Make your website easier to be crawled by SE bots

Webinar hosted by serpact

26th February 2019

Who am I?

VP of Growth at Brainly - since 2018

SEO Consultant - since 2015 *(worked with 70+ worldwide clients)*

Search Quality and Webspam at Google - 2010-2015

Startup Advisor - since 2013 *(provided advisory to 5 boards)*

Ambassador at OnCrawl - since 2018

Judge on UK (since 2016) + European + MENA (since 2017) + US (since 2018) **Search Awards**

Agenda

1. Life of a URL
2. What is crawl budget?
3. Significant factors affecting crawl
4. Building crawl friendly domains

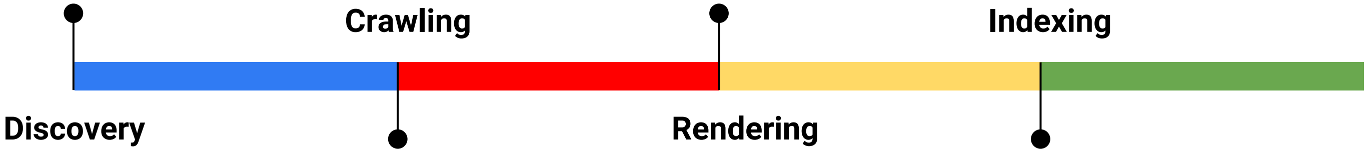
Life of a URL

Where are the new URLs or new content on existing URLs?

Example of a source:
- External links

Let's analyze the web content and interpret it like a user

On the 2nd wave of indexing, bots render URLs by fetching scripts



Discovery

Crawling

Rendering

Indexing

Are there more to find behind this initial lead provided by discovery?

Bots will crawl the site based on the statement delivered by crawl scheduler

What are we going to show our users?

Upon successful discovery, crawl and rendering of URLs, they get to stored and categorized based on their context

"Sorry, what was crawling again?"

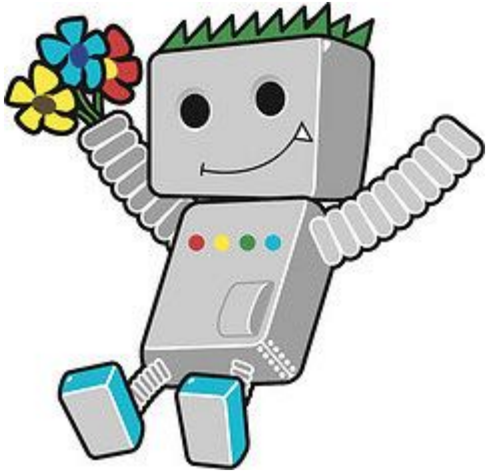
THE CONCEPT BEHIND CRAWLING

- Web consists of ever growing information
- Search Engines have crawlers (aka. bots) to discover publicly available URLs
- Crawlers look at URLs and follow links on those pages
- They go from link to link and bring data about those pages back to search engines' servers

What is crawl budget?

CRAWL BUDGET

Definition & parameters



- Crawl can create burden of a domain's server - SE bots need to **limit** their crawls
- SE bots decides how many URLs of a certain domain will be crawled for a given period of time by **crawl schedulers**
- This plan refers to ***crawl budget*** in SEO industry
- Crawl budget is not a static asset i.e. a URL can be crawled **more** or **less** frequently
- *Long story short*: Crawl budget means the number of URLs Googlebot can and wants to crawl

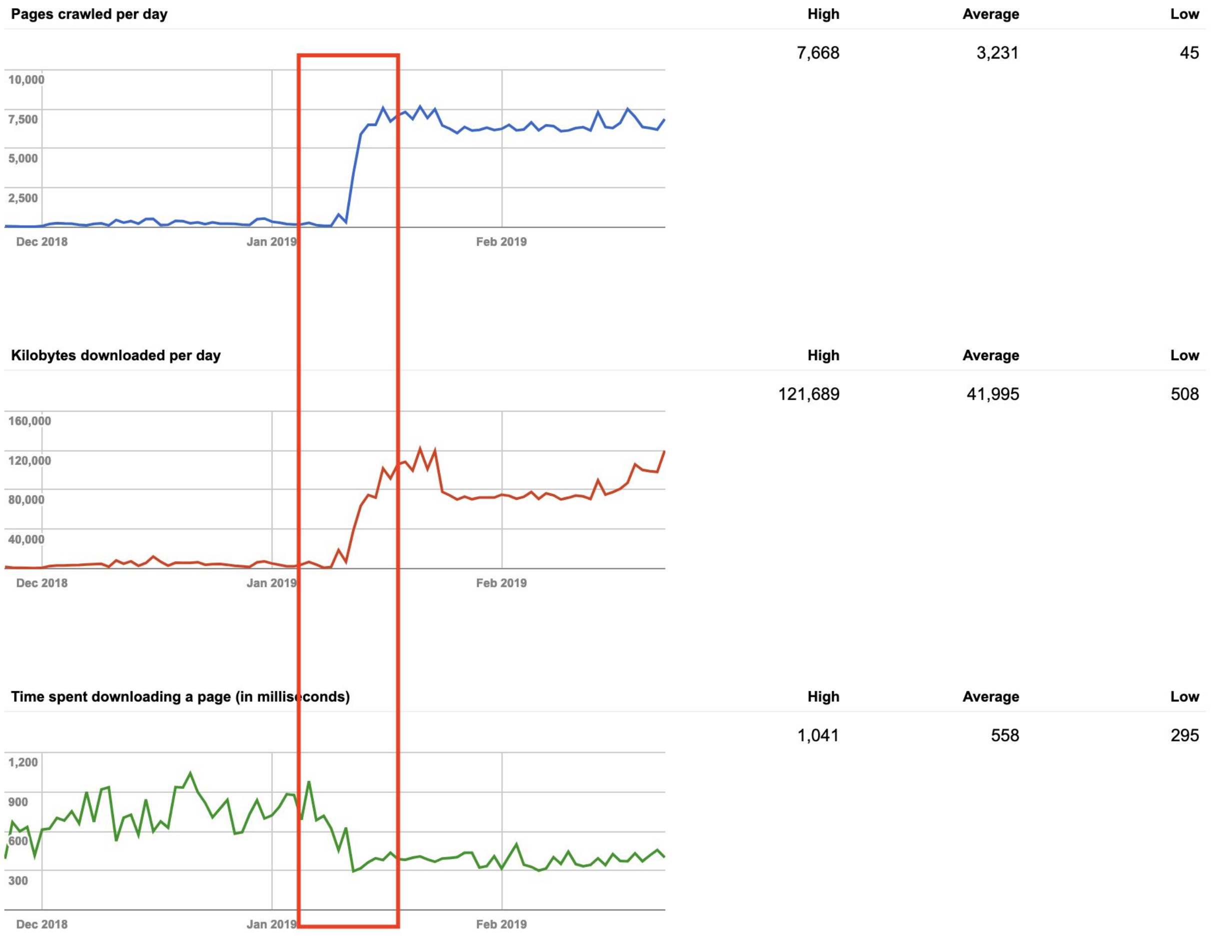
Should I really be concerned about my site's "crawl budget" ?

Sites with < 10K URLs	Sites with > 10K URLs
Unless your new pages are crawled around the same day of publication	YES!

Significant factors affecting crawl

#1	**Capacity to respond quickly**	● Time to last byte - TTLB ● Average time the bot downloads a page
#2	**Domain health**	● 50x / 40x / Soft 404 errors ● Redirection chains & faulty redirects ● Incorrect use of robots.txt + sitemap XML files
#3	**Quality of pages**	● Low quality, no added value or spam pages ● Duplicated pages *(index bloating)*
#4	**Internal linking structure**	● Too complex & deep website architecture ● # of internal & external links pointing to URLs
#5	**Freshness**	● Users tend to prefer fresh content ● PageRank decays in time

Capacity to respond quickly - *Page speed optimization*



Pages crawled per day

High	Average	Low
7,668	3,231	45

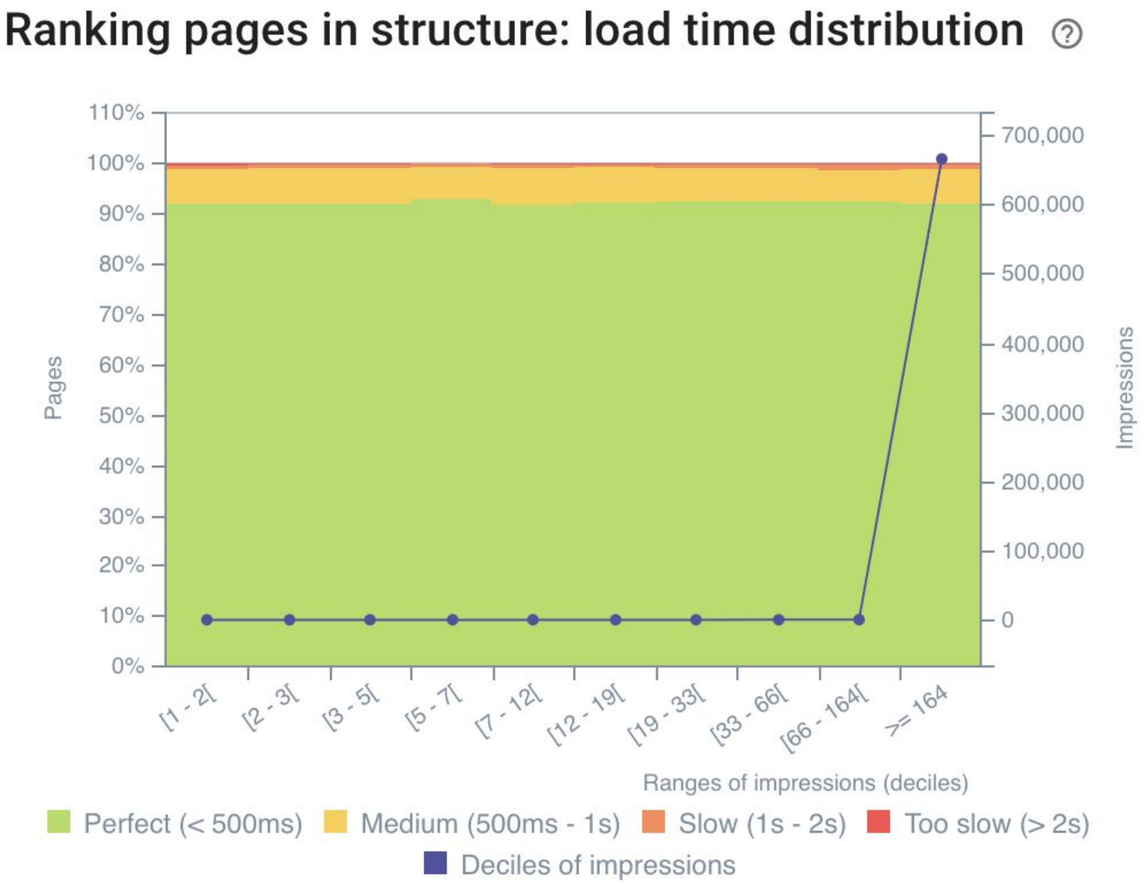
Kilobytes downloaded per day

High	Average	Low
121,689	41,995	508

Time spent downloading a page (in milliseconds)

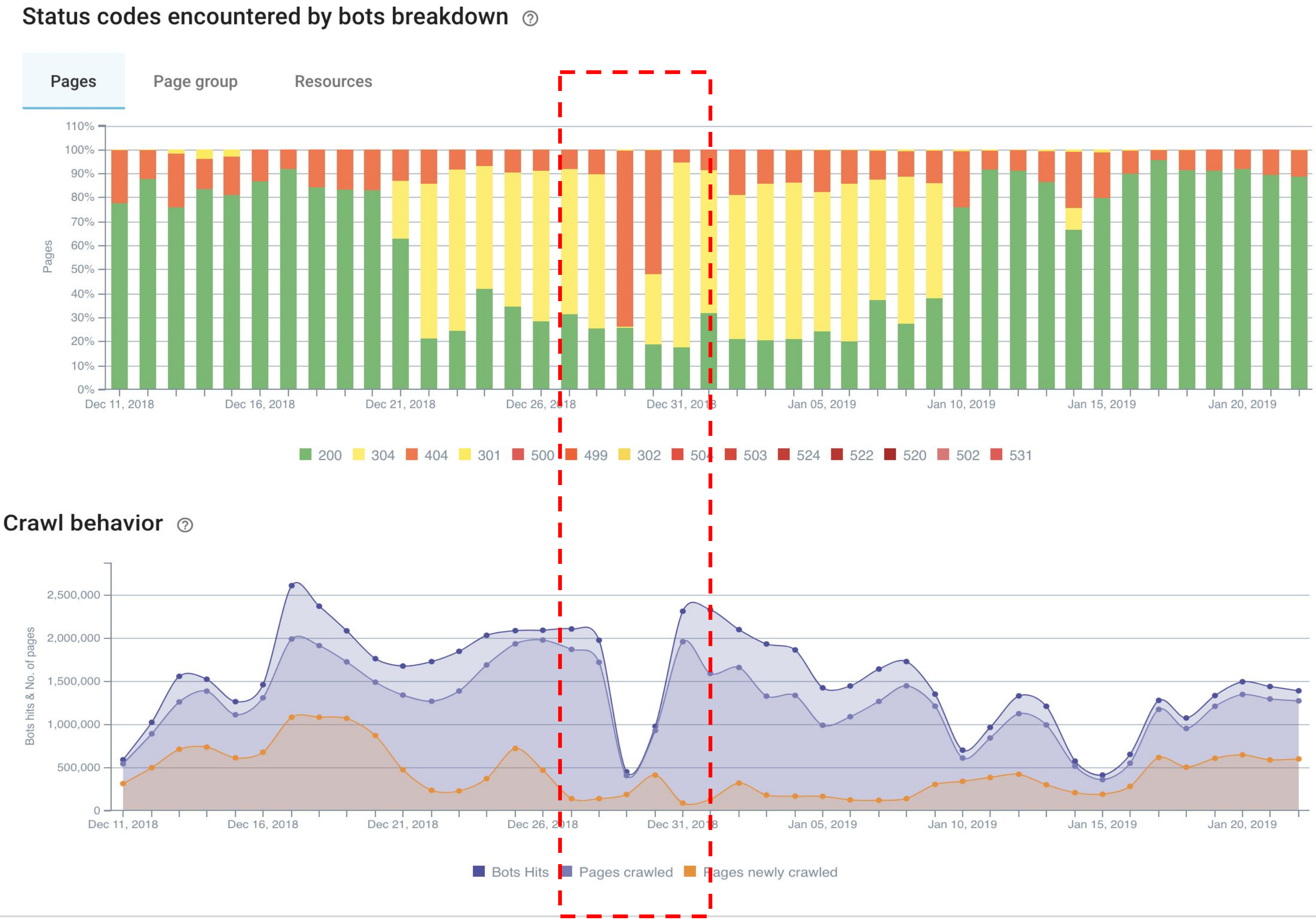
High	Average	Low
1,041	558	295

Enabling Googlebot to crawl URLs faster increases the # of crawls



Optimizing TTLB will not only help the site's URLs get crawled more but it's correlated with better ranking too

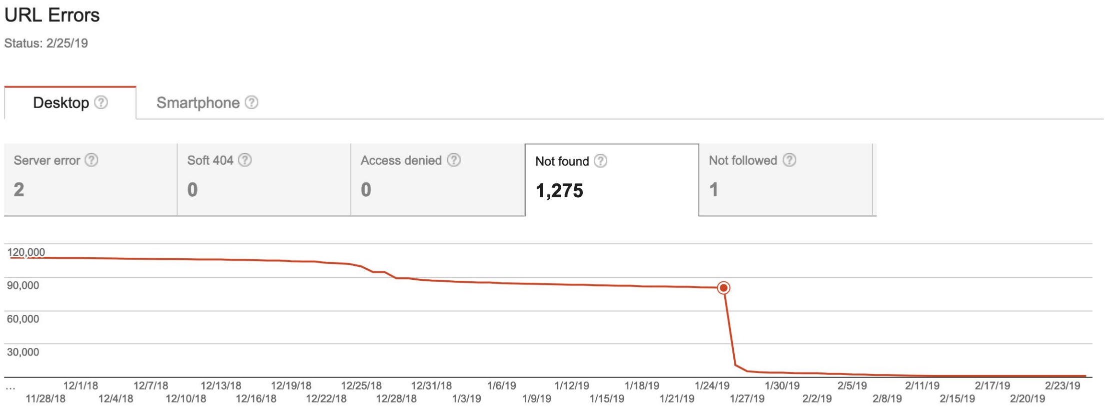
Domain health - *Site sanity exploration*



Status codes encountered by bots breakdown

Crawl behavior

When a site produces significant amount of 4xx or 5xx status codes, crawl frequency will be negatively affected

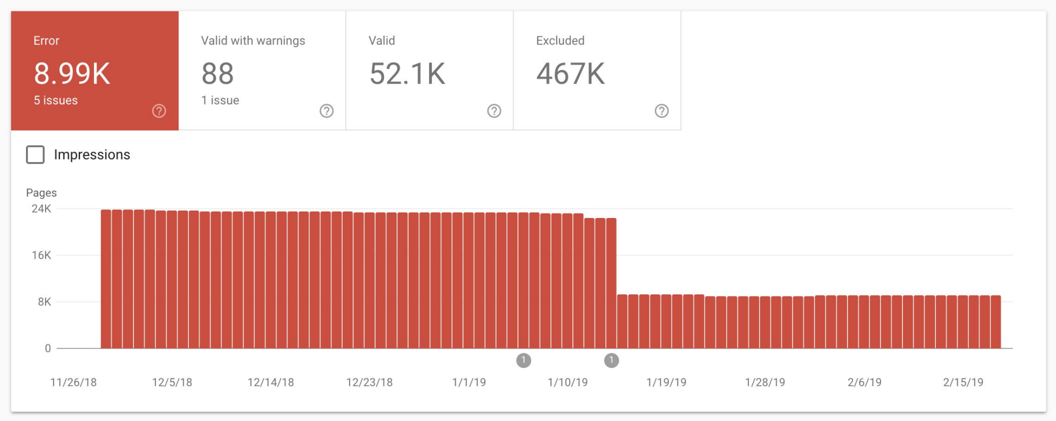


URL Errors

Status: 2/25/19

Desktop ⓘ | Smartphone ⓘ

Server error ⓘ	Soft 404 ⓘ	Access denied ⓘ	Not found ⓘ	Not followed ⓘ
2	0	0	1,275	1

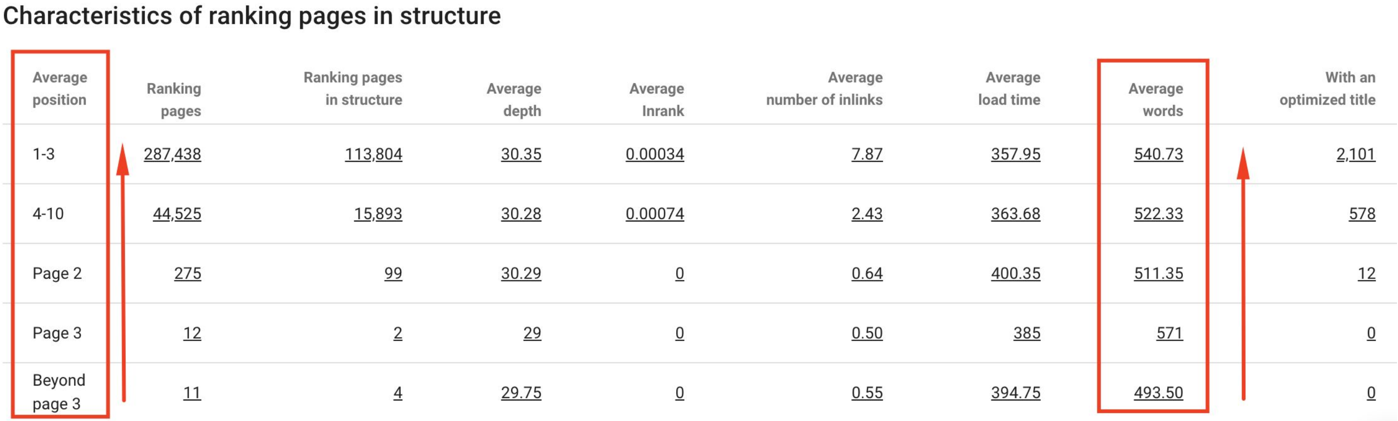


Error	Valid with warnings	Valid	Excluded
8.99K 5 issues	88 1 issue	52.1K	467K

☐ Impressions

Apart from your logs, monitor SC reports on URL errors to make sure you clean them up

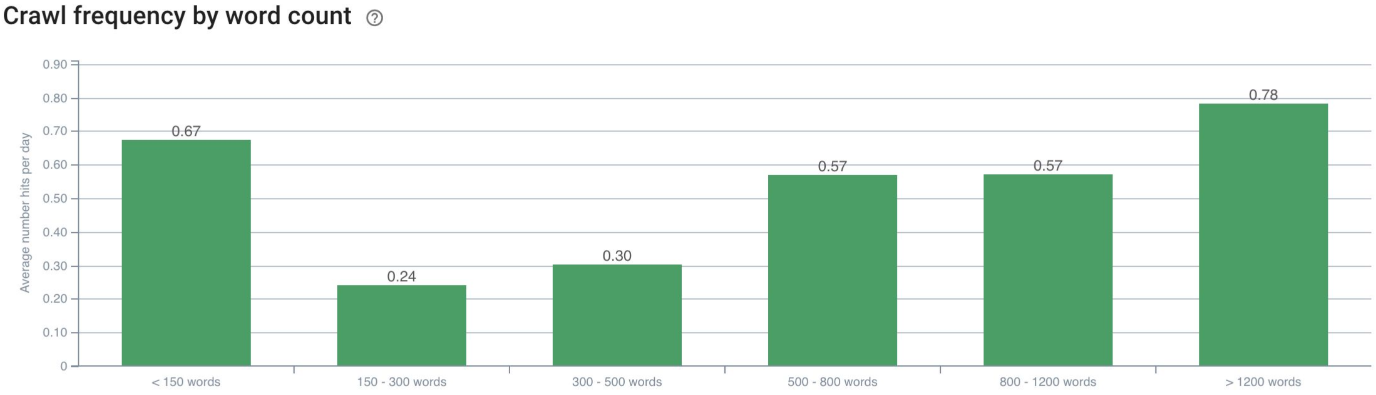
Quality of pages - *Quantify the quality*



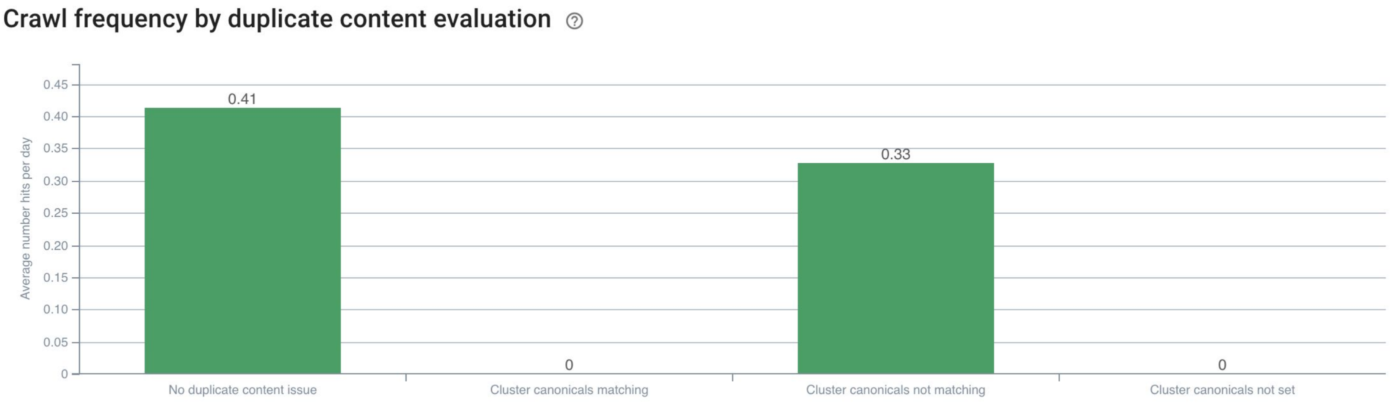
Characteristics of ranking pages in structure

Average position	Ranking pages	Ranking pages in structure	Average depth	Average Inrank	Average number of inlinks	Average load time	Average words	With an optimized title
1-3	287,438	113,804	30.35	0.00034	7.87	357.95	540.73	2,101
4-10	44,525	15,893	30.28	0.00074	2.43	363.68	522.33	578
Page 2	275	99	30.29	0	0.64	400.35	511.35	12
Page 3	12	2	29	0	0.50	385	571	0
Beyond page 3	11	4	29.75	0	0.55	394.75	493.50	0

Crawl frequency by word count



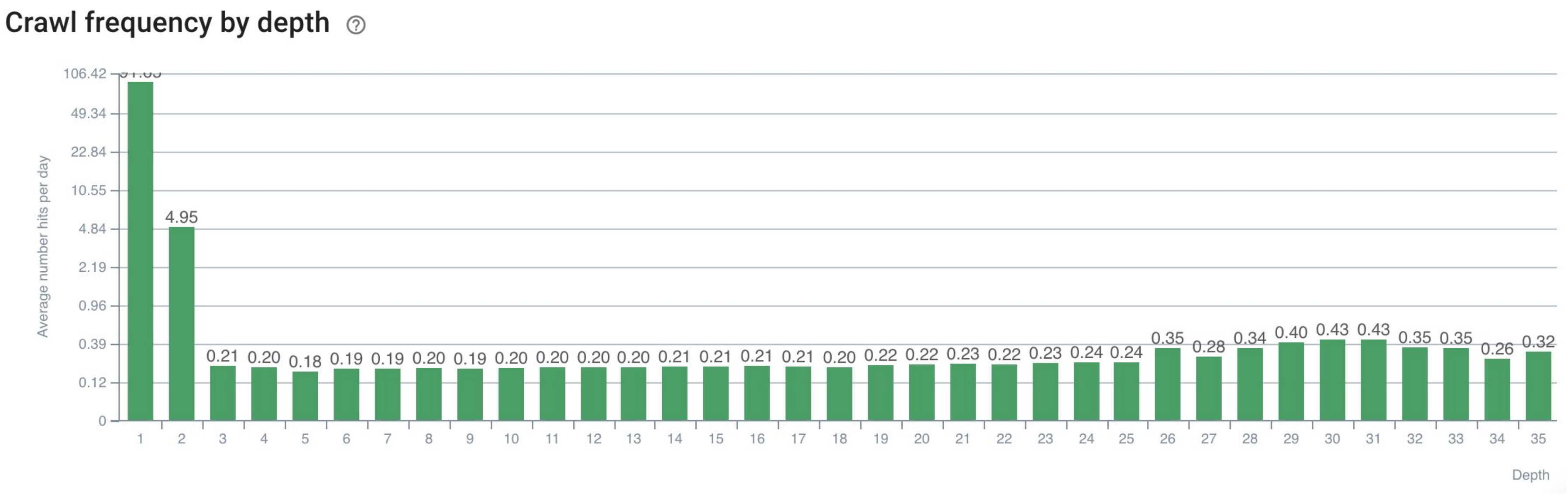
There is a clear correlation between unique + rich content and ranking, as well as crawl frequency and ranking



Watch out for duplicate content on your site. Reducing duplication will enable your important pages to be crawled more often

Internal linking structure - *PageRank propagation*

Crawl frequency by depth



Pages that are deep in the site structure (# of clicks away from HP) will less likely to get crawled more, unless you plan a clever site architecture based on user needs

Freshness

As web is a dynamic environment, PageRank decays in time.

Also, remember Googlebot keeps crawling to either discover new pages or find new content on existing pages.

Keep your website up to date!

Building crawl friendly domains - *A few tips to recap*

Build your way towards a crawl optimized domain

Priority 1 - Focus on user

- Produce high quality and value adding content
- Fulfill users' need by providing them useful features like search, navigation menu etc.
- Update your pages with appealing and fresh content

Priority 2 - Improve performance

- Improve the load time by optimizing sites' resources and increase server performance

Priority 3 - Fix on-page and technical issues

- Avoid duplication, site errors, redirection chains, orphan URLs, index bloating *(pagination, facets, spider traps)* etc.
- Ensure robustness of robots.txt and sitemap XML files

Priority 4 - Accessibility of valuable URLs

- Identify your value driving URLs and organize your site's internal linking structure accordingly

THANK YOU

Murat Yatagan