusted sites.**

In practice, this means is that IF you get 10 links from randomsite.com pointing to your site with the anchor "casino slots", it will think it's spam.

However, if you then get a trusted site, harvard.edu to link to you with "casino slots", then Google will not only accept it... it will ALSO validate and pass the power from randomsite.com



In the example above, NONE of the links are trusted. This is because none of the websites linking with the shadier anchor text are considered trusted.



In this instance, ALL of the links are trusted. The link from the trusted site (in this case, Harvard), validates the link anchors from the other websites.

In practice, **if you're trying to rank for terms that might be fall into this "high risk" category, then it is critical that you mix in links from trusted websites in order to validate other links.**

*(I personally used this technique extensively in the past when I was building link for HIGHLY competitive terms in a sensitive industry. At one point I had links from 17 different sites that I owned pointing to a single place and in order to get it 'over the top', I also acquired 1 link from an industry site via a contribution I made. This one link seemed to have made all the 17 other links work better and the page shot up to #1 for my target keyword.)*

**trustedScore** - Fraction of pages with newsy anchors on the site, >0 for trusted sites

Finally, we also see they have a trustedScore that verifies if the site is in the news (news links!). Here's how it can help you do better SEO...

## Better Link Building

Google TRUSTS sites that link out frequently using "newsy" anchor texts.

This is HUGE for determining if a site is a good source for a link.

**YOU WANT LINKS FROM SITES THAT HAVE MANY "NEWSY" related outgoing links.**

For example, if a site is frequently linking out with anchors such as:

*"Website"*

*"Here"*

*"Click here"*

**Then it is likely a news website that can be trusted.**

Conversely, if it's a site that is frequently linking out with commercial terms such as:

*"best drone reviews"*

*"blue t-shirts"*

*"best CBD oil"*

**Then the trusted score for the entire website will be VERY low.** (And then you don't want a link from it)

The implications are HUGE... and I almost don't want to say it...

Knowing that TrustedScore is determined by the anchor texts profile of the website means that...

## Manipulating TrustedScore

*Disclaimer: I do not recommend doing this. Proceed at your own risk. You are responsible for any changes you make to your site.*

Link sellers, PBNs and other sites that want to increase their "TrustedScore" *might* be able to artificially link out with a TON of anchor texts with:

"website"

"click here"

"read here"

"site"

"here"

And that should, *in theory*, increase the TrustedScore of the site by altering the ratio of commercial to 'newsy' anchor text... which will make all the other links work better.

*Note: The more outgoing links there are, the more it dilutes Pagerank. Hypothetically, if I were to do such a thing, I would create a category with only 1 link to minimize Pagerank loss. Within that category, I would create all the anchors text links.*

## Bad Links

Due to the importance of links and anchor texts, SEOs have been artificially creating them since the dawn of SEO. Here's how Google deals with link abuse.

`penguinLastUpdate` - BEGIN: Penguin related fields. Timestamp when penguin scores were last updated. Measured in days since Jan. 1st 1995  
`anchorCount`  
`badbacklinksPenalized` - Whether this doc is penalized by BadBackLinks, in which case we should not use improvanchor score in mustang ascorer  
`penguinPenalty` - Page-level penguin penalty (0 = good, 1 = bad)  
`minHostHomePageLocalOutdegree` - Minimum local outdegree of all anchor sources that are host home pages as well as on the same host as the current target URL  
`droppedRedundantAnchorCount` - Sum of anchors\_dropped in the repeated group RedundantAnchorInfo, but can go higher if the latter reaches the cap of kMaxRecordsToKeep

Nearly every reference in here is important.

First, it confirms that **Penguin updates periodically** which means that if you are hit by a Penguin penalty, you'll have to wait until the next update to know if you recovered or not.

In addition, we can clearly see that one of the major components is counting anchor text. This indicates that **they likely penalized webpages when too many repeating anchor text are detected.**

We also see that **pages can have a 'bad links' flag.** While it isn't clear how a page gets this flag, I speculate that it might be manually assigned. In the early days of Penguin, Google used a lot of manual reviewers.

Finally, **they check homepage links between sites hosted on the same server**. For instance, if you have 10 or more sites all on the same host, and all the homepages link to a central location, they'll likely penalize the site for link abuse.

```
nonLocalAnchorCount
mediumCorpusAnchorCount
penguinEarlyAnchorProtected - Doc is protected by goodness of early anchors
droppedHomepageAnchorCount
redundantanchorinfoforphrasecap
forwardedOffdomainAnchorCount
droppedNonLocalAnchorCount
perdupstats
onsiteAnchorCount
droppedLocalAnchorCount
penguinTooManySources - Doc not scored because it has too many anchor sources
forwardedAnchorCount
anchorSpamInfo - This structure contains signals and penalties of AnchorSpamPenalizer
```

Here we see that **Penguin is looking specifically at the external anchor text links**. While we haven't seen any benefits to repeating the same internal anchor text link over and over again, **internal links seem to be spared from any Penguin penalty**.

Another interesting reference is that there is a mechanism by which **a page can be "protected" by early, trusted anchors texts**. This was likely put in place to shield against negative SEO.

Finally, it's interesting that there is actually an anchor text limit! **You can have too many links from different sources.** For example, if a page receives 500,000 links (don't laugh, it happens), then Google will not calculate the score.

`lowCorpusAnchorCount`  
`lowCorpusOffdomainAnchorCount`  
`baseAnchorCount`  
`minDomainHomePageLocalOutdegree` - Minimum local outdegree of all anchor sources that are domain home pages as well as on the same domain as the current target URL  
`skippedAccumulate` - A count of the number of times anchor accumulation has been skipped for this document  
`topPrOnsiteAnchorCount` - According to anchor quality bucket, anchor with pagerank > 51000 is the best anchor. anchors with pagerank < 47000 are all same  
`pageMismatchTaggedAnchors`  
`spamLog10Odds` - The log base 10 odds that this set of anchors exhibits spammy behavior  
`redundantanchorinfo`  
`pageFromExpiredTaggedAnchors` - Set in SignalPenalizer::FillInAnchorStatistics  
`baseOffdomainAnchorCount`  
`phraseAnchorSpamInfo` - Following signals identify spike of spammy anchor phrases  
`anchorPhraseCount` - The number of unique anchor phrases  
`ondomainAnchorCount`  
`totalDomainsAbovePhraseCap` - Number of domains above per domain phrase cap

When it comes to backlink abuse, Google really hates homepage links from PBNs. They have a dedicated section checking for PBN homepage links. **If most of your incoming links come from homepages... then you're in trouble.**

We see that there's a **special trust factor for anchors with PR5+ of power.** This is interesting and lines up with what I have personally experienced in the past: *Having one single link from a high PR page with relevant anchor text makes a huge difference.*

In addition, we see that Google pays attention to anchors coming from pages marked with "expired". The mechanism behind this is that Google **likely applies a flag to an entire domain if they notice that an expired domain has been revived with the same content**. Then all the outgoing links from that domain will be marked with "expired" for a certain period of time.

*As there are surely legitimate cases where someone might re-use an old expired domain, I speculate there must be some time limit associated with the expired flag.*

And finally, as we have previously seen, Google does have limits for the quantity of incoming link. Here they clarify that **you can have a maximum of 5000 anchors pointing to a page**. After that, they might ignore them or perhaps assign a penalty.

totalDomainsSeen - Number of domains seen in total  
 topPrOffdomainAnchorCount  
 scannedAnchorCount - The total number of anchors being scanned from storage  
 localAnchorCount  
 linkBeforeSitechangeTaggedAnchors  
 globalAnchorDelta - Metric of number of changed global anchors computed as, size - intersection  
 topPrOndomainAnchorCount  
 mediumCorpusOffdomainAnchorCount  
 offdomainAnchorCount  
 totalDomainPhrasePairsSeenApprox - Number of domain/phrase pairs in total  
 skippedOrReusedReason - Reason to skip accumulate, when skipped, or Reason for reprocessing when not skipped  
 anchorsWithDedupedImprovanchors - The number of anchors for which some ImprovAnchors phrases have been removed due to duplication within source org  
 fakeAnchorCount  
 redundantAnchorForPhraseCapCount - Total anchor dropped due to exceed per domain phrase cap  
 totalDomainPhrasePairsAboveLimit - The following should be equal to the size of the following repeated group, except that it can go higher than 10,000  
 timestamp - Walltime of when anchors were accumulated last



When it comes to link building, **Google calculates the quantity of redundant anchors texts.** (Repeating the same anchor likely gets you into trouble.)

They also look at the **total quantity of domains linking, go too far and it can set off red flags.**

And of course, **they look at how many anchors per domain you have.** (ie: In other words, **avoid site wide links** from another domain as it will surely set off a red flag within the Penguin algorithm.)

## Avoiding Link Penalties

The MAIN thing you should remember should you engage in link building is that you want to **VARY** your anchor text.

**Too many exact anchor texts, built too quickly, from too many sources... will trigger a penguin penalty.**

Therefore, you'll likely want to **build links slowly over time while varying the anchor text**. While it *might* be fine to have a few repeating anchor texts I am very cautious to avoid excessive repetition.

*(In the past, I have personally acquired websites that have never expired / never dropped from the index. I then made sure to maintain them for a long period of time, 6 months to a year, before using them to build links. I never linked from the homepage and instead, I created hyper-relevant articles and linked out from the main content in a contextual manner. In addition, I timed my links so I never received more than 1 link per day and ideally, I only acquired 1 link every few days. Of course, the anchor text always changed in order to avoid anchor text issues.*

*Finally, when I did acquire sites with a potentially troubled past, I swiftly redirected them to sub-directories of new sites in order to filter out any flags associated with them. I avoided redirecting websites to the root of other websites and I always redirected into a directory. None of this is ranking advice as it is against Google's official guidelines to partake in any link building...*

*Plus, these days I have migrated to acquiring more 'whitehat' links from large authority sites.)*

## Negative SEO

I was pleasantly surprised to discover a section dedicated to protecting websites from negative SEO attacks.

(It appears as if Google cares about SEOs)

`demotedEnd` - End date of the demotion period  
`demotedStart` - Start date of the demotion period  
`phraseCount` - Following fields record signals used in anchor spam classification. How many spam phrases found in the anchors among unique domains  
`phraseDays` - Over how many days 80% of these phrases were discovered  
`phraseFraq` - Spam phrases fraction of all anchors of the document  
`phraseRate` - Average daily rate of spam anchor discovery

It's very cool to see this: It appears as if there's an active "spam shield" that is activated when it detects a negative SEO attack.

**If an abnormal quantity of links begins to point towards a specific page, it will temporarily demote the page while the attack is underway. Once the attack is over, it preserves the good anchor text links (built before the attack) while eliminating the bad ones created during the attack phase.**

This approach is clever because it makes the negative SEO attacker *think* their techniques are working, as they might see a temporary demotion of their target. Then, once they stop, the page magically bounces back to where it was.

# Site Quality - Rewards & Penalties

## Core Updates, Panda and Topical Authority

This might be one of the most important sections of the entire API reference repository, as it explains how Google evaluates entire websites. **Site quality dictates how all the pages on a website will rank**; thus, understanding the site quality algorithm is critical for good SEO practices.

**Furthermore, if you've been impacted by a recent penalty, the site's quality is likely the root cause.**

`ugcDiscussionEffortScore` - UGC page quality signals

`productReviewPDemoteSite` - Product review demotion/promotion confidences

`exactMatchDomainDemotion` - Page quality signals converted from fields in proto QualityBoost in `quality/q2/proto/quality-boost.proto`

The first API reference discusses user generated content which is likely **used to evaluate the effort that goes into forum threads, comments and possibly sites like Reddit**. What is interesting here is that they are measuring the "effort" of the discussion.

**I suspect that the page might get a boost if the discussion is deemed high effort.**

Next, we have "Product Review site" promotion or demotion used to identify and rate **product reviews**. *(Secretly, we all know this is demotion as it has been difficult to rank product reviews unless you're on a major authority website.)*

Last, we have an "Exact match domain demotion" which was likely introduced to prevent sites like: **best-drone-reviews.com** from ranking. The act of registering exact match domains for ranking purposes was quite popular back a decade ago... these days, I personally recommend creating a memorable brand name for you

**nsrConfidence** - NSR confidence score: converted from quality\_nsr.NsrData

**lowQuality** - S2V low quality score: converted from quality\_nsr.NsrData, applied in Qstar

**navDemotion** - nav\_demotion: converted from QualityBoost.nav\_demoted.boost

**siteAuthority** - site\_authority: converted from quality\_nsr.SiteAuthority, applied in Qstar

Now we get to the "site wide" elements that can make an entire site tank.

We have "nsrConfidence" which is the confidence in the "Normalized Site Rank" score. **As we previously covered, the Normalized Site Score is most likely a measure of how the site performs compared to the rest of the industry.** I believe it is one of the most important metrics and this "nsrConfidence" evaluates how trustworthy the score is.

Then we have "lowQuality" which is likely a flag for a bad site. We see that it is pulled from the NSR data which means that **when the normalized site rank is too low, the site receives a "low quality" flag...** and then likely doesn't rank at all. If you've seen websites remain in the index but refuse to rank, then this is probably the reason.

The "navDemotion" is likely a demotion related to NavBoost, perhaps calculating how much it should drop a site in rankings.

And of course... *\*drumroll\** they have a measure of site authority in the form of "siteAuthority".

This is notable for two reasons:

1. Google has denied having a site authority metric in the past. They LITERALLY have an API reference called SiteAuthority so I don't feel as if the Google spokesperson was being straightforward.
2. Once again, **we see that the "site authority" metric is derived from the normalized site rank!** If you weren't already convinced, hopefully you can see how important normalized site rank is within the entire algorithm.

## Site Quality

**BOTH the "Low Quality" flag and the "Site Authority" metric come from NSR data.**

It appears as if a large portion of Google's algorithm revolves around user interactions with the website which would explain why some sites are struggling to rank after the latest core updates.

Websites can be hit by MULTIPLE demotions originating from one measurement: NSR.

`babyPandaV2Demotion` - New BabyPanda demotion, applied on top of Panda

`authorityPromotion` - authority promotion

`anchormismatchdemotion` - anchor\_mismatch\_demotion

`crapsAbsoluteHostSignals` - Impressions, unsquashed, host level, not to be used with compressed ratios

I knew that they had worked on revisions of Panda a few years ago as I was involved in recovering websites affected by Panda (I created case studies on how to systematically recover websites from Panda)

However, I wrongfully assumed that they had simply updated the existing Panda algorithm... It seems as if introduced a NEW panda algorithm called: "Baby panda". Surprisingly, it seems as if **this new baby Panda (user experience related algorithm) is applied ON TOP of the original Panda?**

**Yikes! This means you can potentially be penalized twice for poor user experience.**

Conversely, there also seems to be a boost for 'authority sites' so perhaps if your user signals are great, it go the other way. And of course, **a demotion if your anchor text links don't match the content.**

**Finally, raw click signals are used to evaluate the site's performance as well. Click signals (from Chrome) are incredibly important for determining how a site performs in search.**

`topicEmbeddingsVersionedData` - Versioned TopicEmbeddings data to be populated later into superroot / used directly in scorers

`scamness` - Scam model score

`unauthoritativeScore` - Unauthoritative score

We see that the topics of the site can be accessed by the algorithm. This makes sense as **Google needs to be able to retrieve sites that are topically relevant to the query.**

Then we have a measure of "Scamness" for a website. While we don't know much about this API reference, we can infer that they are using AI to measure it.

And finally, an unauthoritative score. Perhaps this is based on links, user experience or originality of content.

`pandaDemotion` - This is the encoding of Panda fields in the proto SiteQualityFeatures in `quality/q2/proto/site_quality_features.proto`

Here we find the original Panda!!

## The Original Panda Algorithm

*The Panda algorithm has a soft spot in my heart as I was on the main stage of Traffic & Conversions in front of 2000 entrepreneurs explaining how to recover websites from a Panda penalty. I spoke around the world explaining how to take advantage of Google algorithms to rank better online.*

*For those wondering, the trick to recovering from Panda is to trim the fat from the site while focusing on user accomplishment / experience. We want every 'landing page' from Google to provide a good user experience. In my presentation, I would share a repeatable process for making a site 'sticky' and share how to remove all the redundant/duplicate/low quality pages as determined by visitor analytics.*





While the Google algorithm has changed significantly since then, **Google is still focused on the overall user experience...** albeit they are measuring it in slightly different ways.

The recent Helpful Core Update, March Core update and more are all focus on the user... and they are likely using additional signals (ie: from clicks) to measure it.

**It does seem somewhat unfair that all of these penalties stack on top of each other.**

From unauthoritative score, to click signals, to normalized site rank, to Panda, and even BabyPanda—if you're affected by one, then you're likely to be affected by all of them. This is the main reason why some websites impacted by one algorithm change notice a subsequent drop after the next update, and so forth.

## Site Score

I found a section that discussed an explicit "siteScore" and dove into what that could possibly entail.

```
siteFrac - What fraction of the site went into the computation of the site_score  
siteScore - Site-level aggregated keto score  
versionId - Unique id of the version
```

First, I think it's notable because just like site authority, Google has also mentioned there is no site score...

Well, there is.

Interestingly, **they don't use the entire site to calculate the site score.** Instead, they note the percentage of the site used to estimate the score. This is likely to save resources.

```
pageEmbedding  
siteEmbedding - Compressed site/page embeddings  
siteFocusScore - Number denoting how much a site is focused on one topic  
siteRadius - The measure of how far page_embeddings deviate from the site_embedding  
versionId
```

Here we have a section that many SEOs will be interested in... topical authority.

When Google refers to embeddings in the context of search, it means transforming words and phrases from web content into vector representations. These vectors help Google's algorithms understand and quantify the relationships and relevance among

different textual entities, enhancing the accuracy of search results.

**In plain English, it's like Google creating a digital map of all the words and phrases found on the website.**

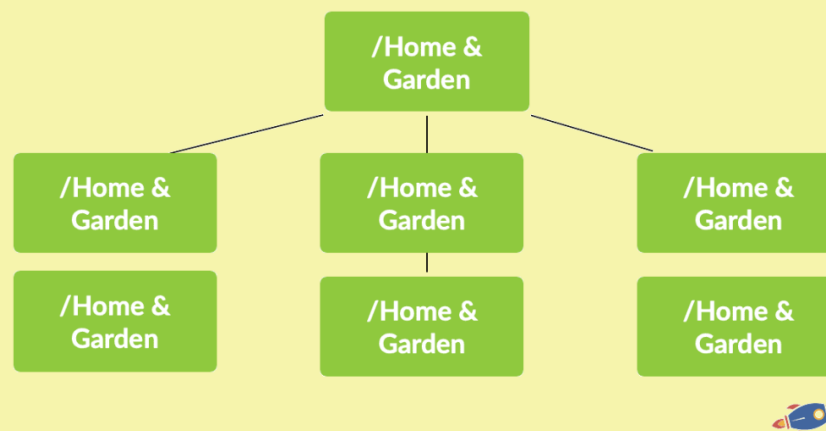
We see with the "SiteEmbedding", "SiteFocusScore" and "SiteRadius" that **Google looks at both the page embedding AND the site embedding to determine the topic. This means that other content on your site will dictate how well you rank.**

**In addition, it also measures how focused on a topic a site is... very likely providing a significant ranking boost for sites that have a narrow focus.**

And finally, it will measure how 'unrelated' a page is to the rest of the site. Creating unrelated content on a website will likely not rank as well.

## Ranking With Topical Authority

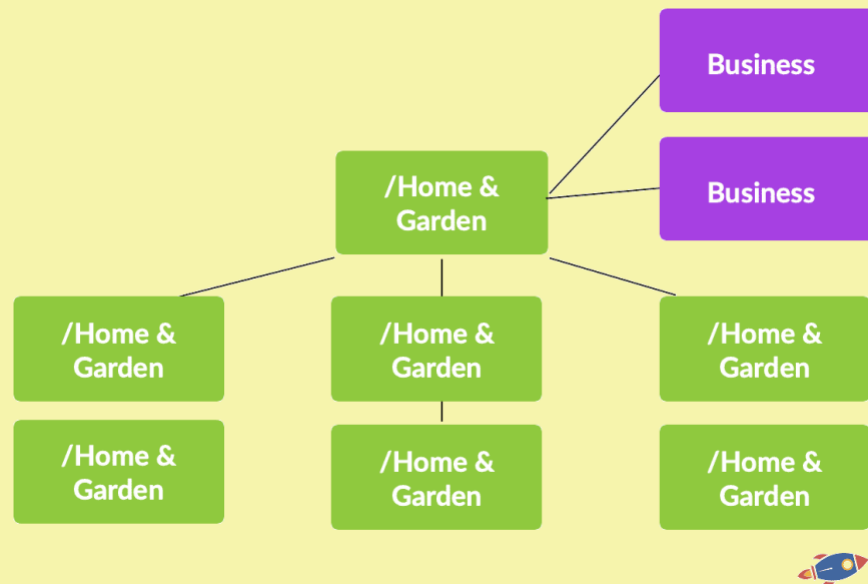
Building a topical authority site plays a key role in ranking for terms. **Sites with a narrow focus will fare better as Google measure how related (or unrelated) a page is when compared to the rest of the site.**



This boost might be necessary to compete against giant websites.

So it is likely a **good idea to start a site with a narrow focus** and expand over time as the site builds up power.

Eventually when the site accumulates enough links and authority, you'll likely want to branch out to other topics. These topics might take a little time to take off as you build up content in that section however they too, will end up ranking.



*(I personally start with 1 single piece of content that acts as the centerpiece for my site.*

*I then map out a list of subsequent, semantically related content. To find these semantically related ideas, I'll typically go to Wikipedia and read the page most closely related to my main topic. For example, if the center of my cluster is on SEO, I'll find the Wikipedia page most closely related to Search Engine Optimization.*

*In addition, I try determine the Google NLP category of the keyword by looking analyzing at the top ranking results on Google. (If I see that the top 3 results for that topic all land in the same category, I can reasonably predict that if I write an article on that topic, it will also land in that category).*

*It isn't perfect, however I try to make sure that follow-up content lines up with existing categories.*

*Finally, clusters also require internal links so I will interlink similar and complimentary topics together with relevant anchor text within my content.)*

## Content That Ranks

One of the my favorite areas of SEO, on-page optimization! In this section of the API reference documentation, we dive into how Google looks at content.

`entity` - Entities in the document

`semanticNode` - The semantic nodes for the document represent arbitrary types of higher-level abstractions beyond entity mention coreference and binary relations between entities

`hyperlink` - The hyperlinks in the document. Multiple hyperlinks are sorted in left-to-right order

`lastSignificantUpdate` - Last significant update of the page content, in the same format as the contentage field.

The first thing we notice is that **they are looking at both the entities AND the semantic node in which this document lies when evaluating content.** In other words, if this document is part of a related, relevant set of documents.

This is just another reason to have topically aligned content.

Next, **they are looking at the outgoing links in a document.** While we had already tested for this (pages with relevant outgoing links ranked better than pages without any links), it's nice to confirm that they are monitoring it.

And of course, **they are looking for fresh & updated content**, as noted by "lastSignificantUpdate".

Google knows the difference between minor updates and significant updates.

## Optimizing Content For Rankings

When optimizing a page, if you want Google to recrawl and rescore the page, you have to modify a *significant* amount of text on the page.

**I have found that adding an extra paragraph of text will typically be enough to trigger a full re-evaluation of the content.**

In contrast, adding 4-5 words won't be enough. Google likely saves resources when there are only small modifications to the page.

So if you're optimizing and trying to add more relevant entities into your content, aim to perform a significant update.

*(In my experience, it typically takes approximately 3-4 weeks for Google to fully re-crawl and re-calculate a page score after you've performed a significant update. You might get a freshness boost before that though!)*

`entityLabel` - Entity labels used in this document

`topic`

`golden` - Flag for indicating that the document is a gold-standard document

Once again, we see that **Google sort documents by using entities.**

Entities are specific words or phrases that are recognized as representing distinct and well-defined concepts or objects, each carrying an associated meaning based on real-world references.

For example, the word "party" can have multiple meanings.

1. A party of 5 people
2. A political party
3. Let's go to the party!
4. I like to party with friends

While the word stays the same, **the entity identifies which one of these categories the word falls into.** Within a machine learning context, it is very important to make a distinction between a political party... and politicians having a party! *(And that's why*

*Google uses entities to classify and rank all documents on the web)*

Closely related to entities are topical categories, which they mention when discussing "topics".

Surprisingly, **they can also flag "Golden" documents that human reviewers deem important or as a gold standard.** I'm not sure to which capacity the golden flag is used however it would surely give the document an unfair advantage over all documents.

**focusEntity** - Focus entity. For lexicon articles, like Wikipedia pages, a document is often about a certain entity

**syntacticDate** - Document's syntactic date

**privacySensitive** - True if this document contains privacy sensitive data. When the document is transferred in RPC calls the RPC should use

Here we see that **when Google is analyzing a document, they try to identify ONE main focus entity.** I'll sometimes refer to this as main keyword of a document even though it should probably technically be called the "focus entity" (somehow that just doesn't have the same ring to it).

An interesting bit is that they also note **if a date is mentioned in the title / URL, then they check if this aligns with the other dates found within the document.**<sup>[1][2]</sup> A while ago, a Google engineer mentioned that you should avoid just updating the date in the title without updating any other information on the page... this is likely why.

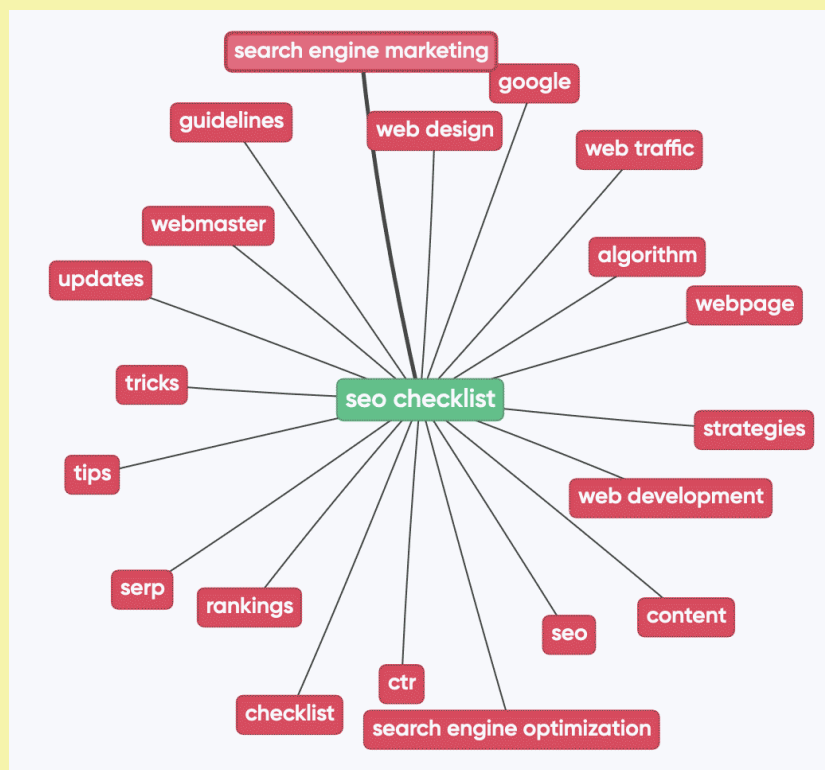
Finally, there's a special note for checking if the page contains private information. (For example, a person's home address, credit card, maybe social security number or phone number.) In my experience, **when Google finds private and sensitive information, the page is less likely to rank.** <cough> negative seo tactic</cough>

## Better On-Page Ranking

Google uses entities, topics and semantic nodes to classify a document.

This means that **your page can appear for queries even if the words don't appear on the page** (because that term might be recognized as a topic or might appear in the semantic nodes).

They also have a focus entity... which tries to identify the MAIN focus of your page.

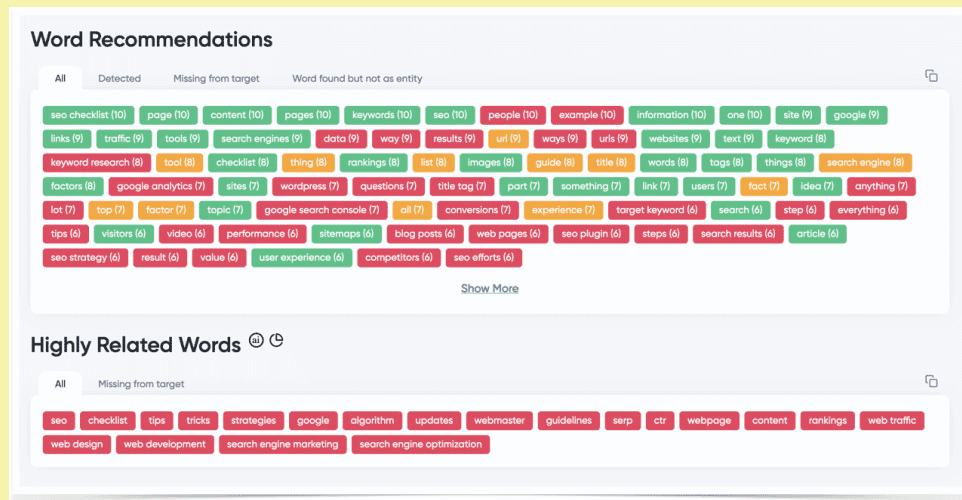


*Highly Related Entities*

<http://on-page.ai>

**I personally try to add all the top related entities multiple times on my page.**

In addition, I will add the focus entity (the one I want to rank for) within the title, headline, text and images.



**Top Entities (Analyzed With Google's NLP Model)**

<http://on-page.ai>

This helps Google build a highly relevant embedding, it sets the correct focus entity and will maximize my chances of ranking.

*(During the recent core updates, I have noticed a trend in which pages that have a higher exact & related entity density ranked above pages with a lower density of related entities. I try to use all the top entities multiple times within my content and compare the density of my pages against my competitors. I don't stop until I have a higher entity density.)*



## Page Information

Within this section of the API reference material, Google shows us the information they store about webpages.

`cdoc` - This field contains reference pages for this entity

Apparently they have reference pages for entities... like Wikipedia!

That's cool... because **it means if you want to be highly relevant for an entity, you might be able to look up the Wikipedia page associated with the entity and include similar terms.**

`linkInfo` - Contains all links (with scores) that Webref knows for this entity. Links are relationships between entities

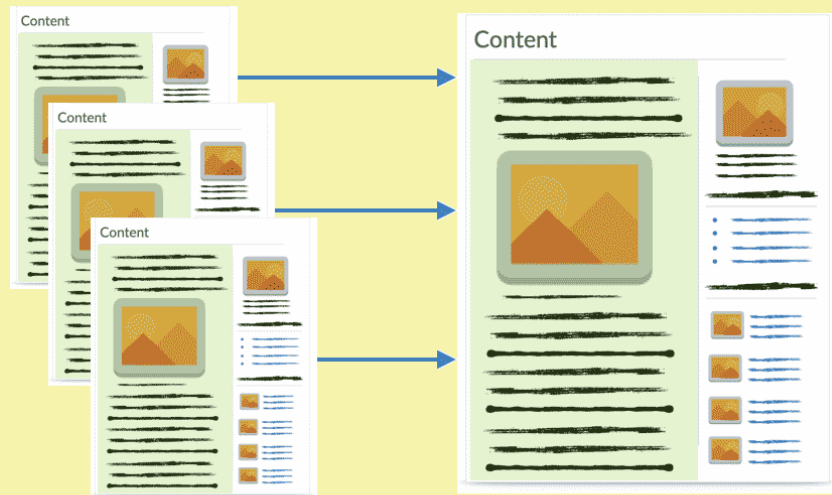
`nameInfo` - Contains all names (with scores) that Webref knows for this entity

In addition, there's an API to list all the links with scores associate with a page! *(This would be cool to use. Imagine Ahrefs site explorer... but with Google's own data. It would be incredible to see the actual scores associated with each link built to a page. This would make link building SO much easier.)*

**We can also confirm that links are relationships between entities. We already knew this however it re-confirms that you want to get links from related content.**

We also see that Google makes a note of all the names associated with page. This is likely very useful when deciding which documents to retrieve from the index.

## Better Ranking



Ranking on Google can sometimes get fairly complex. So when strangers ask me how to rank, **I usually share that it comes down to 3 primary things:**

- 1. Good site**
- 2. Good content**
- 3. Good links**

If you can deliver on all 3, then you'll usually be just fine (*Assuming Chrome visitors agree that your content is good... which leads to Google thinking you have a good site.*)

Plus, once I know I'm working on a site that is in good standing... then it becomes really easy. **All I need to do is produce highly optimized content (with a lot of entities) that lines up with the pre-existing topical content. Add a few relevant internal links... and it ranks!**

We continue digging into the page information and see they have a section for original content.

`originalcontent`

Personally, I believe this refers to original content vs duplicate content.

And not, 'how original' content is.

`badSslCertificate` - This field is present iff the page has a bad SSL certificate itself or in its redirect chain

**Google will likely not rank you if you have a bad SSL certificate.**

`registrationinfo` - Information about the most recent creation and expiration of this domain

Google cares a LOT about the recent creation and expiration date of a domain. In this section they elaborate (not shown) on something called a "DomainEdge signal" that I suspect is likely used to fight PBNs.

`richsnippet` - rich snippet extracted from the content of a document

It's interesting to see that **every webpage has a rich snippet** (even if it's not shown). More on how Google evaluates rich snippets later on.

## Document Info

`sitemap` - Sitelinks: a collection of interesting links a user might be interested in, given they are interested in this document

`csePagerankCutoff` - URL should only be selected for CSE Index if it's pagerank is higher than `cse_pagerank_cutoff`

Interestingly, **they store a list the related pages to a document which is likely determined by user behavior**. Within patents, they describe that they create associations between subsequent user searches.

*Perhaps if you search for document X and THEN search for document Y, an association is created between X and Y.*

Another small discovery within the document information section is that Google has an option to NOT show a page if it's Pagerank is lower than a pre-determined amount.

## Auto-Suggest SEO

*While this isn't exactly what the API is referencing, in the past, I might or might not have manipulated Google search suggestions.*

*Using mobile devices, I may have instructed searchers to search for an initial term, click search and then return to Google to search for a different, associated term. Through consistent daily searches, the association may have eventually linked up and auto-suggest may have shown suggestions for the complimentary search.*

*This is how many of the search terms now have "reddit" at the end... except this works for different website brand names. The only pitfall is that it requires consistent searches over a long period of time so it could be a hassle to get going and maintain.*

## Avoiding Penalties - Content Evaluation

This section dives deep into content, penalties and spam. Our aim is to understand what constitutes great optimized content while avoiding over-optimization

**uacSpamScore** - The uac spam score is represented in 7 bits, going from 0 to 127

**spamtokensContentScore** - For SpamTokens content scores. Used in SiteBoostTwiddler to determine whether a page is UGC Spam

**trendspamScore** - For now, the count of matching trendspam queries

**ScaledSpamScoreYoram** - Spamscores are represented as a 7-bit integer, going from 0 to 127

Inside the documentation, there are quite a few spam flags. The first, "uacSpamScore" might stand for User Automation or User Activity.

The next "Spam tokens content score" & "trendspamScore" references suggests that **Google might have a list of spammy words they use to measure spam**. Perhaps many mentions of casino / viagra might trigger it.

### Trending Spam Topics

Each year, there are new spam topics that trend on the internet.

From the age old viagra... to new protein powders, gummy CBD edibles, new bitcoin slot machines, etc.

**Spam evolves through the years and it appears as if Google keeps track of it. (TrendSpamScore)**

Unless you are explicitly targeting a high risk term, avoid having comments / multiple spam recognized entities on the page.

`datesInfo` - Stores dates-related info (e.g. page is old based on its date annotations)

`ymlHealthScore` - Stores scores of yml health classifier as defined at [go/yml-classifier-dd](#)

`ymlNewsScore` - Stores scores of yml news classifier as defined at [go/yml-classifier-dd](#)

Within the date section, Google mentions a "FreshnessTwiddler" (Twiddlers are modifiers used by Google, usually to boost rankings). In spite of what Google has claimed in the past, **this is very likely a freshness ranking boost given to fresh content.**

In addition, we see that they do, in fact, have a "YourMoneyYourLife" score. Whenever you're posting content on the web, Google is checking to see if it falls within this category and if it does, there might be an additional layer of verifications and/or requirements.

The next API reference is incredibly important:

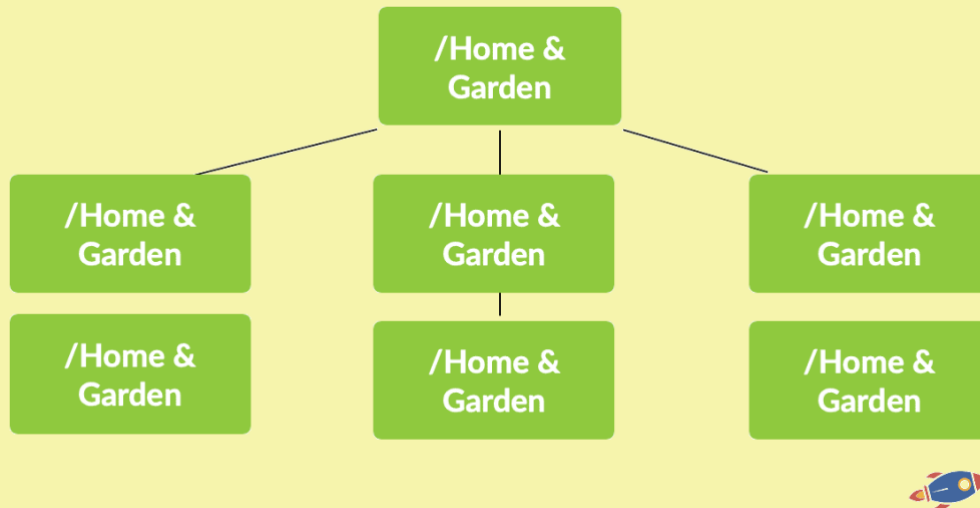
`topPetacatTaxId` - Top petacat of the site

The 'TopPetacatTaxID' API reference, in my opinion, indicates that topical relevance is very important when it comes to Google ranking. **This suggests that Google is classifying websites and assigning ONE MAIN category to them.**

**It is likely that this categorization is used throughout Google's algorithm, influencing both content and link building.**

**The bottom line is that content aligned with the site's primary topic receives a ranking boost.** For example, imagine a user searching for 'best dog food for puppies.' The 'SiteboostTwiddler' would analyze the query, recognize that it pertains to pet food, and then use the 'TopPetacatTaxID' to prioritize results from the top category of pet food.

## Topical Authority Gets A Boost



Google rewards topical authority sites in multiple ways and **TopPetacatTaxID, used in SiteBoostTwiddler is just another boost used reward sites that are focused on a specific topic.**

For example, if the query is in /home & garden/ and your site's main category is /home & garden/, then it would be logical to assume that the site should get a boost.

While we don't know exactly the new categories that Google is using, I believe this is the old list they might have used in the past (sample down below):

/Arts & Entertainment/Celebrities & Entertainment News  
/Arts & Entertainment/Other  
/Arts & Entertainment/Comics & Animation/Anime & Manga  
/Arts & Entertainment/Comics & Animation/Cartoons  
/Arts & Entertainment/Comics & Animation/Comics  
/Arts & Entertainment/Comics & Animation/Other  
/Arts & Entertainment/Entertainment Industry/Film & TV Industry  
/Arts & Entertainment/Entertainment Industry/Recording Industry  
/Arts & Entertainment/Entertainment Industry/Other  
/Arts & Entertainment/Events & Listings/Bars, Clubs & Nightlife  
/Arts & Entertainment/Events & Listings/Concerts & Music Festivals  
/Arts & Entertainment/Events & Listings/Event Ticket Sales  
/Arts & Entertainment/Events & Listings/Expos & Conventions  
/Arts & Entertainment/Events & Listings/Film Festivals  
/Arts & Entertainment/Events & Listings/Food & Beverage Events  
/Arts & Entertainment/Events & Listings/Live Sporting Events  
/Arts & Entertainment/Events & Listings/Movie Listings & Theater Showtimes  
/Arts & Entertainment/Events & Listings/Other  
/Arts & Entertainment/Fun & Trivia/Flash-Based Entertainment  
/Arts & Entertainment/Fun & Trivia/Fun Tests & Silly Surveys  
/Arts & Entertainment/Fun & Trivia/Other  
/Arts & Entertainment/Humor/Funny Pictures & Videos



While they might have a new category system that is being used internally, we're still using the old category list to classify websites. It isn't perfect however it does give us a fairly decent idea of the top category.

Domain Relevance Report			
We highly recommend exporting the full report			
		Export	Compare category...
<input type="radio"/>	#	Domain	Category
<input type="radio"/>	1	dangerousroads.org	/Autos & Vehicles /Reference/Geographic Reference/Maps /Travel
<input type="radio"/>	2	continenthop.com	/Food & Drink/Cooking & Recipes /Food & Drink/Restaurants /People & Society /Travel
<input type="radio"/>	3	anglotopia.net	/Arts & Entertainment
<input type="radio"/>	4	belaroundtheworld.com	/Online Communities/Blogging Resources & Services /Shopping/Apparel /Travel
<input type="radio"/>	5	52perfectdays.com	/Food & Drink/Cooking & Recipes /Food & Drink/Restaurants /Travel /Travel/Tourist Destinations/Beaches & Islands
<input type="radio"/>	6	asoulwindow.com	/Arts & Entertainment /People & Society /People & Society/Religion & Belief /Travel /Travel/Tourist Destinations
<input type="radio"/>	7	creativetravelguide.com	/People & Society/Family & Relationships/Family /Travel /Travel/Tourist Destinations/Beaches & Islands
<input type="radio"/>	8	bluedreamer27.com	/Travel/Air Travel /Travel/Cruises & Charters
<input type="radio"/>	9	awanderfulsole.com	/Online Communities /Travel

### Top Categories (Analyzed With Google's NLP Classification Model)

<https://on-page.ai>

(Topical authority requires a significant quantity of content. However, the benefit is that when Google recognizes your authority on a subject, ranking becomes significantly easier and requires less links.

My strategy:

1. I will, as previously stated, look up the Wikipedia page for my main focus entity in order to get ideas for the topics I want to cover.
2. I scrape competitors' website sitemaps to gather more topically related keyword ideas. Since not all articles will be relevant, I run 100-200 of their pages through the NLP category checker. This allows me to easily identify content that falls within the same category.

When building topical authority sites, remember to create relevant internal links. You can't have clusters & nodes without links!

Continuing with the content, we see that they have an explicit mention of: "OriginalContentScore".

**OriginalContentScore** - The original content score is represented as a 7-bits, going from 0 to 127

**DocLevelSpamScore** - The document spam score is represented as a 7-bits, going from 0 to 127

Although it might still be a ratio of duplicate content versus original content... I believe that this API reference is **measuring how original the content is versus your competitors**. Google has long encouraged publishers to create new, original content and this seems to be an attempt to measure that effort.

I believe that if you're just creating a carbon copy of the existing search engine results, there is no real incentive for Google to rank you above the rest. **That's why I encourage webmasters to go above and beyond the #1 result, using entities that your competitors aren't currently using and adding unique information on your subject.**

**freshnessEncodedSignals** - Stores freshness and aging related data, such as time-related quality metrics predicted from url-pattern level signals

**ScaledSpamScoreEric**  
**biasingdata**  
**ScaledExptSpamScoreEric**

Once again, we see another freshness signal that indicates that Google is rewarding new and recently updated content.

And...

There's spam score if your name is Eric? That's not nice ☹

Now here's something I didn't expect:

`biasingdata2` - A replacement for `BiasingPerDocData` that is more space efficient

`spamCookbookAction` - Actions based on Cookbook recipes that match the page

To my surprise, they seem to have "Biasing data" entry. **This might be a signal that measure how neutral or bias the content is.** ie: overly salesy affiliate pages might not rank as well.

And finally, `spamCookbookAction` hints that **there are a specific set of rules that trigger spam on pages.** For example, invisible text and other shadier tactics might fall into this category.

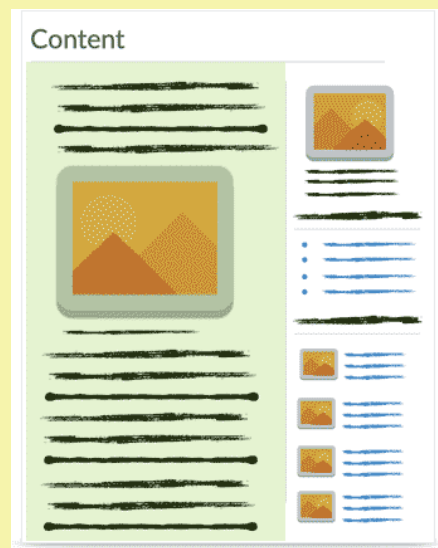
## Document Quality: Bias

It's super interesting to discover a mention of bias in the algorithm.

Anecdotally, I noticed that overly promotional / highly biased affiliate articles weren't performing as well as my more neutral articles.

This might be why!

From now on, I will avoid overly positive articles and will adopt a more professional, neutral tone when reviewing products.



As we continue looking into the content documentation...

**localizedCluster** - Information on localized clusters, which is the relationship of translated and/or localized pages

**KeywordStuffingScore** - The keyword stuffing score is represented in 7 bits, going from 0 to 127  
**spambrainTotalDocSpamScore** - The document total spam score identified by spambrain, going from 0 to

We see even more indications that **topical relevance plays a significant role in ranking**. Whenever Google speaks of clusters (the relationship between local pages), they are denoting the topical content. The idea is that if you're an expert on a specific cluster, then you will likely rank better when there are queries within that cluster.

Next, they have a specific keyword stuffing score! We've long known **that Google prohibits keywords stuffing** so it's nice to see it here in the flesh. Even today, I still see instances of webmasters keyword stuffing their titles (don't do this).

Finally, the **spambrainTotalDocSpamScore is an indication that they also use AI to estimate the document spam score**. Whenever Google mentions "brain" in an algorithm, it is their way of saying they are using machine learning to accomplish the task. This implies that they have trained a machine learning algorithm on a slew of spammy documents and then ask the AI to classify your document's spam level.

We don't know exactly what kind of training data Google originally provided for the AI spam algorithm however it's logical to assume that if your document looks like a spammy document, then it will likely be classified as one.

**spamrank** - The spamrank measures the likelihood that this document links to known spammers

As we continue digging deeper into the data used within content classification, we discover that "spamrank" is a measure of the likelihood that document links to spam.

Suffice to say, **if you link to bad places within your content, this will hurt your rankings**.

This also means we have to watch out for accidental links within our content. Sometimes there are malicious actors that will try to add hidden links within content, sometimes people will add redirects after a link is placed and sometimes we can make typos in our URLs leading to incorrect placements.

```
compressedQualitySignals
```

```
crowdingdata
```

QualitySignals & Crowdingdata both likely represent user signals. Google *cares a lot* about user signals, to the point of saying that "*they don't understand content, they fake it*".

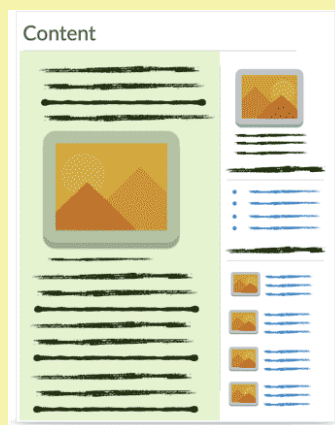
New websites have a sandbox period to prevent spam.

## Crowd Data

We've known that Google measures how humans react to the content and this confirms it.

When creating content, I try to make it as captivating as possible for humans.

1. I aim to make the reader feel *"as if they are at the right place for the information"* as quickly as possible.
2. I quickly establish myself as a **trusted authority** on the subject.
3. I include **charts, tables and other imagery** to captivate my reader's attention.



*(One of my favorite tricks is to include a relevant chart, graph or image that cuts off at the fold.*

*This gets people scrolling and I noticed that when people start scrolling down a page, they are MUCH more likely to consume the rest of the page. The most frequent bounces come from people that never start scrolling!)*

**hostAge** - The earliest firstseen date of all pages in this host/domain. These data are used in twiddler to sandbox fresh spam in serving time

We've long know that there was a sandbox period for new websites however some Google employees denied this in the past. Now we see that **new websites in fact DO have a sandbox period to prevent spam**. It is determined by hostAge.

## Spam & Content

As we continue on our journey into Google's spam filters for content, there are a few more interesting tid-bits:

**GibberishScore** - The gibberish score is represented in 7 bits, going from 0 to 127

**freshboxArticleScores** - Stores scores of freshness-related classifiers

**onsiteProminence** - Onsite prominence measures the importance of the document within its site

First, we have "Gibberish score". This is likely to address lorem ipsum testing and/or just random characters on a page. *(Anecdotally, this must not be the strongest ranking signal as I still have pages ranking with pure 0000 and entities which should, in theory, trigger a relatively high gibberish score.)* Still, it's nice to know that it at least plays some role in ranking.

Next, we have fresh again... however this **freshness might be measured on various different levels. For example, freshness at the host level, blog level and finally, page level.** Perhaps it's not enough to have a single fresh article if the website hasn't been updated in ages? (That said, if an article is updated, then wouldn't the site also be updated?)

And last, this is a really cool discovery: onsiteProminence: **Google tries to determine how important a page is on a website.** They determine it by computing from **SIMULATED TRAFFIC FROM HOMEPAGE!**

## Better Rankings

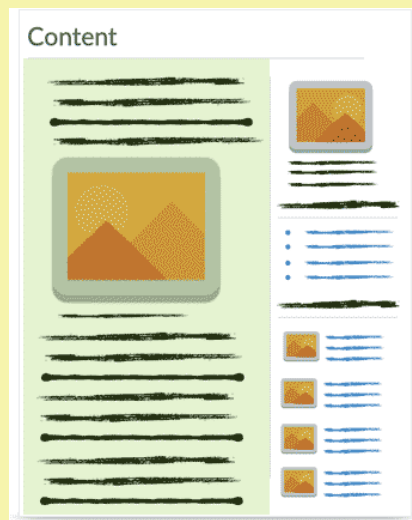
Contrary to what Google has stated in the past...

**Fresher content is usually better.**

However this appears to apply to more than just a single article and might take into account the overall site's freshness. A full site "update" might help with this.

**In addition, Google measures how deep in your site the content is... the closer to the homepage it is, the better it will rank!**

*(I typically include a high quantity of internal links from the homepage. I feel as if the homepage should act as portal to the rest of the site and I link my best content directly from the homepage.)*





**commercialScore** - A measure of commerciality of the document Score > 0 indicates document is commercial (i.e. sells something)

Another small discovery is that **Google measures the commerciality of a page**. This implies that pages selling a product might not rank as well for certain queries if that query is not a commercial term. I suspect that Google would also analyze the incoming query to determine the intent:

**If a user's search term indicates they are seeking to purchase products, then results with a higher commercial score may be shown.** Conversely, if the search term only seeks information, then pages with a high commercialscore might be avoided.

**SpamWordScore** - The spamword score is represented in 7-bits, going from 0 to 127

**spambrainDomainSitechunkData** - Domain sitechunk level scores coming from spambrain

Next, we also have a **SpamWordScore** which implies that Google is looking for a specific set of words on a page.

We also seem "SpamBrainDomainSiteChunkData" which indicates that Google is using AI to process and identify spammy domains. Across all the references that we have seen, it appears as if Google is using a combination of fixed word seeking and artificial intelligence to determine spammy content.

## Normalized Site Rank

When Google is evaluating content on a website, it looks at the hostNSR which very likely stands for Host Normalized Site Rank.

`hostNsr` - Site rank computed for host-level sitechunks

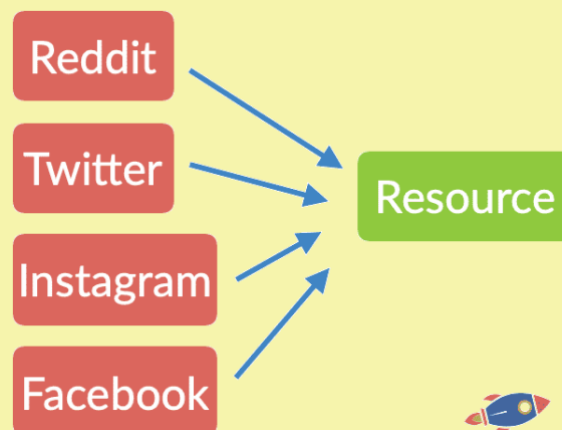
Here's how I believe it works:

### Better Rankings

Host Normalized Site Rank is... in my opinion, the most important driver of rankings after the March Core update.

**I believe it is an evaluation of how your site performs compared to industry sites.**

It is very likely derived from Chrome views and interactions on your site.



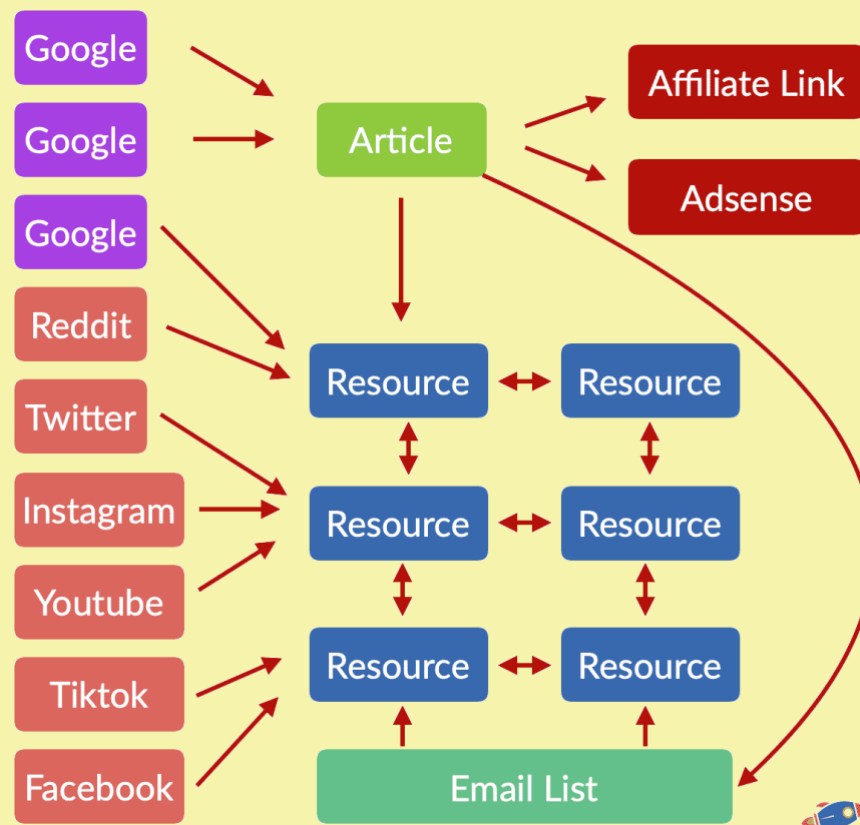
In order to rank well, **I would focus on driving significant quantities of engaged, "sticky" traffic from users to drive up my Chrome views/engagement.**

I suspect this might also be a ratio of how many pages you have on your site.  
*ie: If you have 5 pages, you might be expected to have X quantity of Chrome visitors that are engaged for Y quantity of time. (In comparison with other industry sites)*

**If you've been recently affected by one of the Google updates, I believe that you want to focus on improving your Normalized Site Rank / HostNSR.**

Here's how I have adapted to the new changes...

### New Visitor Focused Ranking Strategy



My new ranking model that **focuses on driving MORE Chrome visitors than my competitors** to improve my normalized site rank.

**Social media funnels visitors to resources on the site and a mailing list captures emails in order to encourage returning visitors.**

If you can prove to Google that your website receives more engaged Chrome visitors than your competitors, then Google will assume you have a higher quality site.

Google has mentioned NSR all over their documentation and it seems to play a key role in everything from content, to links.

## Firefly Site Signal

Next, we have our first mention of "Firefly". **This is an important signal based on the frequency of posting / how many people click on the new articles.**

Let's dive into the ramifications of this signal.

`fireflySiteSignal` - Contains Site signal information for Firefly ranking change

```
dailyClicks
dailyGoodClicks
dataTimeSec
firstBoostedTimeSec
impressionsInBoostedPeriod
latestBylineDateSec
latestFirstseenSec
numOfArticles8
numOfArticlesByPeriods - number of articles
numOfGamblingPages
numOfUrls
numOfUrlsByPeriods - number of urls sliced by 30 days
recentImpForQuotaSystem
siteFp - Hash value of the site totalImpressions
```

Firefly seems to be a major ranking factor that **measures how users react to new content being published on the website.**

Here's a quick tentative overview of the process:

1. New content published
2. Google artificially boosts the content to users to see how they interact with it.
3. Google measures interaction and calculates score for site.

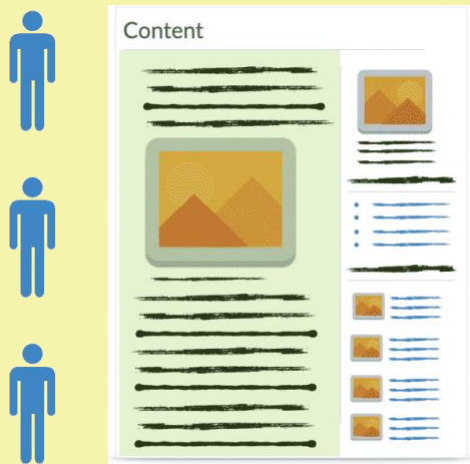
## Better Rankings

Google appears to reward "active" sites by measuring how readers interact with newly published content within a 30 day window.

For example, if you post 5 new articles, Google shows it to random people and then evaluates how your site performs.

**In order to maximize my rankings, I would prioritize publishing at least ONE good piece of content per month.**

*Because I don't have time to actively maintain some websites (but still want them to rank as well as possible), I will often prepare a series of 12 high quality articles in advance and schedule them to be posted on a monthly interval throughout the year. That way, I can 'forget' about a site for a year and still have it rank.*



## Keto Score

We jump into a section called Keto score which is **likely a power prediction made so that Google can rank content without fully processing a final link score**. This allows Google to quickly rank pages (in minutes) for breaking stories without having to go through extensive link scoring (this is done later and then the rank is re-adjusted).

`contentEffort` - LLM-based effort estimation for article pages

`deltaLinkIncoming`

`deltaLinkOutgoing`

`deltaSubchunkAdjustment` - Total deltaNSR adjustment based on subchunks

`keto` - Keto score

`linkIncoming`

`linkOutgoing`

`numOffdomainAnchors` - The total number of offdomain anchors seen by the NSR pipeline for this page

`page2vecLq`

`predictedDefaultNsr` - Predicted default NSR score computed in Goldmine via the NSR default predictor

`rhubarb` - Site-URL delta signals based quality score computed in Goldmine via the Rhubarb model

`subchunkData`

`tofu` - URL-level tofu prediction

`unversionedRhubarb` - The delta score of the URL-level quality predictor

First, when it comes to predictions, **it appears as if Google uses AI to estimate the effort that went into creating the page**. Crazy! *(This can likely be manipulated by writing about the effort spent in the introduction.)*

The rest is likely related to the power predictions based on existing signals for the website. I believe that the final "keto" score is the default predictor value that is used for all brand new content on the site.

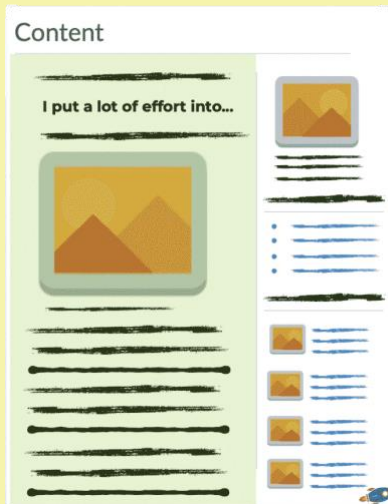
## Better Content

Google is measuring effort on a page using AI. Crazy!

They are likely using total page statistics such as word count, images, links and so forth... (longer content is likely higher effort).

And since they are using an LLM, we can assume they are also reading the page on some level.

**THEREFORE, we might be able to TRICK the "effort" LLM by including a quick mention in the introduction paragraph about how much effort we put into creating the content.**



*"I put a lot of effort into creating this content for you"*

I speculate that have such sentences in an introduction paragraph *might* skew the opinion of the AI in our favor.



## Content Score

Here we jump into another content-specific section of the API documentation.

`ugcScore`

`titlematchScore` - Titlematch score of the site, a signal that tells how well titles are matching user queries

First, the "ugcScore" most likely stands for **"User Generated Content score" which measures the quality of user generated content.** I believe that this is a critical metric to calculate because when Google wants to show discussions from Reddit, it needs to be able to distinguish between good and bad threads.

It most likely tries to identify useful information, how many people are discussing the topic and so forth.

Then we have an API entry called: **"Title match score" which measures how relevant the queries are to the titles.** You obviously want the title to be relevant to the query... however **I speculate you MIGHT be able to go too far if the titles ALWAYS match the exact query.** Further testing is required in order to confirm if this is used for promoting relevant content or if it's used to penalize overly optimized websites.

And now for the API call that seems to have been the focus of many discussions: `SmallPersonalSite`.

`smallPersonalSite` - Score of small personal site promotion

**First, `SmallPersonalSite` is a BOOST for very small sites.** *(It literally says promotion in the description)*

**When websites have very few pages, they receive a ranking boost in order to be able to compete with larger sites.** This initial boost goes away once your site reaches a certain size.

Google put this in place specifically to help the tiny mom & pop websites that are published without any regard to SEO. It helps professionals with single page websites be found, it helps the tiny 2 page businesses be found, etc.

Without it, people would struggle to rank for their own name.

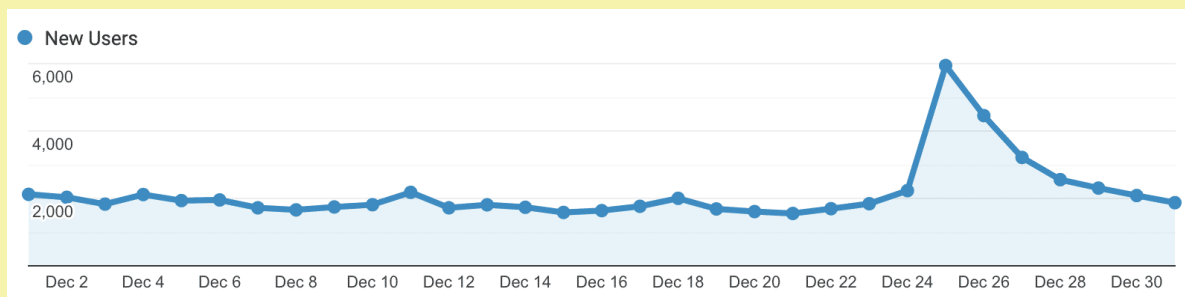
## Boost For Small Sites

I have personally experienced this MANY times before when starting sites.

**For example, sites with 5 pages or less tend to rank abnormally well for huge terms... and I noticed that when I started to scale, adding many more pages, I noticed a drop.**

If you currently have a tiny site that's ranking really well with very few pages, keep it small until you're ready to really scale. (Eventually you have to go for it)

*(One of my favorite tricks for starting a new site is to purposely keep it small for a very long time as I accumulate organic links and traffic. I performed an entire white hat case study where I launched a site with 8 articles and grew it to 109,130 monthly search views.)*



**White hat case study - 0 to 100k+ visitors**

clusterUplift

siteAutopilotScore - Aggregated value of url autopilot scores for this sitechunk

chromeInTotal - Site-level Chrome views

As we continue in the content evaluation section, we see a mention of "clusterUplift". **This is likely a boost for content within a topically aligned cluster.** Having related content is essential for ranking on Google these days and this is just another piece of evidence that explains why this is the case.

Next, we have "siteAutopilotScore". This might be a **measure of how much a site is created by automated processes:** Auto-generated content from RSS feeds.

Finally, we have "chromeInTotal" which is very obviously the total Chrome views a site has received. **I suspect that that the quantity of Chrome views for a site might be a ranking signal...** however it certainly wouldn't protect sites from being penalize.

## Chard Score

Chard seems to be related to Chrome interactions.

chardVariance

chardScoreVariance - Site-level Chard Variance for all pages of a site

chardScoreEncoded - Site-level Chard (encoded as an int)

Even though there is no direct mention of what "chard" stands for, all the descriptions and indications **hint that it is related to user behavior and Chrome.**

With that in mind, my best guess for what chard stands for is: **Chrome Average Duration.**

Even if it stands for something slightly different, one way or another, **we know that they measure how long people stay on the site using Chrome.** This seems to be a VERY important signal.

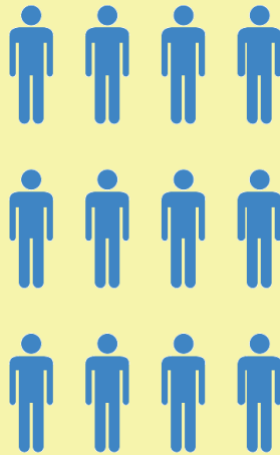
## Site Boost

As stated before, **one of the most important things you can do to increase a site's ranking is to get real traffic, specifically Chrome views, to visit all the pages on your site.**

**Google not only measures the individual views but also the quality of the views.**

How long users stay on the site using Chrome seems to play a role.

*(One of my more recent tricks involves creating a resource section on my website in order to get my visitors trapped in an endless loop of content. I try to compile high value material that people can download and put it all in one place. To my surprise, we've had resources go viral on Pinterest, driving substantial quantities of high quality Chrome visitors to the site.)*



## NSR Override

As I was browsing the site evaluation section, I stumbled upon this API reference that at first glance seemed harmless... however as I discovered the impact of NSR on search results, it became quite apparent that it was immensely powerful.

`nsrOverrideBid` - This signal is used to unconditionally override NSR as a bid in Q\*

Here we have "nsrOverrideBid" which means that it can be used to OVERRIDE the Normalized Site Rank assigned to a site. Essentially, what this means is that they can manually boost a site's rankings if they want... *(in case of an emergency, of course)*

### Secret Google Switch

This is one of those "uh oh" moments as it seems **there's a cleverly disguised variable they can manipulate to override the NSR** (which is one of the primary driving forces in ranking).

This means they have the power to manually override a site that has a low Normalized Site Rank...

**I believe THEY can change the rankings of any website as much as they want with this value.**

They could manually penalize a site (without having an obvious penalty) or they can promote a site (reverse a penalty) with this metric.

## Content Score

In this section, we dive deeper into ranking (calculating a score) for content. While we've already covered many of the elements (ingredients) that Google takes into consideration when evaluating content, this section seems to be more focused on scoring.

`articleScore` - Score from article classification of the site

`site2vecEmbedding` - Site2vec embeddings

First, obviously, we have the **article score that is determined by the classification of the site**. This seems to indicate that **if a site is related to the topic, you automatically get a boost to your score even before Google begins to analyze the content**.

Next, we have Site2VecEmbedding which is related to the topical alignment of the site. This, once again, indicates that **Google is looking at all the entities on a site in order to determine the relevance of a website**.

`isElectionAuthority` - Bit to determine whether the site has the election authority signal, as computed by go/election-authority

I thought some people might be interested in knowing that Google can manually flag sites that are authorities in elections (likely government sites). In an effort to fight misinformation, sites like Whitehouse.gov or local city election sites might be flagged so they appear higher when people are searching for recent election related information.

**clutterScore** - Delta site-level signal in  $Q^*$  penalizing sites with a large number of distracting/annoying resources loaded by the site

Finally, we have a clutter score. The "ClutterScore" is a score based on the layout clutter, used extensively on mobile when you have too many ads / pop-ups.

**If you're running mobile ads, make sure that the ads aren't popping up too much on mobile devices as it will increase the clutter score (and this will affect your rankings).**

While this applies to desktop as well, the clutterscore is a score for intrusive elements on a page.

Clean layouts will typically rank better.

## The Role Of Entities In Search

As words can have multiple meanings, computers prefer to use entities to classify and understand content. In this section, we'll discover the emphasis that Google puts on entities.

**entities** - A list of entities detected on Document.text  
**entityRelations** - Placeholder. Relationship among Document.entities

First, we have a raw list of the entities within a document. Entities play a central role in what Google retrieves whenever there is a search query.

The next thing is looking at the relationship between entities. This can help discover related terms, related topics and help provide context for a document.

Multiple entities can be identified on a document or query. Each entity can be mentioned several times in different positions on the document or query. This message describes a single mention of the entity. Note that a mention can be either explicit or implicit mentions. All explicit mentions refer to exact range in the document where the entity occurred, but implicit mentions may or may not have corresponding range. Next available tag number: 40

Not only does Google cares about the entities mentioned in documents, **they also note the frequency and importance of each entity. Entities that are mentioned more frequently are going to be seen as central to the document and more important.**

Another interesting point is that they share that "implicit" entities are calculated when you mention certain entities.

**isImplicit** - True if the entity is mentioned implicitly

For example, if you mention the word "gym" in the content, then a related entity can be "sports". **This isImplicit is VERY useful as it allows documents to be found even if they don't mention the exact entity.**

You can try it by searching for: "*motion picture about a tsunami*" on Google. You'll likely see results about movies featuring a tsunami... even if the words "*motion picture*" does NOT appear on the page.

This is because Google knows that the motion picture entity is related to the movie entity. (And this is why you don't need to mention every single entity variation in order to rank.)

**confidenceScore** - A probabilistic score describing how certain the annotator is that this exact

Of course, Google will not show the page just because it features an entity. The confidenceScore indicates that **Google is seeking the most important entities in a document.**



SalientTermSet is a collection of terms (unigrams and bigrams) with associated weights that can describe something. The "salient terms"

`docData` - `doc_data` contain additional salient-term-set-level data

`salientTerm` - `salient_term` is the list of terms that are good descriptors, sorted in decreasing order of weight

`version` - `version` is the Salient Terms version used to create the SalientTermSet

Continued here, we see they list all the most important entities in reverse order: From most important to least important.

## Related Words

A list of entities that are latent given this entity. For example, "Lionel Messi" can have the latent entity "FC Barcelona". See [go/refx-latent-entities](#) for detailed description.

`latentEntity` - Latent entities with associated metadata including source of the relationship

And as we previously covered, they list latent entities are going to be the related words associated with the main entity. We can think of these as related words to the main subject.

**This is useful because Google already knows the related entities to the main entity... and therefore Google can seek pages that contain those latent entities.**

## Better On-Page Ranking

As we have seen, Google likely already knows about the main target entities and the latent entities...

And it goes without saying that pages contain BOTH the exact entity and related entities are likely going to be more relevant for the search query.

**Therefore, it might be a good idea to include many latent entities in *addition* to the exact entity in order to maximize the chances of ranking.**

For example, if you have a page on "dog food", I would include mentions of "pets", "nutrition", "protein", "puppies" and so forth.

The screenshot shows two sections from Google's search results. The top section, 'Word Recommendations', has tabs for 'All', 'Detected', 'Missing from target', and 'Word found but not as entity'. It displays a large number of colored tags representing various entities and their frequencies, such as 'seo checklist (10)', 'page (10)', 'content (10)', 'pages (10)', 'keywords (10)', 'seo (10)', 'people (10)', 'example (10)', 'information (10)', 'one (10)', 'site (9)', 'google (9)', 'links (9)', 'traffic (9)', 'tools (9)', 'search engines (9)', 'data (9)', 'way (9)', 'results (9)', 'url (9)', 'ways (9)', 'urls (9)', 'websites (9)', 'text (9)', 'keyword (8)', 'keyword research (8)', 'tool (8)', 'checklist (8)', 'thing (8)', 'rankings (8)', 'list (8)', 'images (8)', 'guide (8)', 'title (8)', 'words (8)', 'tags (8)', 'things (8)', 'search engine (8)', 'factors (8)', 'google analytics (7)', 'sites (7)', 'wordpress (7)', 'questions (7)', 'title tag (7)', 'part (7)', 'something (7)', 'link (7)', 'users (7)', 'fact (7)', 'idea (7)', 'anything (7)', 'lot (7)', 'top (7)', 'factor (7)', 'topic (7)', 'google search console (7)', 'all (7)', 'conversions (7)', 'experience (7)', 'target keyword (6)', 'search (6)', 'step (6)', 'everything (6)', 'tips (6)', 'visitors (6)', 'video (6)', 'performance (6)', 'sitemaps (6)', 'blog posts (6)', 'web pages (6)', 'seo plugin (6)', 'steps (6)', 'search results (6)', 'article (6)', 'seo strategy (6)', 'result (6)', 'value (6)', 'user experience (6)', 'competitors (6)', and 'seo efforts (6)'. A 'Show More' link is at the bottom. The bottom section, 'Highly Related Words', has tabs for 'All' and 'Missing from target'. It displays a smaller set of red tags: 'seo', 'checklist', 'tips', 'tricks', 'strategies', 'google', 'algorithm', 'updates', 'webmaster', 'guidelines', 'serp', 'ctr', 'webpage', 'content', 'rankings', 'web traffic', 'web design', 'web development', 'search engine marketing', and 'search engine optimization'.

### Top Entities (Analyzed With Google's NLP Model)

<https://on-page.ai>

(When optimizing for Google rankings, I like to make sure to include as many recommended related entities as possible within my content, **aiming for a large diversity**. Once I'm confident I have a wide diversity of entities, **I will then focus on increasing the frequency of the top entities** within my text. It's not uncommon for me to repeat the top entities 3-5 times each... sometimes substantially more depending on the context.

**Finally, I make sure to include highly related entities that my competitors might not be using in order to differentiate myself from the pack. )**

## Entity Scoring

So can you just throw a bunch of entities on a page and hope for the best? Well... kinda (it works). However there's a bit more refinement to be done in order to get the absolute best rankings.

**idf** - idf of the original\_term

**label** - label can be two things depending on where this message is

**originalTerm** - original\_term are the different ways we found this normalized term in the signals

**salience** - salience is the importance of the term as a descriptor in [0, 1] (the higher the more important)

**signalTerm** - signal\_term contains extra signal-specific (e.g., body, anchors, clicks) data for this term

**virtualTf** - virtual\_tf is the accumulated corrected term frequency from all the signals

**weight** - weight is the importance of the term as a descriptor in [0, 100] (the higher the more important)

Here we see a mention of IDF which stands for "Inverse Document Frequency". **This is indicate that Google is measuring original content.** (Mention things/entities no one else is discussing). To do so, they analyze how rare an entity is with regards to a corpus of text... if it's something that isn't commonly mentioned, then you're likely writing about something original.

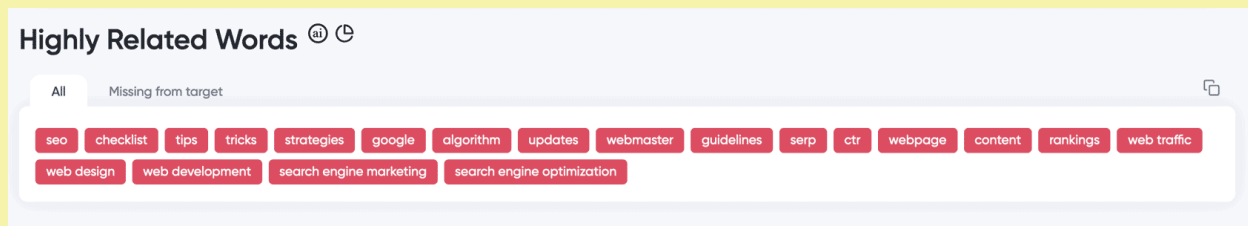
Next, we have mentions of "Salience" which is a measure importance. Google is always asking itself: What is the most important entity in the document.

## Better On-Page Ranking

Create original content and use entities that your competitors AREN'T using.

Original content is rewarded!

In addition, Google seeks out the most important terms on a page and evaluates the salience (importance) so make sure the focus of your article is obvious.



*Relevant entities from our model that competitors might or might not be using*  
<https://on-page.ai>

One of my big gripes against many other optimization tools is that they only return the words that competitors are using... and therefore, at best, you're going to be a worst copy of the top results. **You don't want to be a clone of the current ranking results.**

**Instead, I strongly believe in matching and then going above and beyond the #1 result by including entities your competitors might not be using.**

*(I personally try to include ALL the "Highly Related Entities" as it helps set me apart from my competitors. I will also focus on providing factual and quantifiable data that isn't presented anywhere else on the web.*

*For example, while many of my competitors mentioned the battery life for a product, I added EXTRA information on charging, including the 0% to 50% charging time, the 0% to 80% charging time and the 0 to 100% charge time.*

*The end result was that I wasn't just quoting the manufacturers' battery life... I was going above and beyond, mentioning the battery life AND also including charging time.)*

## Categories

At the higher level, beyond entities is categorization. Which categories does the content fall into?

We still allow legacy use case to exist (no forced migration), but we will not accept any new usage of WMA, incl. from existing clients. UDR has the same features and can be used similarly: - To consume the topical entities

`categoryConfidenceE2` - The confidence of the category

`categoryEncodedMid` - See [go/category-annotations-api](https://cloud.google.com/natural-language/docs/categories-api) about the story behind various types of category annotations that are provided using the catmid token and `category_encoded_mid` fields below

`confidenceE2` - The confidence scores of all entities in the `encoded_mid` array

`topicalityE2` - The topicality scores of all entities in the `encoded_mid` array

According to the documentation, it appears as if Google classifies content into categories.

It appears as if they have recently upgraded categorization and currently have two systems running (and old and one new). **I speculate that the old one might be the list of categories that they published at <https://cloud.google.com/natural-language/docs/categories> as it was a comprehensive list of NLP categories.**

## Site Content Categories

By entering a large sample of links from a site and using Google's NLP classification model, I'm able to get an idea of how the content on the site might be categorized.

I believe Google was using these categories in the past and has now moved on to something new...

However this is the best we have at the moment so it's what I'm personally using.

**Predictive Link Building** DOMAINS ☒ LINKS Clear Load

Enter the domains you want to check (500 max, 1 per line):

```
https://www.waytoosocial.com/make-girl-fall-love/
https://www.waytoosocial.com/how-to-make-a-girl-smile/
https://www.waytoosocial.com/how-to-talk-to-girls/
https://www.waytoosocial.com/top-100-things-that-attract-women-to-men/
https://www.waytoosocial.com/signs-a-girl-likes-you/
https://www.waytoosocial.com/best-dating-sites/
https://www.waytoosocial.com/how-to-approach-girls/
https://www.waytoosocial.com/how-to-flirt-with-a-woman/
https://www.waytoosocial.com/pick-up-artist-guide-from-geek-to-pua/
https://www.waytoosocial.com/what-to-text-a-girl-you-like/
https://www.waytoosocial.com/eharmony-review/
```

Submit reset Import links

**Link Relevance Report** Export /People & Society

We highly recommend exporting the full report

<input type="radio"/>	#	Link	Category	Link Relevance
<input type="radio"/>	1	https://www.waytoosocial.com/make-girl-fall-love/	/People & Society/Family & Relationships	1/1 match
<input type="radio"/>	2	https://www.waytoosocial.com/how-to-talk-to-girls/	/People & Society/Family & Relationships	1/1 match
<input type="radio"/>	3	https://www.waytoosocial.com/top-100-things-that-attract-women-to-men/	/People & Society/Family & Relationships	1/1 match
<input type="radio"/>	4	https://www.waytoosocial.com/how-to-approach-girls/	/People & Society/Family & Relationships	1/1 match
<input type="radio"/>	5	https://www.waytoosocial.com/how-to-flirt-with-a-woman/	/People & Society/Family & Relationships	1/1 match
<input type="radio"/>	6	https://www.waytoosocial.com/what-to-text-a-girl-you-like/	/Online Communities /People & Society/Family & Relationships	1/1 match
<input type="radio"/>	7	https://www.waytoosocial.com/what-women-want-from-men/	/People & Society	1/1 match
<input type="radio"/>	8	https://www.waytoosocial.com/honesty-the-basis-for-a-good-relationship/	/People & Society/Family & Relationships	1/1 match

**Categorization (Analyzed With Google's NLP Classification Model)**

<https://on-page.ai>

**By entering a large sample of links from a site and using Google's NLP classification model, I'm able to get an idea of how the content on the site might be categorized.**

I believe Google was using these categories in the past and has now moved on to something new...

(However this is the best we have at the moment so it's what I'm personally using.)

For example, in the screenshot above, we see that the content on the site that falls into the /people & society/Family & Relationships/ category.

This can be useful when trying to determine the topical alignment of the website and consequently, which content will rank the best.

## Click Signals

One of Google's biggest advantages in search is its unique access to user data. Google monitors how users interact with content to make informed judgments about that content.

`absoluteImpressions` - Thus far this field is only used for host level unsquashed impressions  
`badClicks`  
`clicks`  
`goodClicks`  
`impressions`  
`lastLongestClicks`  
`unicornClicks` - The subset of clicks that are associated with an event from a Unicorn user  
`unsquashedClicks` - This is not being populated for the current format - instead two instances of `CrapsClickSignals` (squashed/unsquashed) are used  
`unsquashedImpressions` - This is not being populated for the current format - instead two instances of `CrapsClickSignals` (squashed/unsquashed) are used  
`unsquashedLastLongestClicks`

This section of API references provides us with many clues about how they measure user clickthrough rate.

First, the obvious: **They measure clickthrough rate (impressions / clicks) for content.**

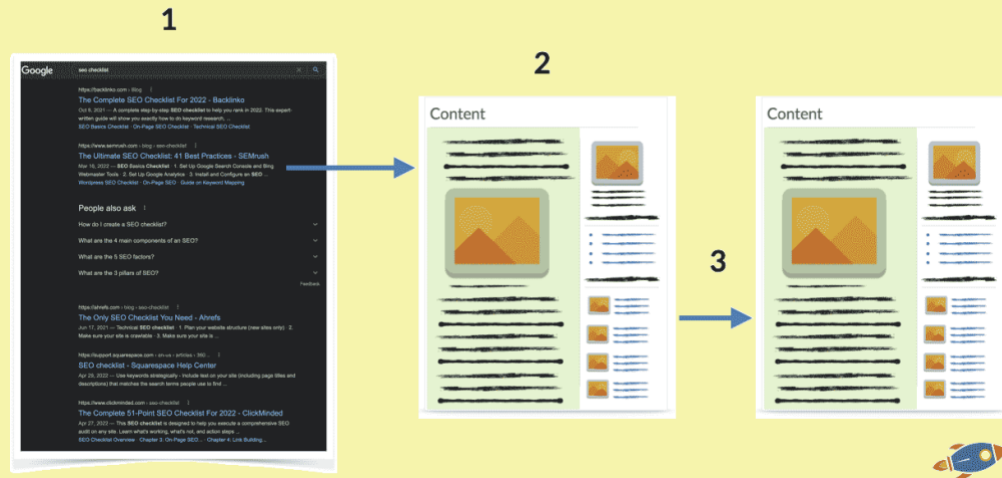
Next, we also note that they have a "bad clicks" (likely people returning to search)

And finally, **unicorn clicks which is clicks from abnormal users trying to manipulate rankings.** (ie: a user clicks on 1000 results a day)



## User Click Data

Google measures the interaction of users with the search engine results. Good catchy titles & good meta descriptions will help increase the clicks to a page.



We want to **avoid "bad clicks"** at all cost... which is when someone goes BACK to Google to search for the same term after they landed on your page.



We also want to **avoid UNICORNCCLICKS** which are clicks from a single user that performs an abnormal quantity of searches / clicks in a day.

*(While I do NOT recommend this, some people have been purchasing 20-50 bulk Android phone lots on ebay and combining them with cheap sim cards. They then create full Google profiles, download Chrome for Android, a GPS spoofer and enable remote access so they can control the phones from one central location.*

*It's a relatively high initial investment however the end result is the ability to a small personal army of Android users. Once again, I do NOT recommend this as there is a steep learning curve, it's costly, you have to deal with power issues, phone issues and much more... and in the best case situation, you end up with limited quantity of users.)*

## Stored Click Data

We just looked at some of the API references for the clicks themselves however there's another section with the accumulated click data.

`averageChanceTime` - Weighted averaged timestamps of the decayed chances  
`chances` - Numbers below are all total in the decayed manner  
`clicksBad`  
`clicksGood`  
`clicksImage`  
`clicksTotal`  
`clicksUnclassified`  
`coverageTimestamp` - Epoch seconds at which this weighted coverage data was calculated  
`ctrWeightedImpressions`  
`dwells` - Dwells from KnowledgePanel and WebAnswers  
`firstBaseCoverageTimestamp` - Epoch seconds at which this url first gets coverage in BASE  
`firstCoveragePagerankNs` - The pagerank when the url was serving for the first time  
`firstCoverageTimestamp` - Epoch seconds at which this url first gets coverage data  
`firstseen`  
`impressions`  
`intervalData` - Interval Data to track the average time between clicks\_total, clicks\_good, and ctr\_weighted\_impression

**This is SUPER important because it is the ACCUMULATED click data for a website that is likely responsible for penalties.** If you've been experienced a recent demotion, it's likely because the data here for your site is sub-par.

Most of it is what we'd expect, good clicks, bad clicks, etc...

However small surprises are:

1. **Image clicks count.** I imagine that these are images from within the normal search engine results however could it also count information from the separate image search? We'll have to test this.

2. **Dwells count.** This means that **even if you don't get a click, if people hover the search snippet, then you get rewarded.** This is a nice addition from Google to reward high ranking sites that currently have the search snippet.

`lastDwellDateInDays` - Indicates the date when this document received the last KnowledgePanel or WebAnswer dwell  
`lastGoodClickDateInDays` - Indicates the date when this document received the last good click  
`lastImpressionDateInDays` - Indicates the date when this document received the last impression  
`lastLuDwellDateInDays` - Indicates the date when this document received the last LocalUniversal dwell  
`lastPseudoImpressionsDateInDays` - Indicates the date when this document received the last pseudo-impression  
`luDwells` - Dwells from LocalUniversal  
`repid` - Repid in Alexandria pipeline  
`totalChances` - Total number of chances on this urls (not decayed)  
`url`  
`urlfp`

Another interesting part is that **they measure the last time a document received a click.** Perhaps this measures the popularity of content within your site as we can assume that if content receive no clicks, then it might not be as relevant.

**The next important part is the mention of "LocalUniversal" dwells.** While we have no additional on what the "LocalUniversal" source is... perhaps this could be an indication they are measuring user data from OTHER sources such as Android.

## Maintaining A High Quality Site

Google measures the clicks sitewide... so it might be interesting to test how Google would react if you had Chrome clicks going to all pages on the site.

**This MIGHT also support the idea of removing old content that gets no clicks/visits... as those pages are likely to have no Chrome views.**

*(One of the **first things I do when I'm recovering a website is look for thin/duplicate and low quality pages.** I will usually use Screaming Frog Spyder to get a quick overview of the site, identifying potentially problematic pages.*

*In WordPress, this often means removing the tag pages, very thin category pages with only 1-2 items, date archives and more. I'll use a plugin such as Yoast to disable these sections.*

*I then search for **all the blog posts with sub-350 words as those are usually (but not always) thin/low quality content.** Finally, I'll look up extremely old content (7 years+) that might be **completely out of date.** This content might need an update or might no longer be relevant.*

*Before I remove content, **I always do a quick backlink check with Ahrefs or Semrush to make sure I'm not accidentally removing any links that might be pointing to the content.** I warn against removing content if you aren't sure what you're doing and recommend working with an experienced SEO professional before making drastic changes. )*

## Mobile & Usage Data

This is yet another section clicks and user engagement on a website... with all this data being used to boost search results that perform well and demote content that doesn't.

```
badClicks
clicks
country - The two-letter uppercase country slice of the CrapsData
device - The device interface and os slice of the CrapsData
features - Contains CrapsClickSignals for specific features
goodClicks
impressions - These fields may become legacy fields
language - The language slice of the CrapsData
lastLongestClicks - The number of clicks that were last and longest in related user queries
mobileData - DO NOT USE: Use the above mobile_signals fields instead
mobileSignals - The portion of this CrapsData aggregated on data from tier 1/2 mobile interfaces in QSessions
packedIpAddress - Contains a packed string in network byte order, as expected by CrapsIpPrior
patternLevel - Level of pattern. More general patterns get higher values
patternSccStats - For pattern data, this will contain stats of the SCC's of the individual urls contributing to the pattern
```

The first thing we notice is the extensive mention of CRAPSdata. This data, gathered via Chrome and mobile devices, is apparently very important.

Unfortunately, Google doesn't share exactly what CRAPS stands for. However, an educated guess might be:

*"Chrome Requests Aggregated Page Statistics"*

The bottom line is: **If users visit and like your site, you will do well on Google.**

query  
sliceTag - This field can be used by the craps pipeline to slice up signals by various attributes such as device type, country, locale etc  
squashed - Not used yet  
unscaledIpPriorBadFraction - Used to assign a prior based on IP address  
unsquashed - We will start using this one for the retuning rollout  
unsquashedMobileSignals  
url  
voterTokenCount - The number of distinct voter tokens

As we know by now, they measure the query and how users react to it. The idea is that every click counts as a vote . The more votes a document receives, the better it likely is.

## NavBoost: Click Testing

In this section, they discuss how Navboost is used in combination with clicks to help improve Google's search results.

The information representing one navboost query for the dataset source\_url

```
impCount - imp_count stores an estimate of the number of impressions for this tuple
lccCount - lcc_count stores an estimate of the number of long clicks for this tuple
query - The query string
queryCount - The query_count stores the counts on this query
queryDocCount - The query_doc_count stores the number of long-clicks on this pair
```

Because we know that Navboost is a method of increasing the rank of pages... and that Google is trying to measure clicks, we can deduce that **Google temporarily increases the ranking of pages in order to measure clicks from it.**

The more "longclicks" a page receives when it is boosted, the better it is. This could help surface pages that otherwise wouldn't be found.

Contains high level search query statistics of the document

aggregatedQueryCount - Total query count for the document from all the query terms.  
Can be used to estimate the popularity of the document

Finally, we have a mention of "aggregatedQueryCount" at the end of the documentation. This might be a ranking factor on it's own... how many times this document appeared in search queries.

More popular pages might appear in more queries, especially if this page covers many topics (might help pages that rank for many different topics)



## Mobile Visits = Site Boost

**Receiving mobile visitors produces an additional BOOST for the entire website... and will likely be applied in that region.**

For example, if you receive many mobile visitors from Canada, then you're likely to receive a boost for google.ca searches.

For example, if you were to run a localized advertisement campaign (ie: A local TV ad that has people searching for your website), then this should translate into higher rankings.

*(Unrelated to mobile traffic... however I also noticed that the hosting location of the server seemed to have an impact on rankings. When I just happened to have a hosting server in Germany, I received more European visitors. After changing the location to the USA, I slowly noticed a shift. Perhaps this is just due to site speed or perhaps it can have a direct impact.)*

## Mobile Penalties

Speaking of mobile, Google an entire section dedicated to mobile related verifications and penalties.

`adsDensityInterstitialViolationStrength` - Indicates if the page is violating mobile ads density interstitial policy and the violation strength

`isSmartphoneOptimized` - Indicates if the page is rendered in a friendly manner on smartphones

`violatesMobileInterstitialPolicy` - Indicates if the page is violating mobile interstitial policy and should be demoted

Bottom line is, in order to rank, it is very important that the page be "Smart Phone optimized" which essentially means "Mobile-ready" in Google speak. You can (and probably should) check the Google search console for your website to see if any pages have issues. When you're in the search console, it's within the "Page Experience" tab.

In addition, too much advertisements on smart phones will likely result in a demotion/penalty.

## Snippet Score

Interestingly, we have a section dedicated to the search snippet.

Query related features used in snippets scoring. Next ID: 7

```
experimentalQueryTitleScore  
passageembedScore  
queryHasPassageembedEmbeddings  
queryScore  
radishScore
```

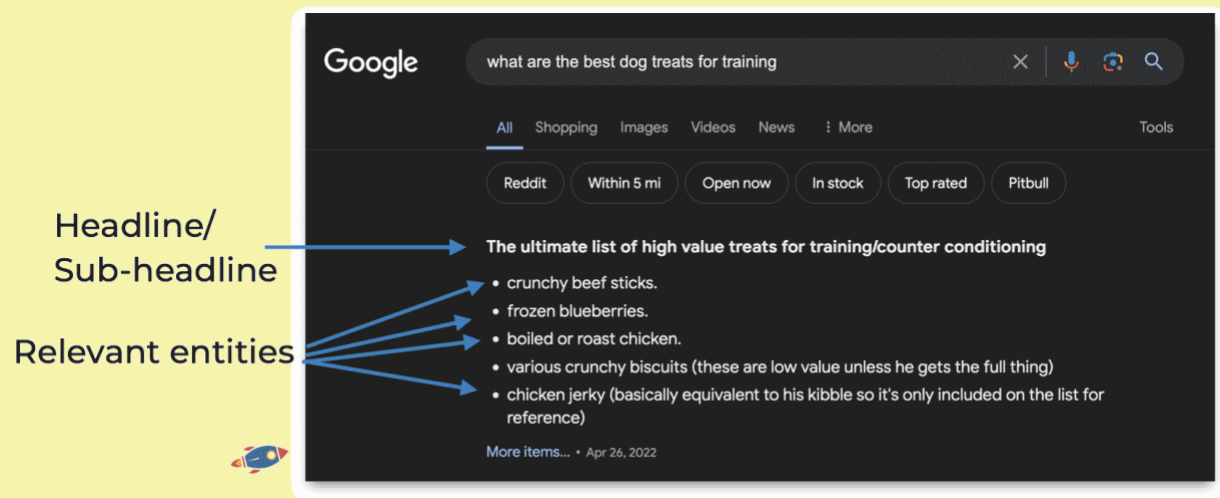
The algorithm surrounding the search snippet seems to fairly straightforward:

- There's a title score that measures the relationship between the title and the query
- There's an embed score for the potential passage and an embed score for the query

And I suspect the **radishScore** at the end might be putting everything together. Perhaps it stands for something like: **R**elevance, **A**uthority, **D**ensity, **I**ntent, **S**pecificity, **H**elpfulness

In which, Google would seek relevant content, from authoritative websites, with a high density of entities, that matches the user intent and provides specific data that is helpful to the user.

## Easier Search Snippet



To increase the likelihood of ranking for the search snippet:

1. Have a **relevant title/sub-headline** above the specific passage addressing the query.
2. Use a **high concentration of relevant entities** to increase your embedding score.
3. Include a **partial or exact query** for which you want to have the search snippet.

*(I personally like to use these two exact format to maximize my chances of getting the search snippet:*

### Format #1:

```
<h2>relevant title addressing the search query</h2>
```

```
<ol>
```

```
<li> List of relevant entities </li>
```

```
<li> List of relevant entities 2 </li>
```

```
<li> List of relevant entities 3 </li>
```

```
</ol>
```

```

```

**Format #2:**

*<h2>relevant title addressing the search query</h2>*

*<p>Passage answering the search query, stuffed with related entities</p>*

*<image src="relevant-image with query filename.jpg" alt="relevant query">*

*I call these "Google Food" because I'm creating a sandwich from the search query, related entities and related image. Google eats it up!)*

## Retrieving Docs

And finally, we have document retrieval which is obviously very important for search.

`latestPageUpdateDate` - The syntactic date of a dataset document that reflects the publication date of the content  
`navboostQuery` - A sequence of Navboost queries for the dataset source\_url  
`pagerank` - The page rank of the document  
`pagerankNs` - The production pagerank value of the document

As we have previously seen, factors that determine which content Google will retrieve when there's a query:

- **Last page update.** Google seems to want to surface freshly updated, new content.
- **Queries with a Navboost**
- And of course, **PagerankNS** (the new pagerank).

`pagerank` - The page rank of the document  
`pagerankNs` - The production pagerank value of the document  
`petacatInfo` - Petacat classifications for the web document  
`salientTerms` - A set of salient terms extracted from the document  
`scholarInfo` - Science per-doc data for inclusion in websearch  
`sporeGraphMid` - A set of entities from WebRef annotations that are in SPORE\_GRAPH  
`title` - The title of the document  
`topEntity` - A set of top entities from WebrefAnnotation, top is defined by topicality score  
`url` - The url of the document  
`webrefEntity` - A set of entities copied from WebRefEntities on cDoc

Google lists all the terms that a page should rank for and the category in which it belongs.

This is likely what they use to retrieve documents before sorting them to rank them.

## Ranking Essentials

When Google retrieves documents, they look at:

**Freshness** (the last time it was updated)

**Navboost factor** (the boosts from other signals)

**PagerankNS** (the power going to the page)

The **classification of the document**

The **most important terms/entities** of the document.

It appears as if MOST of quality signals end up going into the Navboost.

*(When I'm trying to rank a site, I like to start by focusing on ONE page. I'll determine the specific changes I need to make in order to retain users, provide a good user experience and attract links. Sometimes this involves a complete re-write, sometimes the solution is to add imagery, change the theme, add trust elements, etc.*

***Once I identify and confirm the changes I need to make to one page, I replicate those changes throughout the site, improving all pages. If I determine a page doesn't align with a website, I'll put it in 'draft mode', preventing it from showing up in search for the time.***

*This, in turn, improves the overall site quality, provides a freshness boost and improves the topical alignment of the site. I monitor existing traffic for improvements in user engagement.)*

And that's a look at all the most important API references that can help us learn about the Google algorithm.

## Discussion

### (Thoughts & Surprises)

It's incredible to finally see the inner workings of the Google algorithm. It was also nice to confirm some suspicions, bust some myths and gain a new understanding of how search works. **I personally believe that there is a clear, achievable, and repeatable way to rank websites that can drive profits for years to come.**

In that sense, there is no better time to do SEO.

My biggest surprises was the **emphasis that was put on anchor text, click data and topical authority**. While I knew that all of those were involved in the Google algorithm, the pages and pages of data covering each of these highlighted their importance.

And yes, beyond links, **I believe that NSR, normalized site rank, is likely the strongest ranking factor as it affects everything related to the website**. From how powerful the links are... to how well the content will rank.

I was pleasantly surprised to see such a huge emphasis on topical authority, it really is a big deal. The site's focus is measured and related content does get a boost.

I really enjoyed (and appreciate) the **special effort put forward by the search team to shield people from negative SEO**, remove personal information from the web and help small tiny website. I also really appreciate that they look for original content.

Conversely, **I was shocked to see the quantities of penalties that can stack on top of each other when sites didn't follow Google's ideals**. Double, triple penalties aren't cool.



I understand the reasons behind not wanting to share that Pagerank uses seed sites... and I certainly understand why they wouldn't want the SEO community to know about how much weight Chrome visits have on a site.

**Fortunately, many of the core signals are relatively difficult to fake.**

**For anyone that has read through this paper, you now have an incredible edge when it come to ranking online.**

From knowing exactly how to build a site with topical content, to optimizing individual pages of content with entities, to publishing rate (freshness) to link building from high quality sources with relevant anchor text...

You are armed with the ultimate recipe for Google search. Speaking of recipes...

# My New Ranking Strategy

## (Full Step by Step Process)

Putting it all together, here's my new ranking process combining all this newfound knowledge.

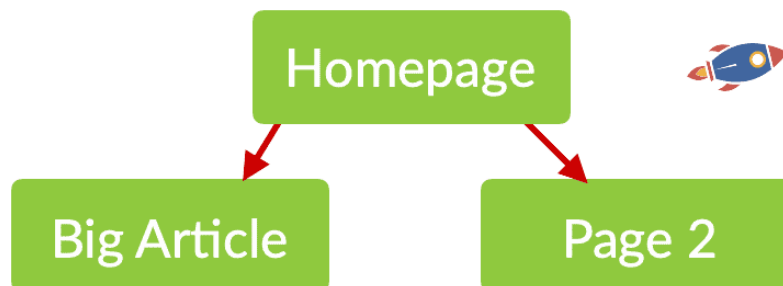
This is **designed to take advantage of all the newly discovered "boosts" that can help a website dominate the search results while following all the rules.** I distilled it down to be as concise as possible in order to make it simple and easy to understand.

*Please note that this is my own personal ranking formula. I have no control over you, your site, nor Google, you are responsible for any changes you make to your website.*

## 1. Starting A New Site

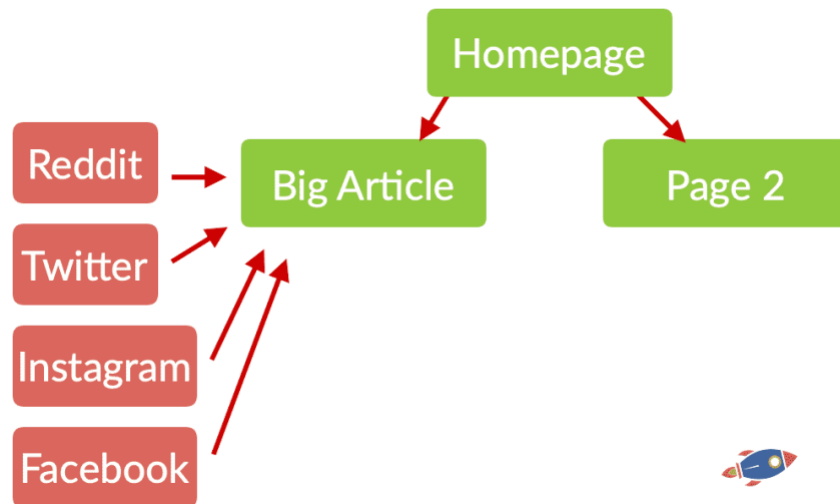
When I'm starting a new site, I'm **purposely going to keep it SMALL at first in order to take advantage of the "Smallpersonalsite" ranking boost.**

I will have a homepage and 2-3 other pages so I can quickly gain traction in the search engine results.



Taking into account that there might be a sandbox period dependant on hostAge, I will be patient even if I don't see instant ranking results.

I will have **ONE flagship article** that is immensely valuable. I will aim to get it shared across social media and get natural visitors to my site.



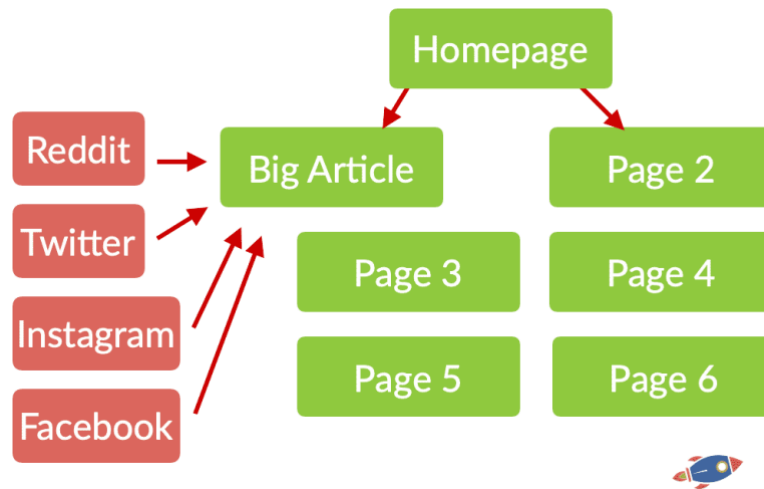
This will, in turn,

1. Increase Chrome views
2. Improve my overall site score
3. Improve my click metrics / visit duration

Essentially, setting up my site for success as all the Site Quality metrics will be positive.

## 2. Topically Aligned Content

Once I'm confident my site quality is set, I'm getting visitors and a I have a few links, I will leap out of the "small" site size and start adding more content.

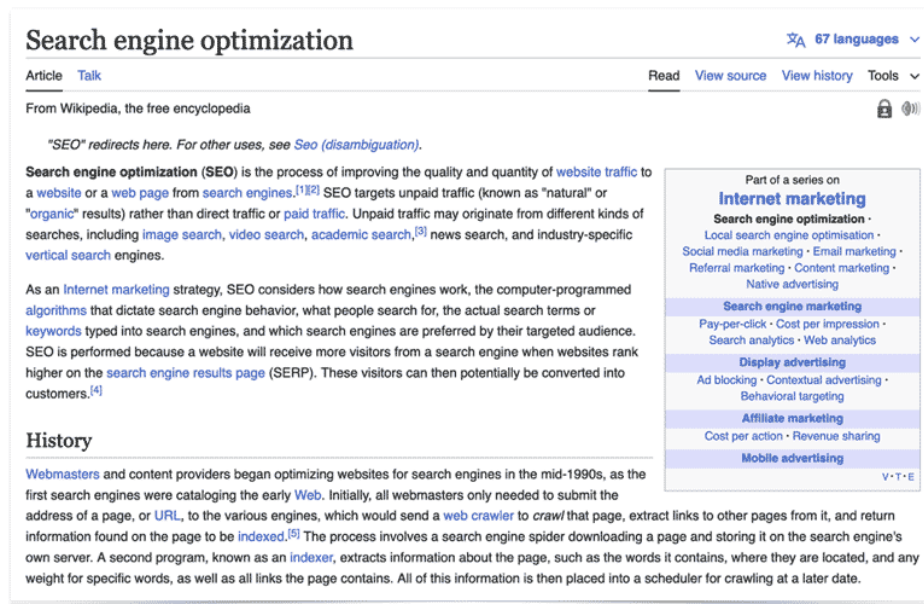


I will not be discouraged when I notice a small drop in rankings as I make this transition because I know it's temporary as I will eventually surpass my small size.

**The content I'm adding will all be HIGHLY related and topically aligned.** I will aim to keep the focus of my site very narrow to take advantage of the second boost.

### 3. Topical Authority

I will focused on writing on content that falls into a specific category and is semantically related.



**Search engine optimization** 67 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

*"SEO" redirects here. For other uses, see [Seo \(disambiguation\)](#).*

**Search engine optimization (SEO)** is the process of improving the quality and quantity of [website traffic](#) to a [website](#) or a [web page](#) from [search engines](#).<sup>[1][2]</sup> SEO targets unpaid traffic (known as "natural" or "organic" results) rather than direct traffic or [paid traffic](#). Unpaid traffic may originate from different kinds of searches, including [image search](#), [video search](#), [academic search](#),<sup>[3]</sup> news search, and industry-specific vertical search engines.

As an [Internet marketing](#) strategy, SEO considers how search engines work, the computer-programmed [algorithms](#) that dictate search engine behavior, what people search for, the actual search terms or [keywords](#) typed into search engines, and which search engines are preferred by their targeted audience. SEO is performed because a website will receive more visitors from a search engine when websites rank higher on the [search engine results page](#) (SERP). These visitors can then potentially be converted into customers.<sup>[4]</sup>

**History**

[Webmasters](#) and content providers began optimizing websites for search engines in the mid-1990s, as the first search engines were cataloging the early [Web](#). Initially, all webmasters only needed to submit the address of a page, or [URL](#), to the various engines, which would send a [web crawler](#) to *crawl* that page, extract links to other pages from it, and return information found on the page to be [indexed](#).<sup>[5]</sup> The process involves a search engine spider downloading a page and storing it on the search engine's own server. A second program, known as an [indexer](#), extracts information about the page, such as the words it contains, where they are located, and any weight for specific words, as well as all links the page contains. All of this information is then placed into a scheduler for crawling at a later date.

Part of a series on  
**Internet marketing**  
**Search engine optimization** ·  
 Local search engine optimisation ·  
 Social media marketing · Email marketing ·  
 Referral marketing · Content marketing ·  
 Native advertising  
**Search engine marketing**  
 Pay-per-click · Cost per impression ·  
 Search analytics · Web analytics  
**Display advertising**  
 Ad blocking · Contextual advertising ·  
 Behavioral targeting  
**Affiliate marketing**  
 Cost per action · Revenue sharing  
**Mobile advertising**  
 V · T · E



For example, I will create a topical authority map of the terms I want to cover.

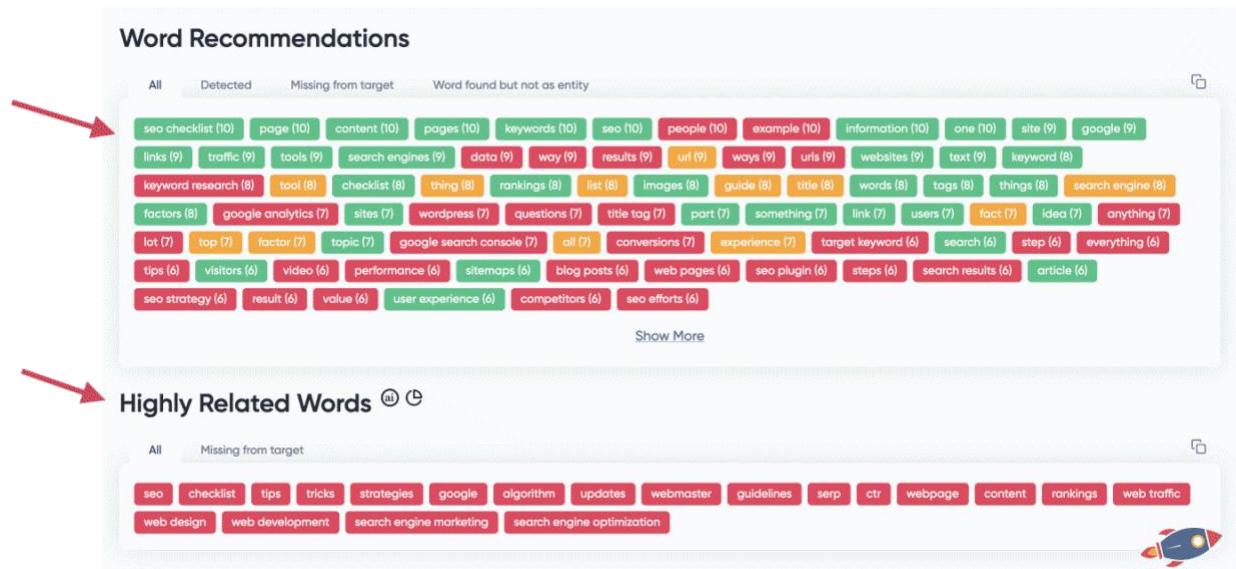
To build my topical map, I will look up:

- Related entities to my main term
- If applicable, I will look up a central wikipedia page on my entity to retrieve semantically related entities.
- I can also look at the knowledge graph for ideas and hints on semantically related entities.

I'll gather a list of approximately 20-50 semantically related topics to cover. These will branch out from my main focus entity and will be inter-linked together.

## 4. Optimized Content

When writing, I will use as many related entities as possible and make it clear what my main "focus entity" is.



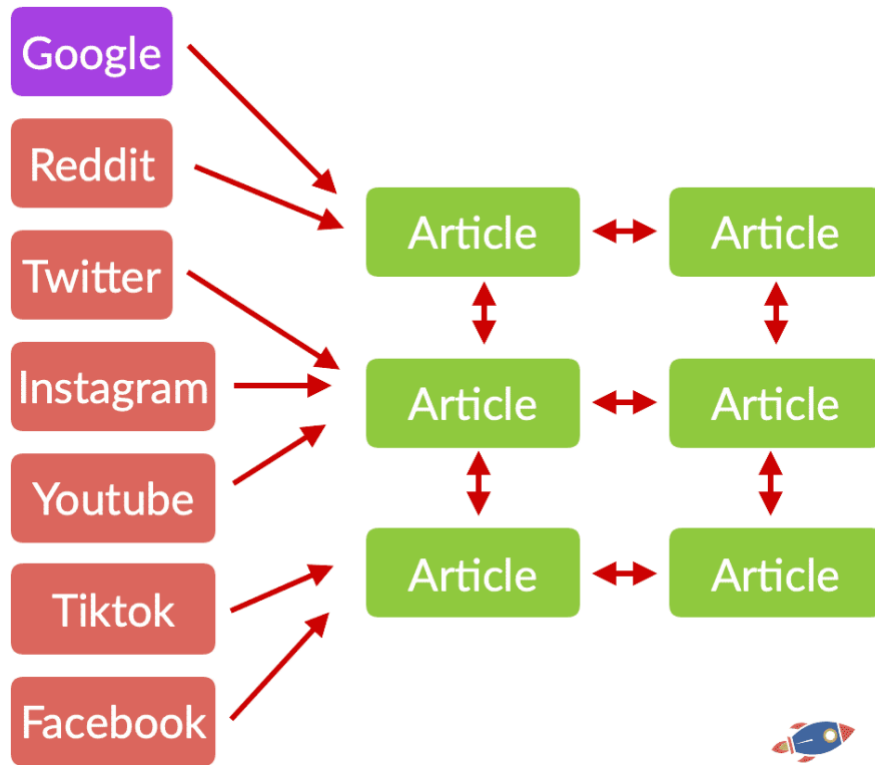
*Entities (Analyzed With Google's NLP Model)*

<https://on-page.ai>

I will ALSO make sure to include entities that my competitors aren't using so that my content is seen as original (and not just a copy of what's already there).

## 5. Distribution

After creating content, I will make sure it receives Chrome views by sharing it on social media.



If an article is not receiving Chrome views, then it's probably not good enough and it will struggle to rank.

## 6. News Links

I will be putting out regular press releases about my site as Google seems to notice when links are coming from news outlets.

However, I will be VERY wary of pointing my press releases directly at my content as link anchor text plays a large role in ranking.



Therefore, I will have a "news" section on my site where I point the press releases.



## 7. Internal Links

Google puts a significant emphasis on anchor text when ranking and internal links are ideal for creating anchor text relevance.

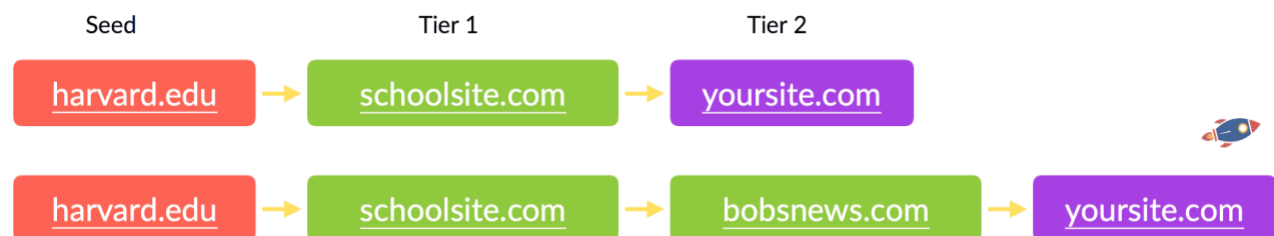


Whenever I create new content, **I will also create relevant internal links that use highly descriptive anchor texts.**

Specifically, I will use the exact match anchor text once and will vary it slightly afterwards. (I don't want to go overboard with the anchor however I do want to keep it related).

## 8. External Links

I will seek relevant links from related content pointing to my site.



Specifically, **I will try to get links from seed sites** (or more realistically, major sites that have links from seed sites.)

## 9. Publishing Rate

**I will aim to publish a minimum of ONE high piece of content per month in order for Google to keep my active boost** and this will let them see how users react to my newly published content.

*(Also not previously mentioned, I will use the same consistent author as Google makes a note of the author used.)*

### Additional Considerations

**I will be very weary of getting links from sites that might have an "expired" tag.** If I'm getting links, I will try to get some from domains that haven't dropped.

**I will also be very careful with anchor text and link velocity.** Too many links with exact anchors will likely trigger a penalty. Same goes for homepage links.

**I will also avoid using click-services as most users are likely going to be flagged as "unicorns"** so IF I do send fake traffic, it's going to be from a unique, low volume source. (Maybe a home made set of phones).

## Ranking For Highly Competitive Terms

One of the biggest realizations is that there's a lot more than just the words on the page that go into ranking a single page.

**The words / entities are immensely important HOWEVER it's also the entities found on *other* pages of your site that will affect the rankings of your main page. (Topical authority).**

**And it's also how users interact with your entire site that will affect how that pages ranks.**

**Finally, it's also the internal links and the external links anchors that will dictate rankings.**

Mastering Google traffic can be challenging at first however once you do it... you have access to the lion's share of traffic for your industry.

# Thank you!

## (References, Further Readings, Support)

This project represents a significant investment—over 158 combined man-hours from my team (including 126 from me) and numerous late-night sessions as I was pushing hard to complete this report for you. This is all I did during the month of June.

The reason I have chosen to make this publicly available at no cost is to help as many people as possible that have been recently affected by Google updates. I believe the SEO community has been pushed around enough and it's finally time to fight back against unfair practices.

To those people, I hope that by sharing some of my work helps you.

**If you run an SEO agency, I hope that some of my theories, data and tests help you better serve your clients. You can be the hero that rescues a struggling website.**

If you're an SEO professional, then the data might make more informed decisions, saving you time, money and effort.

**Should you feel this resources valuable, feel free to share this page within your professional network, within Facebook groups, Slack/Skype/Discord chats and forums.**

And should you wish to further support me in any way, I invite you to check out [On-Page.ai](#). Created with Google's best practices at its foundation, the platform aims to optimize your rankings by staying ahead of Google's frequent updates. One of my main motivations for keeping up with the latest SEO trends is to continually refine [On-Page.ai](#), ensuring it embodies the most effective ranking strategies. From discovering the most relevant entities to link building with relevance, to building topical authority, and creating incredibly well-researched content, [On-Page.ai](#) has you covered.

I'm confident that it can be a valuable asset in ranking your website online.

Here are some useful links:

- [Google API References](#)
- [On-Page.ai](#)

Should you wish to contribute any discoveries, case studies or any corrections, please email me at [team@on-page.ai](mailto:team@on-page.ai)

Sincerely,

- Eric Lancheres