МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра информационных технологий**

**ДИПЛОМНАЯ РАБОТА**

**Применение синтаксического анализа новостного сайта с возможностью автопостинга записей на индивидуальный сайт**

Работу выполнил\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Д.Х. Чан

(подпись)

Направлению подготовки 01.03.02 – «Прикладная математика и информатика» курс 4

Направленность (профиль) «Системное программирование и компьютерные технологии» (Математическое и программное обеспечение вычислительных машин)

Научный руководитель

канд. техн. наук, доц. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_А.Н. Полетайкин

(подпись, дата)

Нормоконтролер ст. преп. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_А.В. Харченко

(подпись, дата)

Краснодар

2020

**РЕФЕРАТ**

Ключевые слова: ВЕБ-СКРАПИНГ, ПАРСИНГ, СБОР ИНФОРМАЦИИ

Цель работы - исследовать основные способы веб-скрейпинга для поиска информации и реализовать соответствующее приложение.

В дипломной работе изучены принципы технологии веб-скрапинга, понятие поискового робота и способы применения синтаксического анализа.

В дипломной работе разработана программа, позволяющая получить всю необходимую информацию с новостных сайтов городов Краснодар, Сочи и Новороссийск и сохранить эти данные в документо-ориентированной базе данных.

Реализация выполнена на языке программирования JavaScript с использованием библиотек платформы NodeJS.

**СОДЕРЖАНИЕ**

[**ВВЕДЕНИЕ** 4](#_Toc39709209)

[**1 Основные понятия технологии веб-скрапинга** 6](#_Toc39709210)

[**1.1** **Технологии веб-скрапинга** 6](#_Toc39709211)

[**1.2** **Краулинг** 12](#_Toc39709212)

[1.2.1 Основные сведение о краулерах 12](#_Toc39709213)

[1.2.2 Файл Robots.txt 13](#_Toc39709214)

[**1.3** **Парсинг** 15](#_Toc39709215)

# **ВВЕДЕНИЕ**

Сегодня, в эпоху бурного развития информационных технологий, знания и информация являются главными факторами успеха в обществе. Благодаря сети Интернет можно получить доступ практически к любой информации, общаться с кем-либо еще практически из любой точки в мире и делать гораздо больше. Все это можно сделать, подключив компьютер, телефон или ноутбук к Интернету. Хотя, формально Интернет представляет собой глобальную сеть из миллиардов компьютеров и других электронных устройств, Интернет становится все более важной частью повседневной жизни людей во всем мире. На сегодняшний день Интернет является самым большим источником информации для человека. Это источник, который ежедневно пополняется все новой и новой важной и полезной информацией. Поэтому человеку овладеть таким количеством информации достаточно сложно. Технология веб-скрапинга может ему в этом помочь.

Веб-скрапинг – это термин, обозначающий различные методы, используемые для сбора информации через Интернет. Как правило, это делается с помощью программного обеспечения, которое имитирует переход человека на сайт для сбора определенной информации по определенным параметрам. Благодаря ей можно получать, обрабатывать, преобразовывать и сохранять в заданном формате нужные вам данные с множества различных веб-страниц, не тратя на это много времени, как если бы это делал человек самостоятельно переходя на каждый сайт и вручную копировал и сохранял нужные данные. Нельзя исключать, что человек может также ошибиться. Вместо этого можно воспользоваться скриптами для получения информации. Возможности использования веб-скрапинга для деловых и личных потребностей бесконечны. Особым спросом эта технология пользуется в журналистике и статистике, что поддерживает её актуальность.

Целью работы является изучение технологии веб-скрапинга на основе библиотек платформы Node.JS и поиск важной и актуальной информации.

Итог проделанной работы – создание приложения с использованием технологии веб-скрейпинга.

# **1 Основные понятия технологии веб-скрапинга**

## **Технологии веб-скрапинга**

Веб-скрапинг — это технология, используемая для извлечения данных с веб-сайтов с помощью скрипта. Она помогает автоматизировать трудоемкую работу по копированию данных с различных веб-сайтов [1]. Эта технология является частью более общего понятия web mining.

Web mining (WM) — программное обеспечение, предназначенное для извлечения знаний из данных, обычно из документов и сопровождающих их гиперссылок, из сведений о пользователях и их активности, зафиксированных в разного рода журналах, в общем, из всего, что может быть доступно в Сети [2]. Основной целью веб-майнинга является обнаружение полезной информации из Всемирной паутины и ее моделей использования. Первые публикации на тему WM датируются серединой 90-х годов, когда мирно сосуществовали два подхода — процессный, представляющий WM как последовательность выполняемых задач, и дата-центричный, привязанный к типам данных. В дальнейшем преимущественное распространение получил второй.

Применение технологии веб-майниг:

1) Веб-майнинг помогает улучшить возможности веб-поисковой системы, классифицируя веб-документы и идентифицируя веб-страницы.

2) Он используется для веб-поиска, например, Google, Yahoo и т. д., а также для вертикального поиска, например, FatLens, Become и т. д.

3) Веб-майнинг используется для прогнозирования поведения пользователя.

4) Веб-майнинг очень полезен для конкретного веб-сайта и электронного сервиса, например, для оптимизации целевой страницы.

Сегодня под WM понимают совокупность методов автоматического систематического обхода сети интернет с целью сбора требуемых данных о компаниях и людях для выработки информации и принятия решений. В WM обычно используются публичные данные и нет необходимости в правах доступа, но следует соблюдать определенные этические нормы и уважать труд других людей. Хотя краулеры (поисковые программы-роботы), распространяемые поисковыми машинами, создают для сайтов дополнительную нагрузку, администраторы им не противодействуют, поскольку в результате создаются индексы, способствующие увеличению посещаемости. Однако WM не придает сайтам дополнительных качеств, поэтому краулеры должны быть более тактичны.

Рисунок 1 показывает три основные части, которые можно выделить в Web Mining.

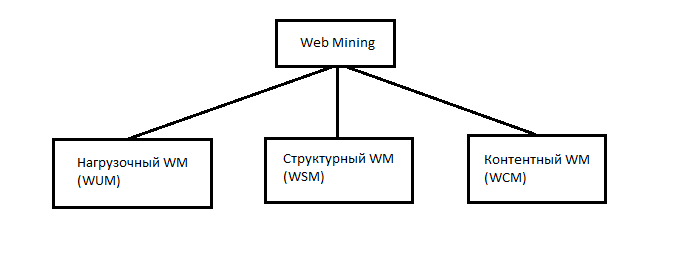


Рисунок 1 – Составные части WM.

Дата-центричность означает акцент WM на работе с тремя основными источниками информации:

* Данные о действиях пользователей, от журналов серверов до отслеживания обращений к браузеру. Так же называется Web Usage Mining (WUM) или нагрузочный WM. Многочисленные данные о пользователях собираются на сторонах пользователя и сервера. Анализируя эти сведения, можно получать обобщенные данные, обнаружить и изучить закономерности поведения групп пользователей и информацию о посетителях сайта. Помимо владельцев браузеров, сбором такого рода данных занимаются сотни специализированных компаний, образуя крупный сегмент бизнеса.
* Веб-графы, описывающие прямые связи между страницами WWW. В веб-графе страницы всемирной паутины — это вершины, а гиперссылки между ними — это дуги. По графу устанавливаются связи между страницами, людьми и любыми иными объектами. Также известный как Web Structure Mining (WSM) или структурный WM. WSM в основном показывает структурированное резюме определенного веб-сайта. Он идентифицирует отношения между веб-страницами, связанными информацией или прямой связью. Для определения связи между двумя коммерческими веб-сайтами может быть очень полезен анализ структуры сети.
* Контент веб-страниц и связанных с ними документов. Также известный как Web Content Mining (WCM) или контентный WM. WCM — это приложение для извлечения полезной информации из содержимого веб-документов, идентификация, сопоставление и интеграция семантически похожих данных, извлечение мнений из онлайн-источников и иерархия понятий, онтология или интеграция знаний. Веб-контент состоит из нескольких типов данных: текстовых, графических, аудио, видео и т. д. Контентные данные — это группа фактов, разработанных веб-страницей. Он может предоставить эффективные и интересные шаблоны о потребностях пользователей. Текстовые документы связаны с добычей текста, машинным обучением и обработкой естественного языка. Этот майнинг также известен как текстовый майнинг. Этот тип интеллектуального анализа выполняет сканирование и интеллектуальный анализ текста, изображений и групп веб-страниц в соответствии с содержанием входных данных. Основные сложности в исследование WCM является неоднородность веб-данных. Очень часто в WCM выполняется процедура перевода данных из формы, предназначенной для чтения человеком, в форму, удобную для обработки компьютером. Эта процедура называется data scraping. Со временем эта технология модифицировалась и стала называться Web Scraping.

Веб-скрейпинг работает с тремя типами информации: структурированными, неструктурированными и квазиструктурированными.

Структурированные данные в Интернете часто очень важны. Структурированные данные также легче извлечь по сравнению с неструктурированными текстами. Для работы со структурированными данными достаточно применить служебные процедуры сначала обхода страниц, затем генерации и исполнения упаковщика, а дальше можно переходить к анализу содержимого страницы.

Сложнее с квазиструктурированными данными. Примером таких данных может быть граф посещений сайта. Для работы с квазиструктурированными данными предложены специальные языки, такие как NEXIR ELOG, предназначенные для программ-упаковщиков. С их помощью описывается процедура выделения данных, завершающаяся созданием объектной модели данных (OEM).

Данные веб-контента представляют собой большую часть неструктурированных текстовых данных. Исследования по применению методов интеллектуального анализа данных к неструктурированному тексту называются «анализ текстовых данных» или «анализ текста». Неструктурированный DM призван облегчить восприятие пользователям больших массивов текстов. Выделяют несколько типов такого рода операций:

* отслеживание тематики – оценка области интересов пользователя;
* свертка – создание резюме документов;
* ранжирование – упорядочение документов и их распределение по заранее определенным категориям;
* кластеризация – объединение схожих документов в группы;
* визуализация данных – решение проблемы коммуникации пользователя с данными.

Идеального веб-скрапера не существует по нескольким причинам [3]:

1. Доступ. Первое, что нужно проверить, это то, что разрешает ли ваш целевой веб-сайт скрапинг, прежде чем вы запустите его. Если вы обнаружите, что он запрещает скрапинг через файл robots.txt, вы можете попросить владельца веб-сайта предоставить разрешение на скрапинг, объяснив ваши потребности и цели получения данных. Если владелец все еще не согласен, лучше найти альтернативный сайт, на котором есть аналогичная информация.
2. Сложные и изменчивые структуры веб-страниц. Большинство веб-страниц основаны на HTML (язык гипертекстовой разметки). Дизайнеры веб-страниц могут иметь свои собственные стандарты для дизайна страниц, поэтому структуры веб-страниц сильно различаются. Кроме того, веб-сайты периодически обновляют свой контент для улучшения взаимодействия с пользователем или добавления новых функций, что часто приводит к структурным изменениям на веб-странице. Поскольку веб-скраперы настроены в соответствии с определенным дизайном страницы, они не будут работать для обновленной страницы. Иногда даже незначительное изменение на целевом веб-сайте требует настройки скрапера.
3. Блокировка IP. Блокировка IP-адресов — это распространенный метод предотвращения доступа веб-скраперов к данным веб-сайта. Обычно это происходит, когда веб-сайт обнаруживает большое количество запросов с одного и того же IP-адреса. Веб-сайт либо полностью запретит IP, либо ограничит его доступ, чтобы нарушить процесс скрапинга.
4. CAPTCHA. Captcha – это полностью автоматизированный публичный тест Тьюринга для различения компьютеров и людей. Он часто используется для отделения людей от инструментов скрапинга, отображая изображения или логические проблемы, которые люди могут легко решить, а веб-скраперы – нет.
5. «Горшочек с мёдом». Это ресурс, представляющий собой ловушку для злоумышленников. Ловушка, которую владелец сайта помещает на страницу, чтобы ловить скраперов. Ловушки могут быть ссылками, которые невидимы для людей, но видны для веб-скраперов. Как только скраперов попадает в ловушку, веб-сайт может использовать полученную информацию (например, свой IP-адрес), чтобы заблокировать этот скрапер.
6. Медленная или нестабильная скорость загрузки. Веб-сайты могут реагировать медленно или даже не загружаться при получении слишком большого количества запросов на доступ. Это не проблема, когда люди просматривают сайт, поскольку им просто нужно перезагрузить веб-страницу и ждать, пока сайт восстановится. Но скрапер может сломаться, поскольку он не знает, как справиться с такой чрезвычайной ситуацией.
7. Динамический контент. Многие веб-сайты применяют AJAX для обновления динамического веб-контента. Примерами являются ленивая загрузка изображений, бесконечная прокрутка и отображение дополнительной информации при нажатии кнопки через вызовы AJAX.
8. Требование входа. Некоторая защищенная информация может потребовать от вас входа в систему. После того, как вы отправите свои учетные данные, ваш браузер автоматически добавляет значение cookie в несколько запросов, которые вы делаете, поэтому веб-сайт знает, что вы тот же человек, который только что вошел в систему ранее. Поэтому при скрапинге веб-сайтов, требующих входа, убедитесь, что файлы cookie были отправлены вместе с запросами.
9. Сканирование данных в режиме реального времени. Очистка данных в реальном времени необходима, когда дело доходит до сравнения цен, отслеживания запасов и т. д. Данные могут измениться в считанные секунды. Веб-скрапер должен постоянно отслеживать сайты и получать данные. Несмотря на это, он все еще имеет некоторую задержку, так как запрос и получение данных занимают время. Кроме того, получение большого объема данных в режиме реального времени также является большой проблемой.

Перечисленные выше факторы сильно затрудняют процесс веб-скрапинга. Однако все инструменты для Web Content Mining имеют довольно схожую структуру.

Практически любой инструмент предназначенный для web-scrapping включает в себя следующие элементы:

* краулинг;
* парсинг;
* хранение информации.

## **Краулинг**

### Основные сведение о краулерах

Поисковый робот или краулер — это программа для сканирования и индексации сайтов. Он позволяет поисковой системе получить сведения о веб-страницах и внести их в базу для последующей выдачи пользователям при запросе [4]. Поисковых роботов называют по-разному: краулеры, боты, веб-пауки, пауки. Краулинг позволяет автоматизировать такие функции, как сбор особых видов информации, например, адресов электронной почты. Существует два типа Web Crawler, которые называются внешний и внутренний веб-краулер. Внешний паук сканирует неизвестные сайты. Внутренний сканер сканирует внутренние страницы сайта, которые возвращаются внешним пауком. Функционирование поискового робота строится по тем же принципам, по которым работает браузер. Боты запускаются с начального источника или из списка известных URL-адресов. Сначала он сканирует веб-страницы по этим URL. Когда он закончит сканировать эти веб-страницы, он находят гиперссылки на другие URL-адреса и добавляют их в список страниц для последующего сканирования.

Результатом является перебор веб-ресурсов в строгой последовательности.

Учитывая огромное количество веб-страниц в Интернете, которые можно проиндексировать для поиска, этот процесс может продолжаться почти бесконечно. Однако краулер будет следовать определенным политикам, которые делают его более избирательным в отношении того, какие страницы сканировать, в каком порядке их сканировать и как часто они должны сканировать их снова, чтобы проверить наличие обновлений контента. Большинство краулеров не сканируют весь общедоступный Интернет и не предназначены для этого; вместо этого они решают, какие страницы сначала сканировать, основываясь на количестве других страниц, которые ссылаются на эту страницу, количестве посетителей этой страницы и других факторах, которые указывают на вероятность того, что страница содержит важную информацию. Идея состоит в том, что веб-страница, на которую ссылаются многие другие веб-страницы и которая посещает большое количество посетителей, скорее всего, будет содержать высококачественную, авторитетную информацию. Поисковые роботы также решают, какие страницы сканировать, основываясь на протоколе robots.txt (также известном как протокол исключения роботов). Перед сканированием веб-страницы они проверят файл robots.txt, размещенный на веб-сервере этой страницы, и дальнейшие поиски должен проводить согласно правилам, установленным в robots.txt.

### Файл Robots.txt

Файл robots.txt или индексный файл — обычный текстовый документ в кодировке UTF-8, действует для протоколов http, https, а также FTP. Если файл будет содержать символы не в UTF-8, а в другой кодировке, поисковые роботы могут неправильно их обработать. Правила, перечисленные в файле robots.txt, действительны только в отношении того хоста, протокола и номера порта, где размещен файл [5]. Эти правила определяют, какие страницы могут сканировать боты, и по каким ссылкам они могут переходить. Правила включают в себя:

* какие разделы сайта закрыты и открыты для роботов;
* интервалы, которые робот обязан выдерживать между последовательными запросами. Некоторым веб-сайтам может понадобиться использовать этот параметр, если боты поглощают трафик и замедляют загрузку вашего сайта для обычных посетителей.
* расположение карты сайта (sitemap.xml), что может сделать ваш контент более доступным.

Серия правил может относиться ко всем без исключений краулерам или же к какому-то определенному [6]. На рисунке 2 представлена часть файла robots.txt от Яндекса. Полностью его можно посмотреть, открыв http://yandex.ru/robots.txt.

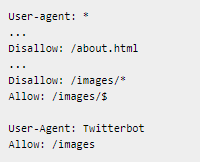


Рисунок 2 – Часть файла robots.txt от Яндекса

Это означает, что:

1. ни одному из краулеров не разрешается заходить на страницу /about.html;
2. боту, идентифицирующему себя как Twitterbot, разрешается просматривать любой ресурс, чей адрес содержит сегмент /images;
3. остальным разрешено посещать адреса, заканчивающиеся на /images/, и запрещено спускаться ниже.

После обращения к роботу, с помощью директивы user-agent, следуют команды, запрещающие или разрешающие обращение к определённым разделам сайта. Для запрета к просмотру используется директива Disallow. В основном эта директива применяется для запрета посещения пауком служебных разделов и конфиденциальной информации. А для разрешения директива Allow, параметрами, которой является URL раздела или страницы сайта.

## **Парсинг**

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1 Веб-скрапинг с помощью Node.js – URL: <https://medium.com/nuances-of-programming/%D0%BA%D0%B0%D0%BA-%D0%B2%D1%8B%D0%BF%D0%BE%D0%BB%D0%BD%D0%B8%D1%82%D1%8C-%D0%B2%D0%B5%D0%B1-%D1%81%D0%BA%D1%80%D0%B0%D0%BF%D0%B8%D0%BD%D0%B3-%D1%81-%D0%BF%D0%BE%D0%BC%D0%BE%D1%89%D1%8C%D1%8E-node-js-7285b5d8d53c> [4 мая 2020].

1. Инструменты анализа Web – URL: <https://www.osp.ru/os/2014/06/13042313/> [4 мая 2020].
2. Веб-скрапинг — что это и как он работает – URL: <https://jetruby.com/ru/blog/web-scraping-chto-eto/> [5 мая 2020].
3. Что такое поисковый робот, и как он работает – URL: <https://netpeaksoftware.com/ru/blog/chto-takoe-poiskoviy-robot-i-kak-on-rabotaet> [6 мая 2020].
4. Что за роботс ти экс ти? – URL: <https://netpeak.net/ru/blog/chto-takoe-robots-txt-i-zachem-on-voobshche-nuzhen/> [6 мая 2020].
5. О веб-краулерах – URL: <http://crawlers.info/pages/crawlers.html> [6 мая 2020].