GROWING WEB SPIDERS

Juozas Kaziukėnas // juokaz.com // @juokaz

300’000’000 products

/ 24 hours = 12’500’00 products

/ 3600 seconds = 3’472 products

/ **3000** nodes = 1.1 sec. per product

24’000 cores on Amazon = $300/h

Juozas Kaziukėnas, Lithuanian

You can call me Joe

More info http://juokaz.com

WHY CRAWL?

WE NEED DATA

1. Get data
2. ???
3. Profit

IF PEOPLE ARE SCRAPPING YOUR SITE, YOU HAVE DATA PEOPLE WANT. CONSIDER MAKING AN API

Russell Ahlstrom

DATA SCIENCE

1. FIGURE OUT WHAT TO REQUEST
2. MAKE A REQUEST
3. PARSE THE REQUEST
4. STORE RESULTS

WHAT TO EXTRACT

AS LITTLE AS POSSIBLE

MAKE A REQUEST

FILE_GET_CONTENTS($URL);

HANDLING HTTP ERRORS

OPTIMIZE HTTP REQUESTS

```php
function get($url) {
    // Create a handle.
    $handle = curl_init($url);

    // Set options...

    // Do the request.
    $ret = curl_exec($handle);

    // Do stuff with the results...

    // Destroy the handle.
    curl_close($handle);
}
```

```php
function get($url) {
	// Create a handle.
	$handle = curl_init($url);

	// Set options...

	// Do the request.
	$ret = curlExecWithMulti($handle);

	// Do stuff with the results...

	// Destroy the handle.
	curl_close($handle);
}
```

```php
function curlExecWithMulti($handle) {
    // In real life this is a class variable.
    static $multi = NULL;

    // Create a multi if necessary.
    if (empty($multi)) { $multi = curl_multi_init(); }

    // Add the handle to be processed.
    curl_multi_add_handle($multi, $handle);

    // Do all the processing.
    $active = NULL;
    do {
        $ret = curl_multi_exec($multi, $active);
    } while ($ret == CURLM_CALL_MULTI_PERFORM);

    while ($active && $ret == CURLM_OK) {
      if (curl_multi_select($multi) != -1) {
      do {
          $mrc = curl_multi_exec($multi, $active);
       } while ($mrc == CURLM_CALL_MULTI_PERFORM);
     }
    }

    // Remove the handle from the multi processor.
    curl_multi_remove_handle($multi, $handle);

    return TRUE;
}
```

QUEUES FOR EVERYTHING

ASYNCHRONOUS PROCESSING

DO NOT BLOCK FOR I/O

RETRIES

REGULAR EXPRESSIONS

REGULAR EXPRESSIONS NOT

XPATH

PHANTOM.JS/SELENIUM

WHAT HAPPENS WHEN THE PAGE CHANGES

ACTING LIKE A HUMAN

HTTP HEADERS

```
$HEADER = ARRAY();
$HEADER[0] = "ACCEPT: TEXT/XML,APPLICATION/XML,APPLICATION/XHTML
+XML,";
$HEADER[0] .= "TEXT/HTML;Q=0.9,TEXT/PLAIN;Q=0.8,IMAGE/PNG,*/
*;Q=0.5";
$HEADER[] =  "CACHE-CONTROL: MAX-AGE=0";
$HEADER[] =  "CONNECTION: KEEP-ALIVE";
$HEADER[] = "KEEP-ALIVE: 300";
$HEADER[] = "ACCEPT-CHARSET: ISO-8859-1,UTF-8;Q=0.7,*;Q=0.7";
$HEADER[] = "ACCEPT-LANGUAGE: EN-US,EN;Q=0.5";
$HEADER[] = "PRAGMA: "; // BROWSERS KEEP THIS BLANK.


CURL_SETOPT($CURL, CURLOPT_USERAGENT, 'MOZILLA/5.0 (WINDOWS; U;
WINDOWS NT 5.2; EN-US; RV:1.8.1.7) GECKO/20070914 FIREFOX/
2.0.0.7');
CURL_SETOPT($CURL, CURLOPT_HTTPHEADER, $HEADER);
```

COOKIES AND SESSIONS

```
 curl_setopt($curl,CURLOPT_COOKIEJAR, $cookieJar);
curl_setopt($curl,CURLOPT_COOKIEFILE, $cookieJar);
```

AVOIDING GETTING BLOCKED

DO NOT DDOS

PROXY NETWORK

HAProxy

ACT LIKE A HUMAN BROWSING THE PAGE

```
curl_setopt($curl,CURLOPT_AUTOREFERER, true);
```

ROBOTS.TXT

LEGAL ISSUES

YOU ARE GOING TO GET SUED

MEASURE EVERYTHING

1. Response time
2. Response size
3. HTTP error type
4. Retries count
5. Failing proxy IP
6. Failing parsing
7. etc.

OPTIMIZE AND REPEAT

WEB CRAWLING FOR FUN AND PROFIT

THANKS!

Juozas Kaziukėnas
@juokaz