AutoThrottle扩展

这是基于Scrapy服务器和您正在抓取的网站的负载自动限制爬网速度的扩展。

设计目标

1. 更好的网站，而不是默认下载延迟为零
2. 自动将scrapy调整到最佳爬行速度，因此用户无需调整下载延迟以找到最佳延迟。用户只需要指定它允许的最大并发请求，扩展完成剩下的工作。

它是如何工作的

AutoThrottle扩展动态调整下载延迟，使蜘蛛`AUTOTHROTTLE_TARGET_CONCURRENCY`平均向每个远程网站发送 并发请求。

它使用下载延迟来计算延迟。其主要思想是：如果一台服务器需要`latency`秒钟响应，客户端应该发送一个请求的每个`latency/N`秒，具有`N`并行处理的请求。

可以设置一个小的固定下载延迟，并使用`CONCURRENT_REQUESTS_PER_DOMAIN`或`CONCURRENT_REQUESTS_PER_IP`选项对并发性施加硬限制，而不是调整延迟 。它会提供类似的效果，但有一些重要的区别：

- 因为下载延迟很小，偶尔会有突发的请求;
- 通常非200（错误）响应可以比常规响应更快地返回，因此当下载延迟较小且硬并发限制时，爬虫将在服务器开始返回错误时更快地向服务器发送请求。但这与爬虫应该做的事情相反 - 如果出现错误，减速更有意义：这些错误可能是由高请求率引起的。

AutoThrottle没有这些问题。

节流算法

AutoThrottle算法根据以下规则调整下载延迟：

1. 蜘蛛总是以下载延迟开始`AUTOTHROTTLE_START_DELAY`;
2. 接收到响应时，目标延迟下载被计算为 ，其中是响应的等待时间，并且是。`latency / N``latency``N``AUTOTHROTTLE_TARGET_CONCURRENCY`
3. 下次请求的下载延迟设置为先前下载延迟和目标下载延迟的平均值;
4. 非200响应的延迟不允许减少延迟;
5. 下载延迟不能小于`DOWNLOAD_DELAY`或大于`AUTOTHROTTLE_MAX_DELAY`

注意

AutoThrottle扩展程序遵循标准的Scrapy设置以实现并发和延迟。这意味着它将尊重 `CONCURRENT_REQUESTS_PER_DOMAIN` 和 `CONCURRENT_REQUESTS_PER_IP` 选择，并且永远不会将下载延迟设置为低于 `DOWNLOAD_DELAY` 。

在Scrapy中，下载延迟测量为建立TCP连接和接收HTTP标头之间经过的时间。

请注意，在协作式多任务处理环境中，这些延迟很难准确测量，因为Scrapy可能正忙于处理蜘蛛回调，例如，无法参加下载。但是，这些延迟仍然可以合理估计Scrapy（最终是服务器）的繁忙程度，并且此扩展基于该前提。

设置

用于控制AutoThrottle扩展的设置包括：

- `AUTOTHROTTLE_ENABLED`
- `AUTOTHROTTLE_START_DELAY`
- `AUTOTHROTTLE_MAX_DELAY`
- `AUTOTHROTTLE_TARGET_CONCURRENCY`
- `AUTOTHROTTLE_DEBUG`
- `CONCURRENT_REQUESTS_PER_DOMAIN`
- `CONCURRENT_REQUESTS_PER_IP`
- `DOWNLOAD_DELAY`

有关更多信息，请参阅其工作原理。

AUTOTHROTTLE_ENABLED

默认： `False`

启用AutoThrottle扩展。

AUTOTHROTTLE_START_DELAY

默认： `5.0`

初始下载延迟（以秒为单位）。

AUTOTHROTTLE_MAX_DELAY

默认： `60.0`

在高延迟的情况下要设置的最大下载延迟（以秒为单位）。

AUTOTHROTTLE_TARGET_CONCURRENCY

版本1.1中的新功能。

默认： `1.0`

Scrapy应该与远程网站并行发送的平均请求数。

默认情况下，AutoThrottle会调整延迟以向每个远程网站发送单个并发请求。将此选项设置为更高的值（例如 `2.0` ）以增加远程服务器上的吞吐量和负载。较低的 `AUTOTHROTTLE_TARGET_CONCURRENCY` 值（例如 `0.5` ）使得爬虫更加保守和礼貌。

请注意， `CONCURRENT_REQUESTS_PER_DOMAIN` 并 `CONCURRENT_REQUESTS_PER_IP` 在启用自动油门扩展选项仍然尊重。这意味着如果 `AUTOTHROTTLE_TARGET_CONCURRENCY` 将if 设置为高于 `CONCURRENT_REQUESTS_PER_DOMAIN` 或 的值 `CONCURRENT_REQUESTS_PER_IP` ，则爬网程序将不会达到此并发请求数。

在每个给定的时间点，Scrapy可以发送或多或少的并发请求 `AUTOTHROTTLE_TARGET_CONCURRENCY` ；它是爬虫尝试接近的建议值，而不是硬限制。

AUTOTHROTTLE_DEBUG

默认： `False`

启用AutoThrottle调试模式，该模式将显示收到的每个响应的统计信息，以便您可以查看如何实时调整限制参数。