Nutch FAQ

Table of contents

How can I stop Nutch from crawling my site?

Please visit our webmaster info page.

Will Nutch be a distributed, P2P-based search engine?

We don't think it is presently possible to build a peer-to-peer search engine that is competitive with existing search engines. It would just be too slow. Returning results in less than a second is important: it lets people rapidly reformulate their queries so that they can more often find what they're looking for. In short, a fast search engine is a better search engine. I don't think many people would want to use a search engine that takes ten or more seconds to return results.

That said, if someone wishes to start a sub-project of Nutch exploring distributed searching, we'd love to host it. We don't think these techniques are likely to solve the hard problems Nutch needs to solve, but we'd be happy to be proven wrong.

Will Nutch use a distributed crawler, like Grub?

Distributed crawling can save download bandwidth, but, in the long run, the savings is not significant. A successful search engine requires more bandwidth to upload query result pages than its crawler needs to download pages, so making the crawler use less bandwidth does not reduce overall bandwidth requirements. The dominant expense of operating a search engine is not crawling, but searching.

Won't open source just make it easier for sites to manipulate rankings?

Search engines work hard to construct ranking algorithms that are immune to manipulation. Search engine optimizers still manage to reverse-engineeer the ranking algorithms used by search engines, and improve the ranking of their pages. For example, many sites use link farms to manipulate search engines' link-based ranking algorithms, and search engines retaliate by improving their link-based algorithms to neutralize the effect of link farms.

With an open-source search engine, this will still happen, just out in the open. This is analagous to encryption and virus protection software. In the long term, making such algorithms open source makes them stronger, as more people can examine the source code to find flaws and suggest improvements. Thus we believe that an open source search engine has the potential to better resist manipulation of its rankings.