Абрамова Т.А. Парсинг как средство для получения скрытых ссылок со страниц социальных сетей. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей XVIII Междунар. научно-техн. конф. – Пенза: ПДЗ, 2018. – С. 234-237.

УДК 004.624

ПАРСИНГ КАК СРЕДСТВО ДЛЯ ПОЛУЧЕНИЯ СКРЫТЫХ ССЫЛОК СО СТРАНИЦ СОЦИАЛЬНЫХ СЕТЕЙ

Т.А. Абрамова

PARSING AS A MEANS OF GETTING HIDDEN LINKS FROM SOCIAL NETWORKING PAGES

T.A. Abramova

Аннотация. В статье обсуждаются вопросы парсинга страниц социальных сетей. С помощью стандартных средств серверного языка программирования PHP разобран пример создания парсера, получающего скрытые ссылки на видеофайлы со страниц социальных сетей.

Ключевые слова: парсинг, web-приложение, web-сервер, скрипт, регулярные выражения, PHP.

Abstract. The article discusses the issues of parsing social networking pages. Using standard tools of the server-side PHP programming language, an example of creating a parser that receives hidden links to video files from social networking pages has been analyzed.

Keywords: parsing, web application, web server, script, regular expressions, PHP.

Перед web-мастером или контент-менеджером сайта зачастую встает задача получения достаточно больших объемов информации с удаленных серверов. При этом имеют существенное значение несколько параметров. Во-первых, объем данных может быть очень большим, и, возможно, информацию необходимо будет получать не с одного сайта, а с нескольких. Во-вторых, имеет значение время. Поэтому мы должны обновлять контент нашего сайта максимально быстро и эффективно. Из этого формируется третье условие – организация автоматизированного сбора данных, при котором человек выполняет только контролирующую функцию.

Есть стандартные программные продукты, позволяющие эффективно и быстро проводить парсинг различных сайтов. Одними из лучших считаются Content Downloader и ZennoPoster. У них много достоинств: простое добавление страниц для парсинга, автоматический поиск контента, многопоточность, фильтрация текста, импорт в популярные CMS и т.д. Но есть один немаловажный и существенный минус – программы платные.

Попробуем разработать свой парсер. Существует так называемый «джентльменский набор web-разработчика»: HTML, PHP, MySQL, JavaScript и CSS [1]. Работать этот набор может в сочетании с web-сервером Apache. Все описанное выше программное обеспечение является свободно распространяемым и доступно к скачиванию в сети. В настоящее время очень популярным для разработчиков тандемом являются PHP и MySQL. PHP прост в использовании, в нем допустимы ссылки на программу базы данных MySQL, он позволяет легко создавать на web-сайтах динамические

элементы и отлично вписывается в HTML-код [2]. Для нас использование языка PHP важно тем, что он имеет в своем составе мощные инструменты для работы с регулярными выражениями [3].

Теперь перейдем непосредственно к решению нашей задачи. Социальная сеть «ВКонтакте» (https://vk.com) предоставляет нам множество изображений, аудио- и видеофайлов для просмотра. Но не все из них доступны для прямого скачивания. Здесь может быть несколько различных ситуаций. Например, видео находится на одном из внешних серверов, например на очень популярном сейчас видеохостинге «YouTube» (https://www.youtube.com), называемом русскоязычными пользователями Интернета «ютуб».

Нам интересен случай, когда видео хранится на сервере «ВКонтакте». Для получения такой ссылки нужно воспользоваться специальными средствами [4].

Попробуем самостоятельно с помощью изучения кода нужного элемента выделить ссылку на видеофайл нужного качества.

Для разбора кода и написания парсера нам понадобятся следующие части кода:

1) url240=https%3A%2F%2Fcs508301.vk.me%2F3%2Fu51591817%2Fvideos%2Fa429dc616a.240.mp4

2) url360=https%3A%2F%2Fcs508301.vk.me%2F3%2Fu51591817%2Fvideos%2Fa429dc616a.360.mp4

3) url480=https%3A%2F%2Fcs508307.vk.me%2F3%2Fu51591817%2Fvideos%2Fa429dc616a.480.mp4

4) url720=https%3A%2F%2Fcs508307.vk.me%2F3%2Fu51591817%2Fvideos%2Fa429dc616a.720.mp4

Всю основную работу по разбору текста выполняет следующее регулярное выражение:

/(url720=)(https%3A%2F%2F\w+\.vk\.me%2F.+%2F\w+%2Fvideos%2F\w+\.720\.mp4)/Ui

Результат разбора текста (проведенный с помощью собственной функции [4]) в виде массива из трех строк можно увидеть на рис. 1.

```
Array
(
    [0] => Array
        (
            [0] => url720=https%3A%2F%2Fcs508307.vk.me%2F3%2Fu51591817%2Fvideos%2Fa429dc616a.720.mp4
            [1] => url720=
            [2] => https%3A%2F%2Fcs508307.vk.me%2F3%2Fu51591817%2Fvideos%2Fa429dc616a.720.mp4
        )
```

Рис. 1. Результат обработки текста в окне браузера

Мы убедились в том, что разбор файла работает корректно и искомая гиперссылка получена нами в третьем элементе массива (с индексом «2»), выданного на экран. Но есть небольшая хитрость. Некоторые элементы строки заменены на спецсимволы. Например, «%3A» – это двоеточие, а «%2F» - прямой слеш. Чтобы отобразить на экране корректную ссылку, напишем еще одну функцию, производящую нужные замены.

На рис. 2 можно увидеть результат работы созданной нами функции, которая возвращает корректную гиперссылку в привычном нам формате.

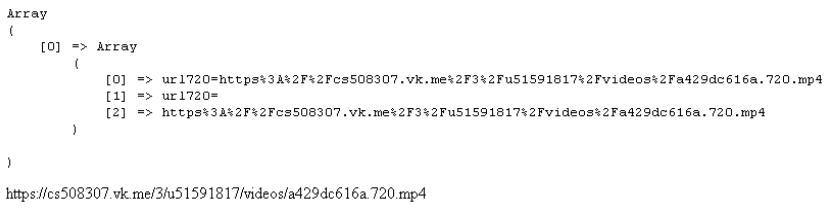

Рис. 2. Результат работы функции в окне браузера

Скопировав полученную ссылку в любой менеджер закачек, например, бесплатный «Download Master», произведем копирование видеофайла на жесткий диск компьютера. Аналогичного результата мы достигнем, если сохраним файл непосредственно в окне браузера. По этой же схеме можно выполнить разбор текста, создав регулярные выражения для закачки видео в нужном нам качестве, выбрав любой вариант из предложенных выше.

Библиографический список

1. Прохоренок П.А. HTML, JavaScript, PHP и MySQL. Джентльменский набор Web-мастера / П. А. Прохоренок, В.А. Дронов. 4-е изд., перераб и доп. СПб.: БХВ-Петербург, 2015. 768 с.

2. Никсон Р. Создаем динамические веб-сайты с помощью PHP, MySQL, JavaScript и CSS. 2-е изд. СПб.: Питер, 2013. 560 с.

3. Янк К. PHP и MySQL. От новичка к профессионалу. М.: Эксмо, 2013. 384 с.

4. Абрамова Т.А. Разработка парсинг-системы для получения скрытых ссылок со страниц социальных сетей // Вестник Пензенского государственного университета. 2016. № 3 (15). С. 41-47

Абрамова Татьяна Алфиевна
Пензенский государственный университет,
г. Пенза, Россия
E-mail: abramova_ta@mail.ru

Abramova T.A.
Penza State University,
Penza, Russia