Broad Crawls

Scrapy defaults are optimized for crawling specific sites. These sites are often handled by a single Scrapy spider, although this is not necessary or required (for example, there are generic spiders that handle any given site thrown at them).

In addition to this "focused crawl", there is another common type of crawling which covers a large (potentially unlimited) number of domains, and is only limited by time or other arbitrary constraint, rather than stopping when the domain was crawled to completion or when there are no more requests to perform. These are called "broad crawls" and is the typical crawlers employed by search engines.

These are some common properties often found in broad crawls:

- they crawl many domains (often, unbounded) instead of a specific set of sites
- they don't necessarily crawl domains to completion, because it would be impractical (or impossible) to do so, and instead limit the crawl by time or number of pages crawled
- they are simpler in logic (as opposed to very complex spiders with many extraction rules) because data is often post-processed in a separate stage
- they crawl many domains concurrently, which allows them to achieve faster crawl speeds by not being limited by any particular site constraint (each site is crawled slowly to respect politeness, but many sites are crawled in parallel)

As said above, Scrapy default settings are optimized for focused crawls, not broad crawls. However, due to its asynchronous architecture, Scrapy is very well suited for performing fast broad crawls. This page summarizes some things you need to keep in mind when using Scrapy for doing broad crawls, along with concrete suggestions of Scrapy settings to tune in order to achieve an efficient broad crawl.

Use the right :setting:`SCHEDULER_PRIORITY_QUEUE`

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst`, **line 44**); ***backlink***

Unknown interpreted text role "setting".

Scrapyâ€™s default scheduler priority queue is `'scrapy.pqueues.ScrapyPriorityQueue'`. It works best during single-domain crawl. It does not work well with crawling many different domains in parallel

To apply the recommended priority queue use:

```
SCHEDULER_PRIORITY_QUEUE = 'scrapy.pqueues.DownloaderAwarePriorityQueue'
```

Increase concurrency

Concurrency is the number of requests that are processed in parallel. There is a global limit (:setting:`CONCURRENT_REQUESTS`) and an additional limit that can be set either per domain (:setting:`CONCURRENT_REQUESTS_PER_DOMAIN`) or per IP (:setting:`CONCURRENT_REQUESTS_PER_IP`).

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst`, **line 60**); ***backlink***

Unknown interpreted text role "setting".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst`, **line 60**); ***backlink***

Unknown interpreted text role "setting".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst`, **line 60**); ***backlink***

Unknown interpreted text role "setting".

Note

The scheduler priority queue :ref:`recommended for broad crawls <broad-crawls-scheduler-priority-queue>` does not support :setting:`CONCURRENT_REQUESTS_PER_IP`.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 65); *backlink*

Unknown interpreted text role "ref".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 65); *backlink*

Unknown interpreted text role "setting".

The default global concurrency limit in Scrapy is not suitable for crawling many different domains in parallel, so you will want to increase it. How much to increase it will depend on how much CPU and memory you crawler will have available.

A good starting point is `100`:

```
CONCURRENT_REQUESTS = 100
```

But the best way to find out is by doing some trials and identifying at what concurrency your Scrapy process gets CPU bounded. For optimum performance, you should pick a concurrency where CPU usage is at 80-90%.

Increasing concurrency also increases memory usage. If memory usage is a concern, you might need to lower your global concurrency limit accordingly.

Increase Twisted IO thread pool maximum size

Currently Scrapy does DNS resolution in a blocking way with usage of thread pool. With higher concurrency levels the crawling could be slow or even fail hitting DNS resolver timeouts. Possible solution to increase the number of threads handling DNS queries. The DNS queue will be processed faster speeding up establishing of connection and crawling overall.

To increase maximum thread pool size use:

```
REACTOR_THREADPOOL_MAXSIZE = 20
```

Setup your own DNS

If you have multiple crawling processes and single central DNS, it can act like DoS attack on the DNS server resulting to slow down of entire network or even blocking your machines. To avoid this setup your own DNS server with local cache and upstream to some large DNS like OpenDNS or Verizon.

Reduce log level

When doing broad crawls you are often only interested in the crawl rates you get and any errors found. These stats are reported by Scrapy when using the `INFO` log level. In order to save CPU (and log storage requirements) you should not use `DEBUG` log level when preforming large broad crawls in production. Using `DEBUG` level when developing your (broad) crawler may be fine though.

To set the log level use:

```
LOG_LEVEL = 'INFO'
```

Disable cookies

Disable cookies unless you *really* need. Cookies are often not needed when doing broad crawls (search engine crawlers ignore them), and they improve performance by saving some CPU cycles and reducing the memory footprint of your Scrapy crawler.

To disable cookies use:

```
COOKIES_ENABLED = False
```

Disable retries

Retrying failed HTTP requests can slow down the crawls substantially, specially when sites causes are very slow (or fail) to respond, thus causing a timeout error which gets retried many times, unnecessarily, preventing crawler capacity to be reused for other domains.

To disable retries use:

```
RETRY_ENABLED = False
```

Reduce download timeout

Unless you are crawling from a very slow connection (which shouldn't be the case for broad crawls) reduce the download timeout so

that stuck requests are discarded quickly and free up capacity to process the next ones.

To reduce the download timeout use:

```
DOWNLOAD_TIMEOUT = 15
```

Disable redirects

Consider disabling redirects, unless you are interested in following them. When doing broad crawls it's common to save redirects and resolve them when revisiting the site at a later crawl. This also help to keep the number of request constant per crawl batch, otherwise redirect loops may cause the crawler to dedicate too many resources on any specific domain.

To disable redirects use:

```
REDIRECT_ENABLED = False
```

Enable crawling of "Ajax Crawlable Pages"

Some pages (up to 1%, based on empirical data from year 2013) declare themselves as ajax crawlable. This means they provide plain HTML version of content that is usually available only via AJAX. Pages can indicate it in two ways:

1. by using `#!` in URL - this is the default way;
2. by using a special meta tag - this way is used on "main", "index" website pages.

Scrapy handles (1) automatically; to handle (2) enable :ref:`AjaxCrawlMiddleware <ajaxcrawl-middleware>`:

> **System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 181); *backlink***
>
> Unknown interpreted text role "ref".

```
AJAXCRAWL_ENABLED = True
```

When doing broad crawls it's common to crawl a lot of "index" web pages; AjaxCrawlMiddleware helps to crawl them correctly. It is turned OFF by default because it has some performance overhead, and enabling it for focused crawls doesn't make much sense.

Crawl in BFO order

:ref:`Scrapy crawls in DFO order by default <faq-bfo-dfo>`.

> **System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 198); *backlink***
>
> Unknown interpreted text role "ref".

In broad crawls, however, page crawling tends to be faster than page processing. As a result, unprocessed early requests stay in memory until the final depth is reached, which can significantly increase memory usage.

:ref:`Crawl in BFO order <faq-bfo-dfo>` instead to save memory.

> **System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 204); *backlink***
>
> Unknown interpreted text role "ref".

Be mindful of memory leaks

If your broad crawl shows a high memory usage, in addition to :ref:`crawling in BFO order <broad-crawls-bfo>` and :ref:`lowering concurrency <broad-crawls-concurrency>` you should :ref:`debug your memory leaks <topics-leaks>`.

> **System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 210); *backlink***
>
> Unknown interpreted text role "ref".

> **System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 210); *backlink***
>
> Unknown interpreted text role "ref".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 210); *backlink*

Unknown interpreted text role "ref".

Install a specific Twisted reactor

If the crawl is exceeding the system's capabilities, you might want to try installing a specific Twisted reactor, via the :setting:`TWISTED_REACTOR` setting.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scrapy-master\docs\topics\[scrapy-master][docs][topics]broad-crawls.rst, line 219); *backlink*

Unknown interpreted text role "setting".