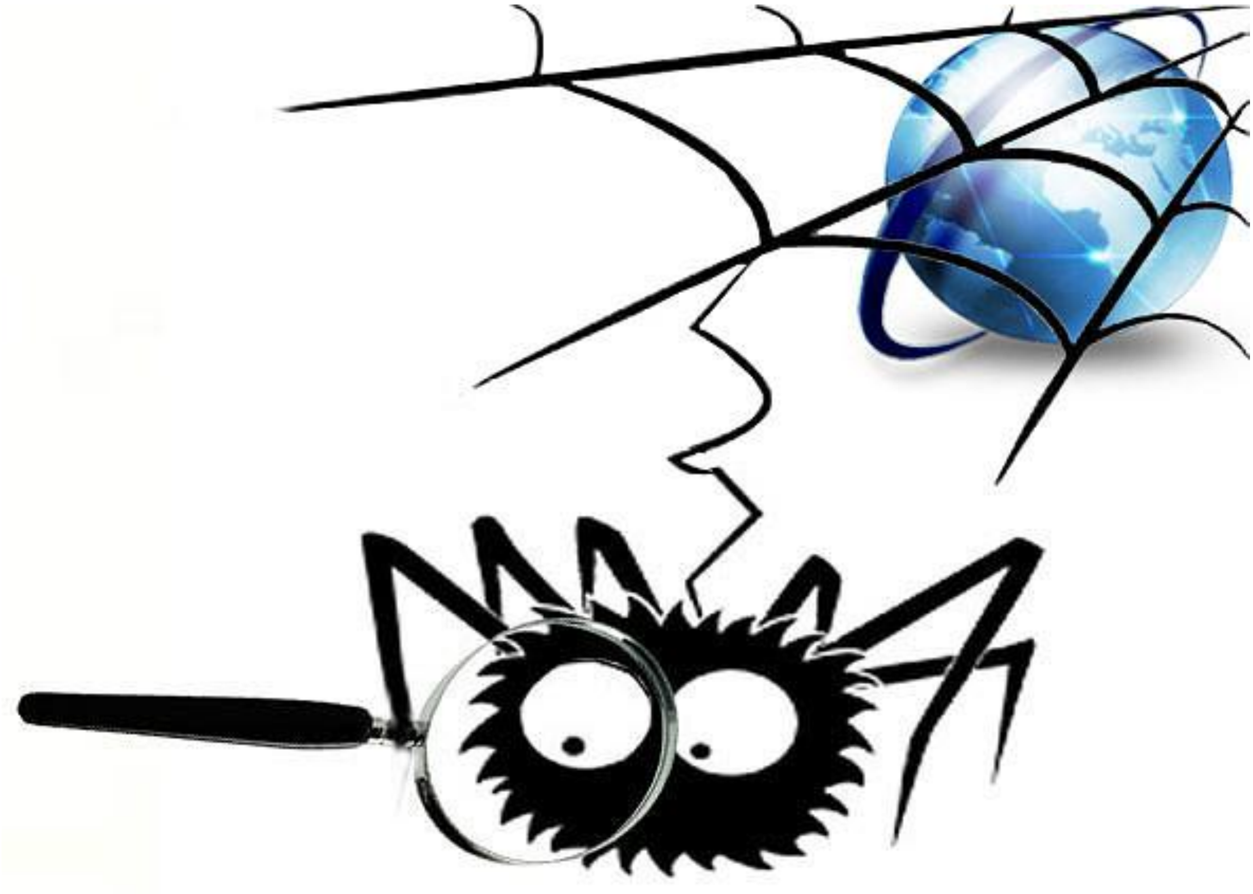
Search spider in a MySQL database



Features:

•The script parsing and search strings on domains and subdomains in large MySQL database

• Finding the line on the page type: htm, html, php, asp, jsp, js

• Line search all the pages belonging to that domain

• Search follows the tree structure, ie in depth and width. You can adjust the depth of search

• The ability to find a new domain on the page and make it into a database table for the next search. Thus, the database has the ability to expand itself

• If you find a string, the event is recorded and marked as *event*- *'found'*. And the event is fixed for the entire domain in its entirety. For example, if you have a domain **test.com** and pages *test.com/ index1.php*, *test.com/ index2.php*, and was found string 'test' on *test.com/ index1.php*, the *event-'found* 'will apply for **test.com**, and not to *test.com/ index1.php* page

• The ability to determine when a string on a domain was removed. This event is defined as an *event-'deleted'*

• Search is in tables with a large amount of data. When testing has been used 469,007 lines

• Do not consume a lot of memory.

• You can send the event *'found'*, *'deleted'* in the CRM system.

Project structure:

MySQL tables:

• *sites*- chart pattern on the basis of which the first search conducted

• *search_result*- result table in which the report is written, ie *events 'found'* and *'deleted'*.

• *string* - table with a row for the search

Files:

• **index.php -** main project file, it starts with the search and parsing data. After its launch table *search_result* wakes filled. The result of the script can be seen also in this way.

You can clear the search results table, typing the address **[domain] / index.php / clear**. For example **myspaider.com/index.php/clear**. In the field of recorded *date_check = NULL*

• **crm.php -** takes the data from the table *search_result* reports and trying to pass in the CRM according to API. Data is transmitted on the condition that if there was not once on this entry, or if the report does not say exceeded the time period (month).

If you need to add your API, you get the *public function api method ($ data){},* remove it lines 146 and 147, and transfer your data:

...

$domain - the domain name

$site_id - id domain

$date - the date when the event was committed

$path - the domain or page on which the event occurred

$string - string search

$event - type events found / deleted

...

You can clear the result in *search_result* table by typing the address **[domain] / crm.php / clear**. For example **myspaider.com/crm.php/clear**. In the field of recorded *send_crm = NULL*

• **helper.crm** - helper class to help write code in a structured object-oriented fashion. It is not recommended to change.

• **crm.data -** data file with which mimic sending CRM. Instead, it is assumed that there is a real CRM.

Installation of the project:

• Using phpMyAdmin, import **sql / parser file. sql**

• Set your time zone in the index.php file and crm.php date_default_timezone_set ('your zone');

• Set up a connection to the MySQL database in the **index.php** file and **crm.php**

```
...

public $ config_bd = array (

        'Server' => 'server name',

        'User_name' => 'user name in database',

        'User_pass' => 'user password',

        'Bd_name' => 'database name'

);

...
```

• Set the maximum number of iterations in the **index.php** file

```
private $max_iteration = 10;
```

The higher the value, the longer it takes the script to process data, but the greater the probability that the line is found.

The sum of all iterations is calculated by the total passage, ie a visit to any site, it is considered as one iteration.

• Set the number of records at a sample of the table

```
public $ limit = 500;
```

General algorithm:

• Runs the first way **index.php**

• Launch a second way **crm.php**

• **index.php** recursively searches in depth and width. To avoid stress on the script and the server, you can adjust the settings:

- ➢ the maximum number of entries on the haul

```
public $ limit = 500;
```

- ➢ the maximum number of iterations for the desired search string

```
private $ max_iteration = 5;
```

• **crm.php** portions trying to send to the CRM (**crm.data**) data. It is important for the settings:

- ➢ max amount to at sample entries

public $ limit = 500;

• **crm.data -** data file with which mimic sending CRM. Instead, it is assumed that there is a real CRM. The data in the file are separated by '|'.

- ➢ Sample lines from the file:

Find | - event *Find / Deleted*

site_id: 1852075 | - id domain

Domain: themeforest.net | - domain name

Page: themeforest.net | - the name of the page or domain, where the string was found

Find string: ThemeForest | - search string

Date: 2016-11-09 21: 40: 54 - time events

Note: It is recommended to run **crm.php** file for cron tasks scheduler, say 1 per hour.

Bud glad if the script and this information will be useful for you!

Yours Motoc Vladimir.

email: motoc_vladimir@mail.ru

skype: it_create