МОДУЛЬ ПОЛУЧЕНИЯ ДАННЫХ ИЗ ВНЕШНИХ ОТКРЫТЫХ ИСТОЧНИКОВ

М.В. Стержанов
Студент кафедры информатики БГУИР

Д.Н. Рожков
Студент кафедры информатики БГУИР

В.Ю. Пресняцкий
Студент кафедры информатики БГУИР

П.Е. Дорошкевич
Студент кафедры информатики БГУИР

А.И. Свито
Студент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: sterjanov@bsuir.by, rdimon2912@gmail.com, presniatski@gmail.com, dpavluha@gmail.com, alexandervirk@gmail.com

Abstract. Web-crawlers (also known as robots or scrapers) enable the process by following the hyperlinks in web pages to automatically download a fractional snapshot of the web site. This paper describes developed web crawler named MMScraper aimed to process informational web resources for further getting statistical properties and performing data analytics.

В настоящее время в связи с бурным развитием сети Интернет наблюдается обилие электронной неструктурированной информации, представленной текстами на естественных языках. Всё более востребованной становится задача автоматической обработки таких текстов с целью извлечения структурированных данных, которые затем используются при решении различного рода проблем: извлечения фактических данных, поиска информации и т.п. Нами решается задача обработки контента информационно-новостных ресурсов с целью анализа лексико-терминологической информации.

Для сбора требуемых данных требуются специализированные инструменты - поисковые роботы, также называемые «веб-пауками» (web-spider), краулерами (web crawler) или скребками (web scraper). Поисковый робот — программный комплекс, осуществляющий навигацию по веб-ресурсам и сбор информации для базы данных приложения-агента [1, 2].

Нами планируется значительная работа по обследованию ряда информационных сайтов, чтобы собрать выборку данных требуемого размера. Анализ имеющихся в свободном доступе решений показал, что открытые реализации зарубежных веб-краулеров слабо приспособлены к решаемой нами задаче, так как требуют весьма трудоемкой настройки, а после нее показывают низкую производительность и существенно нагружают информационный источник. В связи с этим было принято решение разработать собственное решение.

Опишем основные требования, в соответствии с которыми был разработан краулер,

названный MMScraper.

– Получать в качестве исходных данных список доменных имен сайтов, предназначенных для сканирования. Предполагается, что имеется некоторое множество заранее определенных для исследования сайтов.

– Обходить каждый сайт, начиная с главной (индексной) страницы, перемещаясь по внутренним гиперссылкам в заданном порядке обхода «вначале вширь».

– Полученные результаты сохранять в базу данных. Интерес представляют следующие атрибуты: адрес страницы, автор публикации, дата публикации, содержимое публикации.

– Позволять получать данные с сайта через программный интерфейс API.

– Иметь расширяемую архитектуру для последующего развития функциональности.

– Добавление нового сайта должно быть простым и не требовать привлечения квалифицированного программиста.

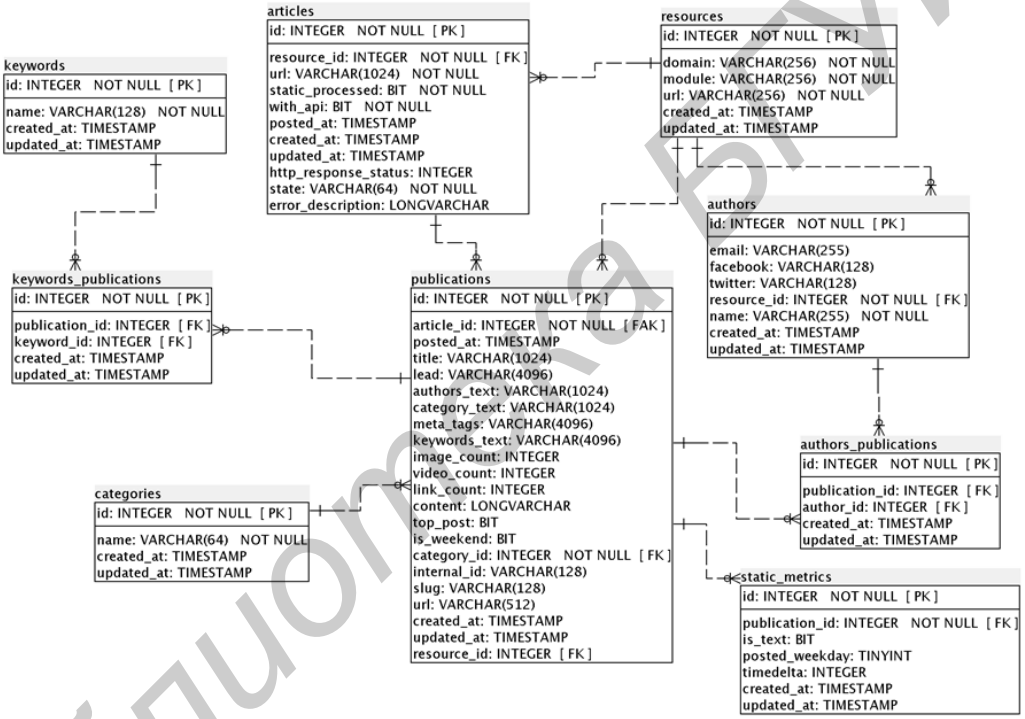

Рис. 1. Модель базы данных

Работу разработанного краулера можно описать следующим образом: сканирование сайта начинается с начальной страницы и затем робот использует ссылки, размещенные на ней, для перехода на другие страницы. Каждая страница сайта анализируется на наличие требуемой информации, которая копируется в соответствующее хранилище в случае обнаружения. Процесс повторяется до тех пор, пока не будет проанализировано требуемое число страниц либо пока не будет достигнута некая цель. Модуль получения данных разработан на языке программирования Ruby и состоит из трех основных частей: блок сканирования и обработки данных, блок управления краулером (команды вводятся через консоль) и база данных. Собираемая роботом информация состоит из ссылочной структуры обрабатываемого ресурса и веб-страниц. В качестве основы для базы данных была выбрана бесплатная СУБД MySQL. Для упрощения взаимодействия с БД нами используется библиотека Sequel, позволяющая представлять данные в виде объектов.

Рассмотрим схему базы данных, содержащую полученную информацию.

Таблица *resources* описывает веб-сайты, которые подлежат краулингу. Атрибут *module* сообщает какой шаблон отвечает за разбор полученной страницы и выделения полей, необходимых для сохранения в БД.

Таблица *articles* содержит ссылки на страницы сайта, которые подлежат скачиванию. Таблица *publications* представляет информацию, полученную путем разбора целевых страниц сайта. Как видно их схемы данных, мы храним заголовок, аннотацию и текстовое содержание документа. Каждая публикация принадлежит категории (таблица categories), имеет ключевые слова (таблица keywords), и написана одним или несколькими авторами (таблица authors). Помимо этого, мы подсчитываем число изображений, видео и ссылок, содержащихся на странице.

В работе приводится описание основных требований, общей архитектуры и конфигурации краулера MMScraper, предназначенного для решения достаточно узкой, но важной задачи, а именно – сбора информации о новостных и информационно-аналитических публикациях.

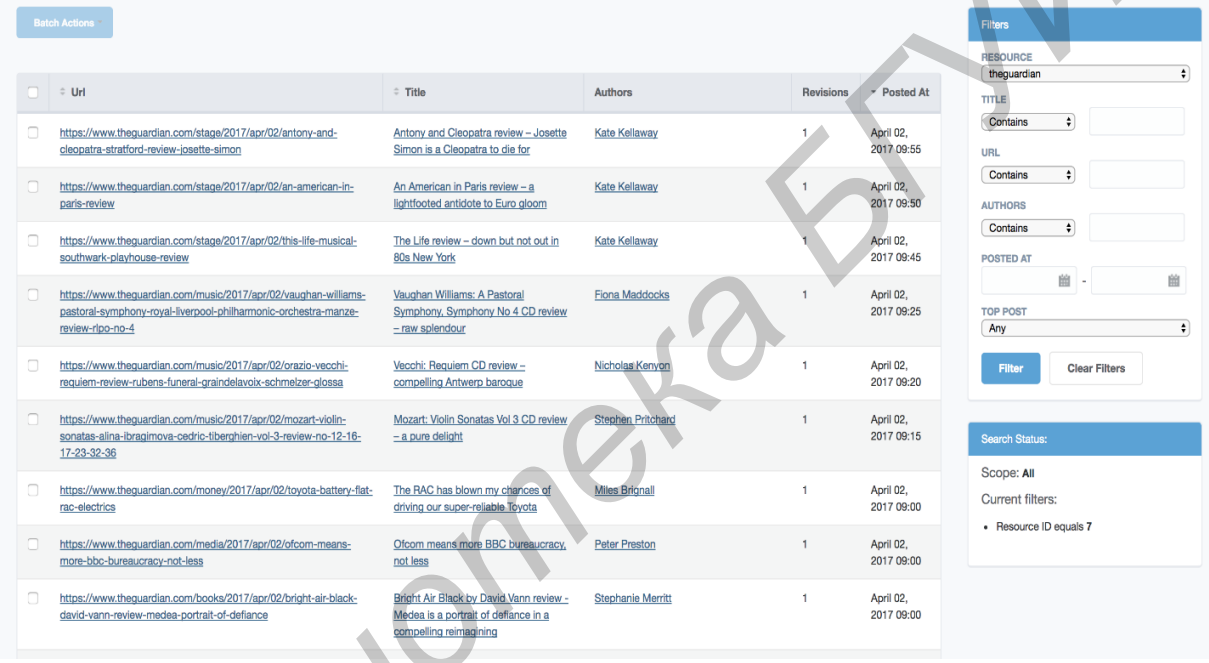

Рис. 2. Графический интерфейс пользователя

В практическом плане разработанный MMScraper позволит собрать экспериментальную базу для исследования задач интеллектуальной обработки текста.

Нам видится важным продолжить исследования результатов краулинга и реализации дополнительных возможностей MMScraper, улучшающих результаты работы на очень больших сайтах. Реализация таких возможностей предусмотрена расширяемой архитектурой краулера.

Литература

[1]. A.H.F. Laender, B. A. Ribeiro-Neto, Juliana S.Teixeria. A brief survey of web data extraction tools // ACM SIGMOD Record 31(2), pp 84-93. 2002

[2]. Baeza-Yates R., Castillo C. Crawling the Infinite Web: Five Levels are Enough // Lecture Notes in Computer Science. Algorithms and Models for the Web-Graph, Third International Workshop. 2004. Vol. 3243. P. 156–167.