

Парсинг сайтов с помощью cURL и phpQuery

Кочитов Михаил Евгеньевич

Приамурский государственный университет им. Шолом-Алейхема

студент

Аннотация

В данной статье рассматривается парсинг сайтов с помощью технологии cURL и библиотеки phpQuery. Парсинг - это сбор необходимой информации с других источников, чтобы саму информацию отобразить на своем сайте. В статье также будет разработан собственный пример, где будет произведен парсинг текущего курса валют с финансового источника, используя язык программирования PHP.

Ключевые слова: парсинг, сбор информации, курс валют, cURL, phpQuery, PHP

Parsing sites using cURL and php Query

Kochitov Mikhail Evgenevich

Sholom-Aleichem Priamursky State University

student

Abstract

This article discusses the parsing of sites using cURL and php Query. Parsing is the collection of necessary information from other sources to display the information on your website. The article will also develop its own example, where the current exchange rate will be parsed from a financial source, using the programming language PHP.

Keywords: parsing, information gathering, currency exchange rate, cURL, phpQuery, PHP

Большинство сайтов в Интернете используют парсинг для сбора информации с других источников, например информацию пользователей, текущих новостей, точного времени, курса валют, цен на вебинары, семинары, курсы и так далее. Для парсинга используется язык программирования PHP с поддержкой технологии cURL, которая позволяет проводить взаимодействие с различными сервисами на различных протоколах: HTTP, FTP и другие. Технология cURL с заданными параметрами берет полностью всю веб страницу сайта, а для поиска определенного блока информации используется php библиотека под названием phpQuery. PhpQuery - это чистый аналог JavaScript библиотеки jQuery, которая имеет почти все функции, что и аналог, но phpQuery работает на серверной части, а jQuery - на клиентской части.

Целью данной статьи является парсинг сайтов с помощью технологии cURL и библиотеки phpQuery. Также будет разработан собственный пример парсинга финансового источника с целью получения текущего курса валют, используя язык программирования PHP.

В статье М.С. Малеванного приводится легковесный парсинг и его использование для функций среды разработки [1]. М.Д. Беззуб и О.И. Чуйко в статье рассмотрели парсинг сайтов как универсальное средство добавление контента [2]. Рассматривая статью Т.А. Абрамовой можно увидеть разработку парсинг-системы для получения скрытых ссылок со страниц социальных сетей [3]. В статье А.Н. Вильданова представляется парсинг HTML - документа на языке Java [4]. Д.Н. Курова и Д.В. Ушаева показывают анализ метода парсинга XML на языке программирования C# [5].

Далее приступим к написанию собственного примера с парсингом информации текущего курса валют с финансового источника. Для начала создадим файл и сохраним его в расширение .php, чтобы его сделать PHP скриптом, который будет обрабатывать веб сервер и брать информацию с другого источника.

Теперь напишем код в этом PHP скрипте, который представлен на рисунке ниже

```
<?php
    require_once 'phpQuery/phpQuery/phpQuery.php';

    $ch = curl_init('https://finance.rambler.ru/currencies/');
    curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
    curl_setopt($ch, CURLOPT_SSL_VERIFYPEER, false);
    curl_setopt($ch, CURLOPT_HEADER, false);
    $html = curl_exec($ch);
    curl_close($ch);

    // echo $html; // Выведет код страницы vk.com

    $pq = phpQuery::newDocument($html);

    $elem = $pq->find('.finance__currency-blocks');
    $text = $elem->html();
    // $text = iconv("windows-1251", "UTF-8", $elem);

    echo $text;
?>
```

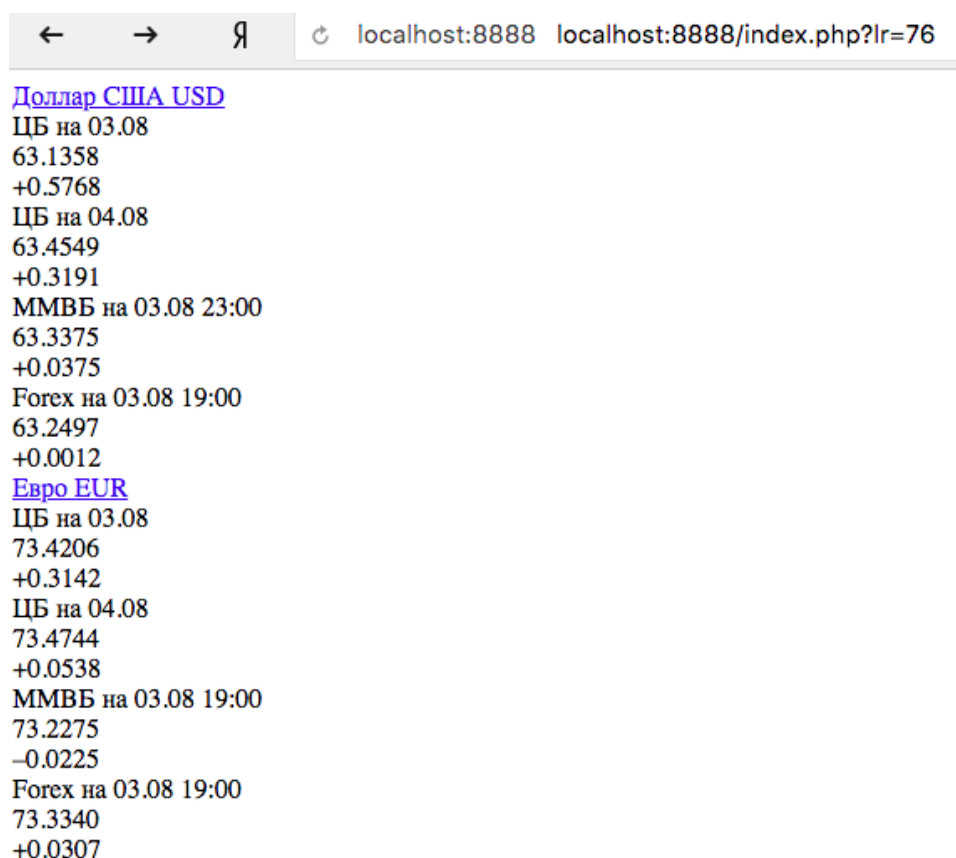
Рис. 1. PHP код парсинга текущего курса валют с финансового источника

На рисунке 1 представлен PHP код парсинга текущего курса валют с финансового источника. Теперь рассмотрим этот код немного подробнее. Функция `require_once` загружает php библиотеку “phpQuery” [6], которая является вытаскиванием нужной информации из сайта. Далее идет функция `curl_init` которая инициализирует cURL и подготавливает его к парсингу, указывая ему параметры через функцию `curl_setopt`. В значении функции `curl_init` указывается ссылка на источник, куда и будет происходить парсинг

для сбора оттуда нужной информации, в нашем случае это раздел финансов поисковика Rambler. Параметр `CURLOPT_RETURNTRANSFER` должен иметь значение `true`, чтобы после парсинга не отобразить всю страницу сайта. Дальше у параметра `CURLOPT_SSL_VERIFYPEER` должно стоять значение `false`, чтобы при обращении к удаленному серверу он не проверял сертификат. Последний параметр `CURLOPT_HEADER` отправляет удаленному серверу заголовки, значение `false` не дает этого сделать `cURL`. Теперь с помощью функцию `curl_exec` происходит исполнению `cURL` и подключение к сайту, чтобы взять у него всю веб страницу.

Далее идет функция `phpQuery::newDocument`, она записывает в переменную `$html` полностью всю HTML страницу с сайта. Функция `find()` ищет в HTML коде блок с классом `“finance_currency-blocks”`, в котором содержится информация о текущем курсе валют доллара и евро в рублях. Последняя функция `html()` преобразует найденный блок в `html` код, чтобы можно было его отобразить на странице браузера. Напоследок функция `echo` отображает кусок `html` кода с нужной информацией.

Теперь исполним готовый PHP скрипт в браузере и посмотрим, что он выдаст



```
← → Я ↻ localhost:8888 localhost:8888/index.php?lr=76
Доллар США USD
ЦБ на 03.08
63.1358
+0.5768
ЦБ на 04.08
63.4549
+0.3191
ММВБ на 03.08 23:00
63.3375
+0.0375
Forex на 03.08 19:00
63.2497
+0.0012
Евро EUR
ЦБ на 03.08
73.4206
+0.3142
ЦБ на 04.08
73.4744
+0.0538
ММВБ на 03.08 19:00
73.2275
-0.0225
Forex на 03.08 19:00
73.3340
+0.0307
```

Рис. 2. Результат работы парсинга

Как видно на рисунке 2 изображен результат работы PHP скрипта, который взял информацию о текущем курсе валют доллара и евро из

финансового источника. Сейчас курс доллара равен 63 рубля, а курс евро - 73 рубля.

Таким образом был рассмотрен парсинг с использованием технологии cURL и библиотеки phpQuery на языке программирования PHP. Также был разработан собственный пример парсинга текущего курса валют с финансового источника. Можно предположить, что парсинг все же нужен множеству сайтов для получения информации с других удаленных веб-источников, чтобы разнообразить содержимое сайта дополненной информацией.

Библиографический список

1. Малеванный М.С. Легковесный парсинг и его использование для функций среды разработки // Информатизация и связь. 2015. № 3. С. 89-94.
2. Беззуб М.Д., Чуйко О.И. Парсинг сайтов как универсальное средство добавления контента // В сборнике: Современные тенденции и проекты развития информационных систем и технологий Материалы Всероссийской научно-исследовательской конференции студентов и школьников. Хабаровский государственный университет экономики и права. 2016. С. 292-296.
3. Абрамова Т.А. Разработка парсинг-системы для получения скрытых ссылок со страниц социальных сетей // Вестник Пензенского государственного университета. 2016. № 3 (15). С. 41-47.
4. Вильданов А.Н. Парсинг HTML-документа на языке Java (на примере расписания НФ БАШГУ) // В сборнике: Достижения и приложения современной информатики, математики и физики материалы V Всероссийской научно-практической заочной конференции. 2016. С. 17-24.
5. Курова Д.Н., Ушаева Д.В. Анализ методом парсинга XML на языке программирования C# // Вестник Димитровградского инженерно-технологического института. 2017. № 1 (12). С. 69-73.
6. Google Code phpQuery URL: <https://code.google.com/archive/p/phpquery/> (дата обращения 05.08.2018)
7. Примеры использования cURL в PHP URL: <http://snipp.ru/view/63> (дата обращения 05.08.2018)
8. 8 примеров использования cURL вместе с PHP URL: <https://ruseller.com/lessons.php?rub=37&id=1370> (дата обращения 05.08.2018)