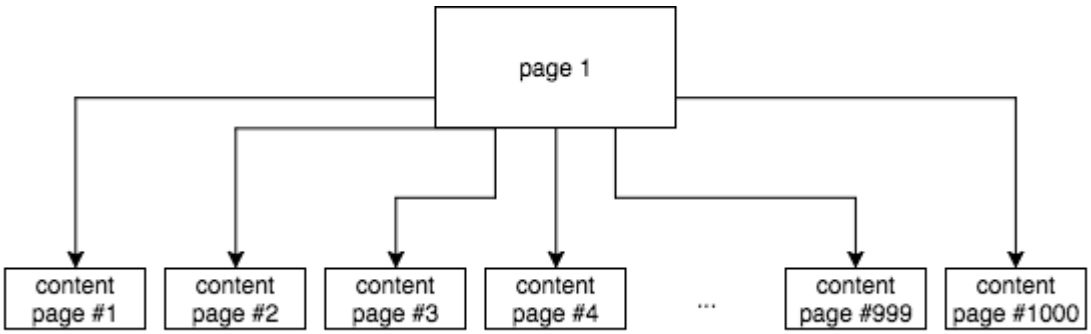
Ready, Set, Go! Googlebot Race

The Googlebot Race is an unusual tournament watched daily with engagement by over 1.8 billion websites. The tournament consists of many competitions commonly referred to as “ranking factors.” Every year, somebody tries to describe as many of them as possible, but nobody really knows what they are all about and how many there are. Nobody but Googlebot. It is he who daily traverses petabytes of data, forcing webmasters to compete on the weirdest fields, to choose the best ones. Or that is what he thinks.

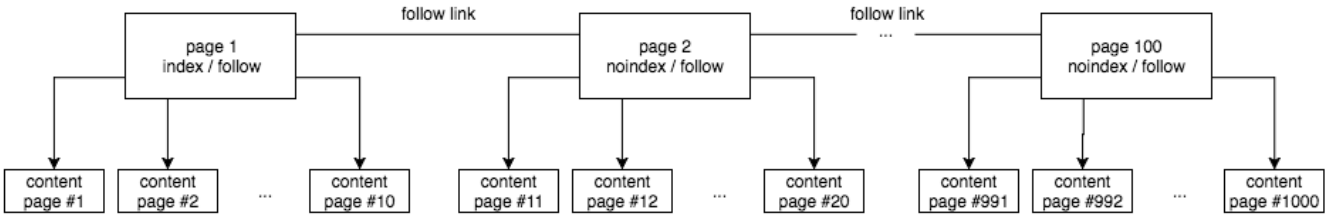
The 1,000 meters run (with steeplechase) – we are checking indexation speed. For this competition, I presented five similar data structures. Each of them had 1000 subpages with unique content and additional navigation pages (e.g. other subpages or categories). Below you can see the results for four running tracks.

This data structure was very poor with 1,000 links to subpages with unique content on one page (so 1,000 internal links). All SEO experts (including me…) repeat it like a mantra: no more than 100 internal links per page or Google will not manage to crawl such an extensive page and it will simply ignore some of the links, and it will not index them. I decided to see if it was true.

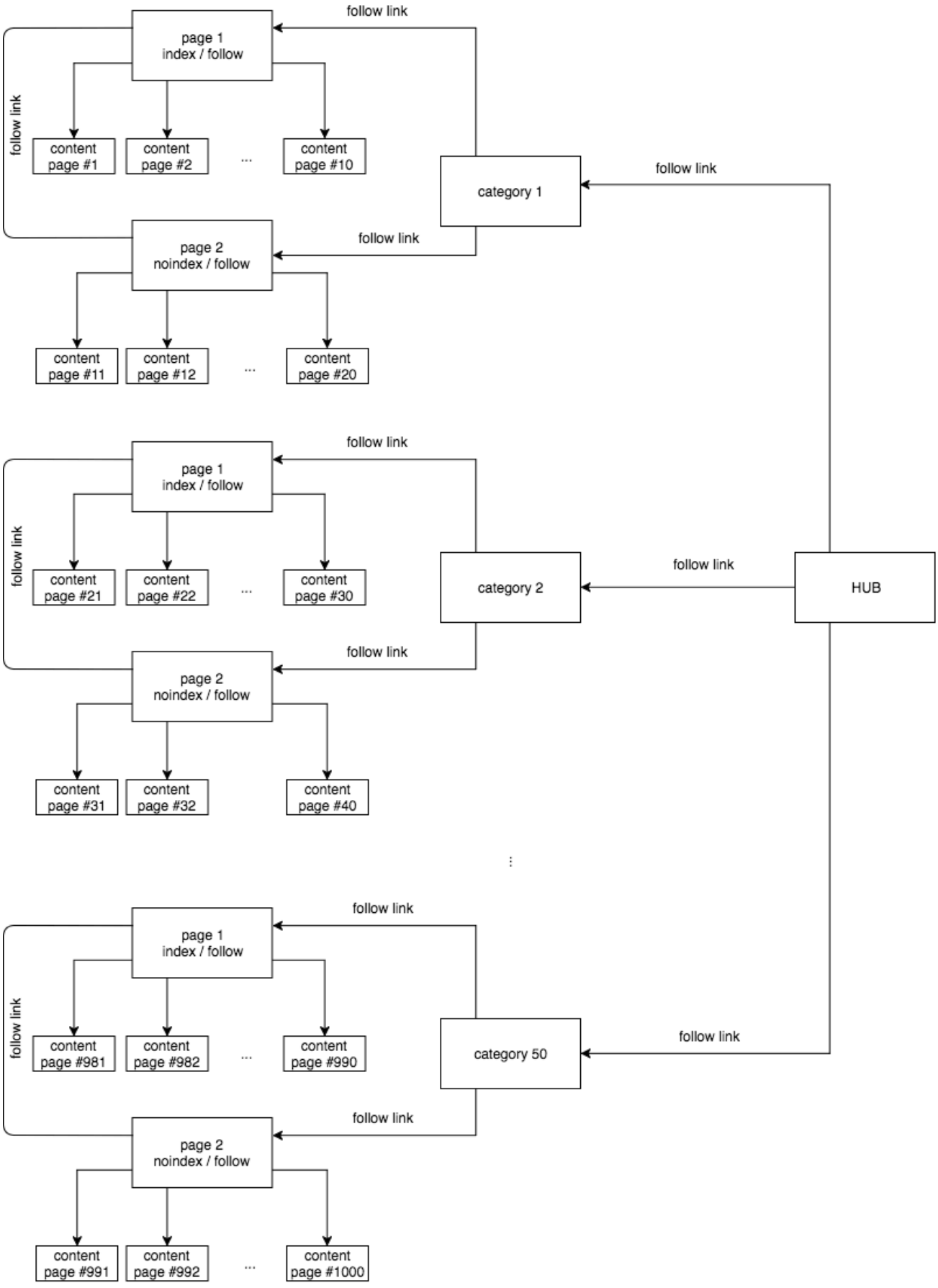


This is an average running track. Another 100 subpages (on

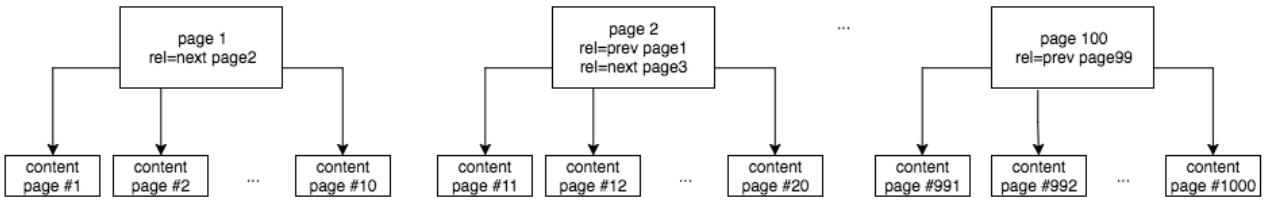
each of them, visible links to a few former pages, a few following pages, to the first one and to the last one). On each subpage, 10 internal links to pages with content. The first page consists of the meta robots tag index/follow, the other one noindex/follow.



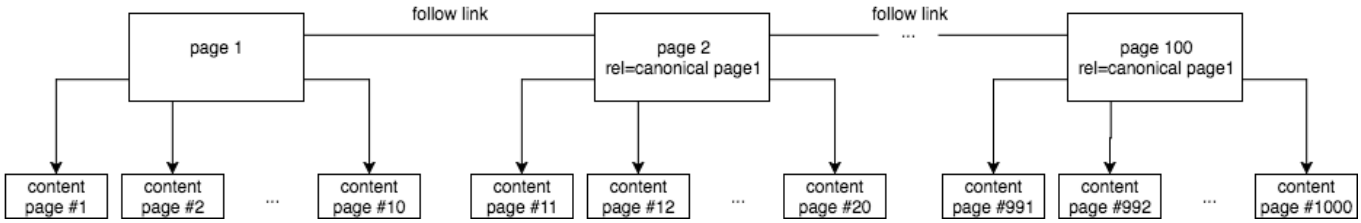
I wanted to introduce a little confusion, so I decided to create a silo structure on the website, and I divided it into 50 categories. In each of them, there were 20 links to content pages divided into two pages.


HUB
category 1
page 1
index / follow
page 2
noindex / follow
content
page #1
content
page #2
content
page #10
content
page #11
content
page #12
content
page #20
follow link
category 2
page 1
index / follow
page 2
noindex / follow
content
page #21
content
page #22
content
page #30
content
page #31
content
page #32
content
page #40
follow link
category 50
page 1
index / follow
page 2
noindex / follow
content
page #981
content
page #982
content
page #990
content
page #991
content
page #992
content
page #1000
follow link

The next running track is the dark horse of this tournament. No normal pagination/paging. Instead, solely rel=”next” i rel=”prev” headlines paging/pagination, defining the following page to which Googlebot should go.



Running track number two is similar. The difference is that I got rid of index/noindex and I set canonical tags for all subpages to the first page.



and they took up…

	Running track 1 (1000 links on the page)		Running track 2 (classic pagination)		Running track 3 (silo structure)		Running track 4 (rel next / prev pagination)		Running track 5 (rel canonical to the first page)	
	hits	indexed	hits	indexed	hits	indexed	hits	indexed	hits	indexed
Day 1	4	0	2	0	**246**	0	4	0	198	1
Day 10	2700	240	1471	121	2644	255	55	10	2684	455
Day 20	3932	861	2182	390	3398	810	121	10	3527	980
Day 30	4252	880	2274	352	3643	920	152	10	3821	992

hits – total number of Googlebot visits

indexed – number of indexed pages

I must admit that I was disappointed by the results. I was very much hoping to demonstrate that the silo structure would speed up the crawling and the indexation of the site. Unfortunately, it did not happen. This kind of structure is the one that I usually recommend and implement on websites that I administer, mainly because of the possibilities that it gives for internal linking. Sadly, with a larger amount of information, it does not go hand in hand with indexation speed.

Nevertheless, to my surprise, Googlebot easily dealt with reading 1,000 internal links, visiting them for 30 days and indexing the majority. But it is commonly believed that the number of internal links should be 100 per page. This means that if we want to speed the indexation up, we should create website's maps in HTML format even with such a large number of links.

At the same time, classic indexation with noindex/follow is absolutely losing against pagination with the use of index/follow and rel=canonical directing to the first page. In the case of the last one, Googlebot was expected not to index specific paginated subpages. Nevertheless, from 100 paginated subpages, it has indexed five, despite the canonical tag to page one, which shows again (I wrote about it here) that setting canonical tags does not guarantee avoiding the indexation of a page and the resulting mess in the search engine's index.

In the case of the above-described test, the last construction is the most effective one for the number of pages indexed. If we introduced a new notion Index Rate defined by the proportion of the number of Googlebot visits to the number of pages indexed, e.g., within 30 days, then the best IR in our test would be 3,89 (running track 5) and the worst one would be 6,46 (running track 2). This number would stand for average number of Googlebot's visits on a page required to index it (and keep it in the index). To further define IR, it would be worth verifying the indexation daily for a specific URL. Then, it would definitely make more sense.

One of the key conclusions from this article (after a few days from the beginning of the experiment) would be demonstrating that Googlebot ignores rel=next and rel=prev tags. Unfortunately, I was late to publish those results (waiting for more) and John Muller on March 21 announced to the world that indeed, these tags are not used by Googlebot. I am just wondering whether the fact that I am typing this article in Google Docs has anything to do with it (#conspiracytheory).

It is worth taking a look at pages containing infinite scroll – dynamic content uploading, uploaded after scrolling down to the lower parts of the page and the navigation based on rel=prev and rel=next. If there is no other navigation, such as regular pagination hidden in CSS (invisible for the user but visible for Googlebot) we can be sure that Googlebot's

access to newly uploaded content (products, articles, photos) will be hindered.

Opinions expressed in this article are those of the guest author and not necessarily Search Engine Land. Staff authors are listed here.

About The Author



“Do not accept ‘just’ high quality. Anyone can do that. If the sky is the limit, find a higher sky.” Max Cyrek is CEO of Cyrek Digital, a digital marketing consultant and SEO evangelist. Throughout his career, Max, together with his team of over 30, has worked with hundreds of companies helping them succeed. He has been working in digital marketing for nearly ten years and has specialized in technical SEO, managing successful marketing projects.

http://feeds.searchengineland.com/~r/searchengineland/~3/Nd_jTmBxRMA/ready-set-go-googlebot-race-314894