DOI 10.37882/2223–2966.2021.12.04

ИССЛЕДОВАНИЕ ЗАЩИТЫ ОТ ВЕБ-БОТОВ: ВЛИЯНИЕ ЗАЩИТЫ НА ТОРГОВЫЕ ВЕБ-САЙТЫ И ЕЁ ВЗАИМОСВЯЗЬ С ПОСЕЩАЕМОСТЬЮ И РЕЛЕВАНТНОСТЬЮ В ПОИСКОВОЙ ВЫДАЧЕ

WEB BOT PROTECTION RESEARCH: THE IMPACT OF PROTECTION ON SHOPPING WEBSITES AND ITS RELATIONSHIP TO VISITS AND SEARCH RELEVANCE

P. Babaritsky
I. Gosudarev

Summary. This article describes the study of the interaction of automated systems of information collectors and the methods and means of protection against them opposed to them. Covers working with both automatic collectors, based on the Python programming language, and also discusses examples of various protections that web bots may encounter. Includes data collection results and statistics on a sample of the online shopping segment, such as traffic, relevance and security. Based on the aforementioned data, a forecast was formed about the likelihood of protection, depending on the popularity of the site or its position in the search engine.

Keywords: web-scrapping, web-crawler, security, web-bots, web-parsing.

Бабарицкий Павел Александрович
Аспирант, ФГАОУ ВО «Национальный исследовательский университет ИТМО»
pavel3345@yandex.ru
Государев Илья Борисович
К.п.н., доцент, ФГАОУ ВО «Национальный исследовательский университет ИТМО»
ilia-gossoudarev@yandex.ru

Аннотация. Данная статья описывает исследование взаимодействия автоматизированных систем сборщиков информации и противопоставленных им методов и средств защит от них. Охватывает работу как с автоматическими сборщиками, на базе языка программирования Python, а также рассматривает примеры различных средств защит, с которыми могут столкнутся веб-боты. Включает в себя результаты сбора информации и статистические данные о выборки сегмента интернет-магазинов, такие как посещаемость, релевантность и наличие защиты. На основе вышеупомянутых данных, был сформирован прогноз о вероятности наличия защиты в зависимости от популярности сайта или положения его в поисковой системе.

Ключевые слова: web-scrapping, web-crawler, security, web-bots, web-parsing.

Введение

Анализ методов извлечения и обработки данных Интернет-ресурсов является актуальным направлением исследования в области прикладной веб-безопасности и коммерческого использования веб-ресурсов. Об этом свидетельствует множество современных научных публикаций, посвященных проработке положений указанной проблемной области. Так, авторами статьи с названием «Exploiting web scraping in a collaborative filteringbased approach to web advertising» [1] раскрывается специфика и назначение такого явления как web-scrapping.

Актуальность данного исследования в рамках поставленной задачи обусловлена отсутствием аналогичных исследований с целью выявления потенциального влияния методов защиты на веб-сайт с использованием теоремы Байса. А также для большей значимости проводимых экспериментов данное исследование проводит сбор метрик с более чем с одной поисковой системы. Исследование проводилось для разработки парсера данных способного собирать информацию о веб-сайтах, для его дальнейшего анализа, дальнейшей модификации данного функционала и нахождения взаимосвязей между существующими параметрами. Данные знания можно использовать для фильтрации из общего количества сайтов для более подробного исследования существующих способов защиты таких как robots.txt [2], CAPTHCA [3], и различные программные методы [4, 5]. Также рамках данной работы были получены результаты показавшие потенциальную взаимосвязь между наличием защиты и числом посещений, полученные результаты могут быть дополнены путём улучшения алгоритмов поиска наличия защитных механизмов путем разделения по способам защиты, подобное знание позволит понять какие типы защит обладают большим воздействие на посещение веб сайтов. Полученные знания может повлиять на сбор статистических сведений о различных способах защиты данных в интернете и дополнить SEO анализ [5]. А информация о воздействии защиты на веб-сайты позволит понять какие способы противодействия веб-бо-

```
def get_shops_links(url, count):
    driver.get(url)

    links = []

    while count > 0:
        all_shop_links = driver.find_elements_by_xpath(
            '//div[@class]/div[@data-hveid and @data-ved]/div[@class]/div[@class]/a')
        for link in all_shop_links:
            if link not in links:
                links.append(link.get_attribute('href'))
            # print(link.get_attribute('href'))

        if count > 1:
            driver.find_element_by_xpath(
                "//tbody/tr//td[@role='heading']/a/span[contains(text(),'Следующая')]").click()

        count = count - 1

    return links
```

Рис. 1. Сбор ссылок торговых магазинов

там наименее приводят к оттоку посетителей. Данная информации позволит разработчикам веб сайтов выбирать лучшие практики относительно выбора способов защиты.

Всё вышесказанное определяет потенциальные развитие данного исследования, которое основываясь на полученных данных и пользуясь инструментами, разработанными в процессе написания данной статьи позволит провести более фундаментальное исследование и рассмотреть каждый элемент защиты исследуемых сетей на предмет возможных их уязвимостей. Установление наличия взаимосвязи между посещаемостью и защитой позволит говорить о том, что разработчикам защитного программного обеспечения рекомендуется разрабатывать механизмы защиты, которые не воздействуют на число посещений и недостаточно агрессивны чтобы приводить к воздействию на поисковых веб-ботов, вследствие которого уменьшается положение в поисковой выдаче.

Объектом исследования выступили торговые онлайн магазины, а предметом исследования средства защиты информации от web сборщиков. В качестве выборки выступили торговые площадки, которые попадают в релевантные выдачи поисковой системы Google. Теорема Байеса, был выбран в качестве основного анализирующего уравнения, так как он позволяет дополнять данные новыми вводными и тем самым позволяет дополнять уже произведенный анализ, и учитывать получение ранее результаты, в более поздних и подробных анализах данных.

Методы исследования

Все современные технологии защиты от сбора данных, направлены на усложнение процедуры сбора и каталогизации, путем усложнения и повышения уровня контроля доступа к web ресурсу. Таким образом, главной задачей web мастера конкретного ресурса, становится поиск точки баланса, между комфортным доступом к ресурсу и сложностью противостояния web-ботам. Для проверки данного предположения, сформируем ряд гипотез, о наличии связи между посещаемостью ресурсом и наличие на нем средств защиты. Проверка будет производится научным методом Байеса, который позволяет произвести расчёты предположений о всемирной сети интернет.

Для сбора данных было разработано приложение на языке программирования Python 3. Собиралась информация о сайтах, которые занимаются торговлей. В качестве основной выборки выступили сайты, которые отдавала поисковая система Google, в результате полученных ответов на «get-запросы» был сформирован массив тематических интересующих сайтов. Для автоматического сбора и эмуляции работы пользователя использовалась библиотека Selenium для языка программирования Python 3, которая позволяет работать с элементами интерфейса сайтов.

Для навигации и поиска данных используются xpath-запросы, позволяющие получить искомые ссылки и совершать переходы по ним (Рис. 1). Для получения информации о собранных сайтах, была создано две функ-

```
def enter_on_site():
    time.sleep(1)
    driver.get("https://a.pr-cy.ru/")
    driver.find_element_by_xpath("//span[@class='badge']").click()
    driver.find_element_by_xpath(
        "//input[1]").click()
    driver.find_element_by_xpath(
        "//input[1]").send_keys("[redacted]")
    driver.find_element_by_xpath("//input[@id='password']").click()
    driver.find_element_by_xpath(
        "//input[@id='password']").send_keys("[redacted]")
    driver.find_element_by_xpath(
        "//button[@type='submit' and @class='btn btn-primary btn-md']").click()
```

Рис. 2. Автоматизированное прохождение регистрации

```
def send_website_for_analysis(url):
    mass = []

    try:
        driver.get("https://a.pr-cy.ru/" + url)
    except:
        try:
            driver.get("https://a.pr-cy.ru/" + url)
        except:
            return mass

    time.sleep(10)
    try:
        driver.find_element_by_xpath("//button[@class='close']").click()
    except:
        time.sleep(0)

    time.sleep(1)

    parse_visit_count = driver.find_elements_by_xpath(
        "//div[@class='analysis-test__content']/table[@class='table-clear table-content-test']/tbody/tr/td[@class='text-right' and position() <=
    for elem in parse_visit_count:
        value = elem.text.replace(u'\xa0', "")
        value = value.replace(" ", "")
        try:
            mass.append(int(value))
        except:
            print('error ', url)
    return mass
```

Рис. 3. Сбор данных для исследования

ции, одна из которых проходит регистрацию на сайте выполняющий SEO анализ (Рис. 2), а вторая принимает ссылки сайтов и отправляет сформированные запросы стороннему сайту для анализа (Рис. 3).

Полученный ответ аналогичным образом подвергается парсингу с помощью xpath-запросов и полученные результаты возвращается в виде массива данных. Эта функция обрабатывает каждый сайт поочерёдно и суммарные результаты проанализированных сайтов будут помещены в словарь, состав которого затем будет использован в качестве базовых данных для построения функциональных графиков зависимости на основе теории Байеса. Для учета сайтов, обладающих средствами защиты от воздействия автоматизированного кода, были так же учтены в итоговых цифрах словаря при помощи функции checkShop (Рис. 4).

Данная функция вызывается и проверяет сайт на наличие отказа в доступе при перемещениях веб-бота на сайте. Учёт отказа происходит в том случае если сайт возвращает отказ в каком-либо из существующих ранее способов, например в виде возможных текстовых сообщений “access denied” или другими способами перебира-

```
def checkShop(url, access_denied_counter):
    try:
        driver.get(url)
        time.sleep(1)
        try:
            driver.find_element_by_xpath(
                "//*[contains(text(),'Access denied') or cont
        except NoSuchElementException:
            info_about_views_visits[url].append(0)
            return access_denied_counter

        info_about_views_visits[url].append(1)
        access_denied_counter = access_denied_counter + 1
        return access_denied_counter
    except:
        info_about_views_visits[url].append(2)
        access_denied_counter = access_denied_counter + 1
        return access_denied_counter
```

Рис. 4. Подсчет отказов в доступе

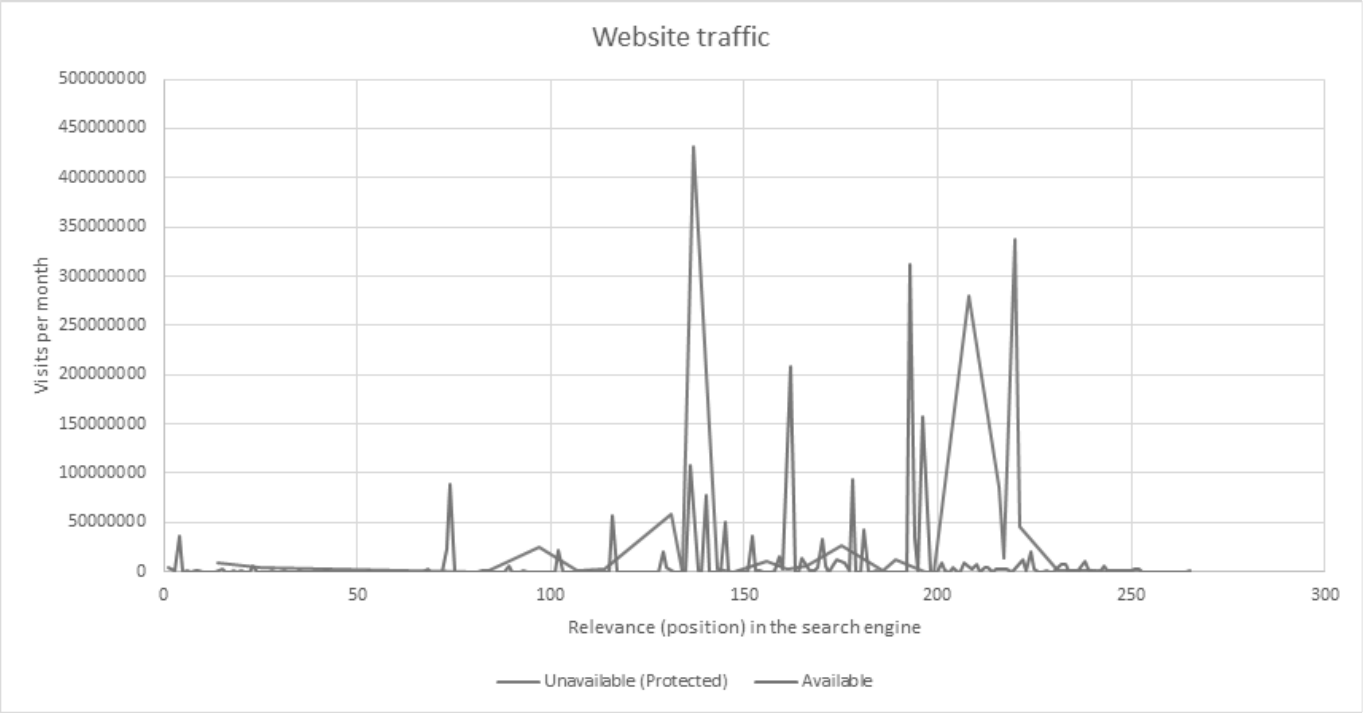

Рис. 5. Зависимость между посещениями и релевантностью за месяц

емых в xpath запросе путем объединения запросов логическими операциями “and” и “or”. Например, в случае необходимости пройти CAPTCHA элемент интерфейса будет найден на сайте за счет классов, содержащих элемент капчи, например класс “rc-anchor-logo-img-portrait”. Если на сайте применялся файл htaccess для ограничения доступа xpath обнаружит текст Forbidden в соответствующем теге с текстом ошибки.

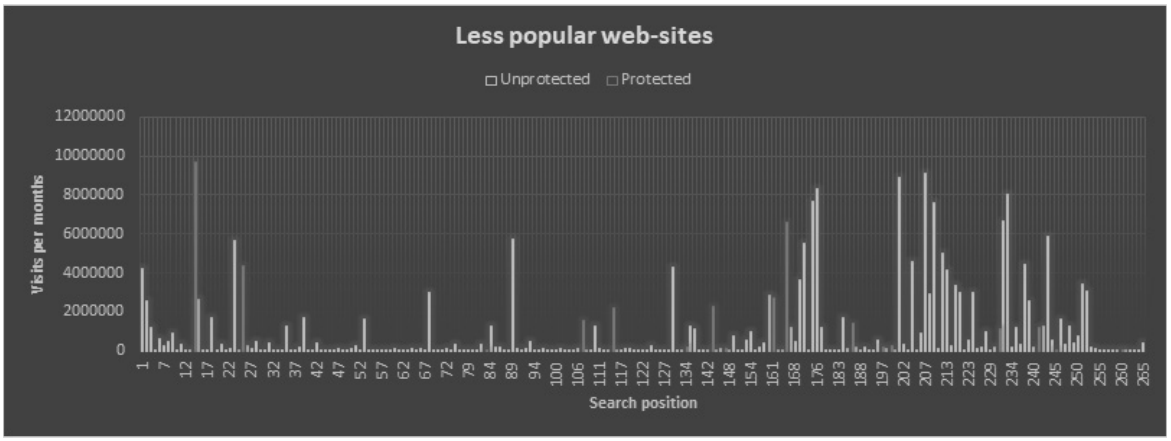

Рис. 6. Зависимость между посещениями и релевантностью (less popular sites)

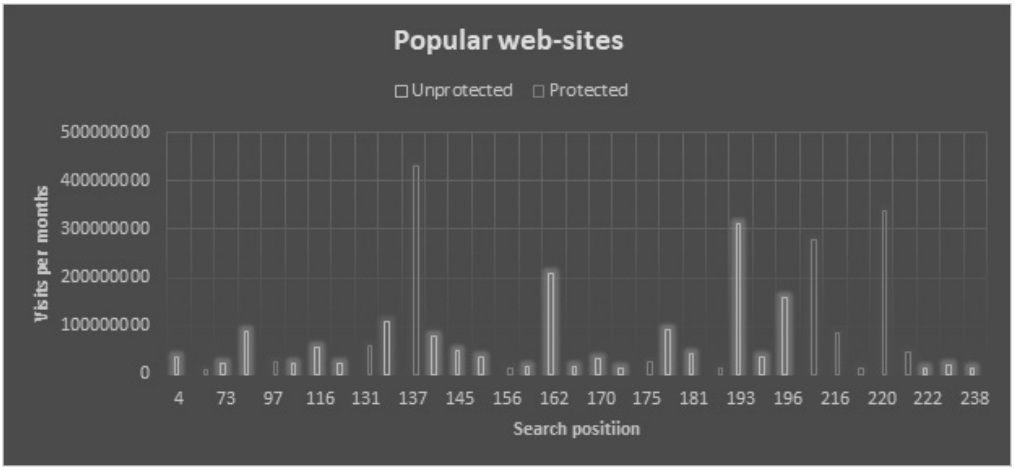

Рис. 7. Зависимость между посещениями и релевантностью (popular sites)

Все полученные результаты сохранены с помощью модуля Pandas в файл-документ в формате таблицы Excel, в которых уже при помощи внутри-операционных инструментов были построены графики: зависимости количества посетителей в месяц (ось Y) и релевантности в соответствии с позицией в поиске (ось X); зависимости количества просмотров от позиции в поисковике с учетом наличия защиты (Рис. 5).

Для анализа полученных результатов была использована одна из наиболее применяемых теорем элементарной теории вероятностей, позволяющая делать предположения с определённой точностью о вероятности того или иного события принимая в учет те условия, что произошли другие статистически взаимосвязанные с ним события.

Формула Байеса, применяемая для получения результатов, выглядит следующим образом:

$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$, где

$P(H)$ — априорная вероятность гипотезы H;

$P(H \mid E)$ — вероятность гипотезы H при наступлении события E (апостериорная вероятность);

$P(E \mid H)$ — вероятность наступления события E при истинности гипотезы H;

Результаты исследования

В результате произведенного исследования, были получены многочисленные данные о популярности сайтов и наличие на них защиты от посещения их web-ботами, и для удобного восприятия, они была занесены в график, где по оси X позиция в поисковой системе, а по оси Y количество посещений в месяц. Цветовая маркировка обозначает наличие защиты, где красная — «protected», а зелёное «unprotected».

Таблица 1. Рассчитанные вероятностные оценки для поисковой системы Google

	P(H)	P(H)	P(E|H)	P(E|H)	P(E)	P(H|E)
1	0,10566038	0,89434	0,42857143	0,09704641	0,13207547	0,34285714
2	0,10566038	0,89434	0,60714286	0,92241379	0,89033582	0,07343815
3	0,10566038	0,89434	0,142857143	0,405063291	0,377358491	0,04
4	0,10566038	0,89434	0,857142857	0,857142857	0,838696398	0,107984293

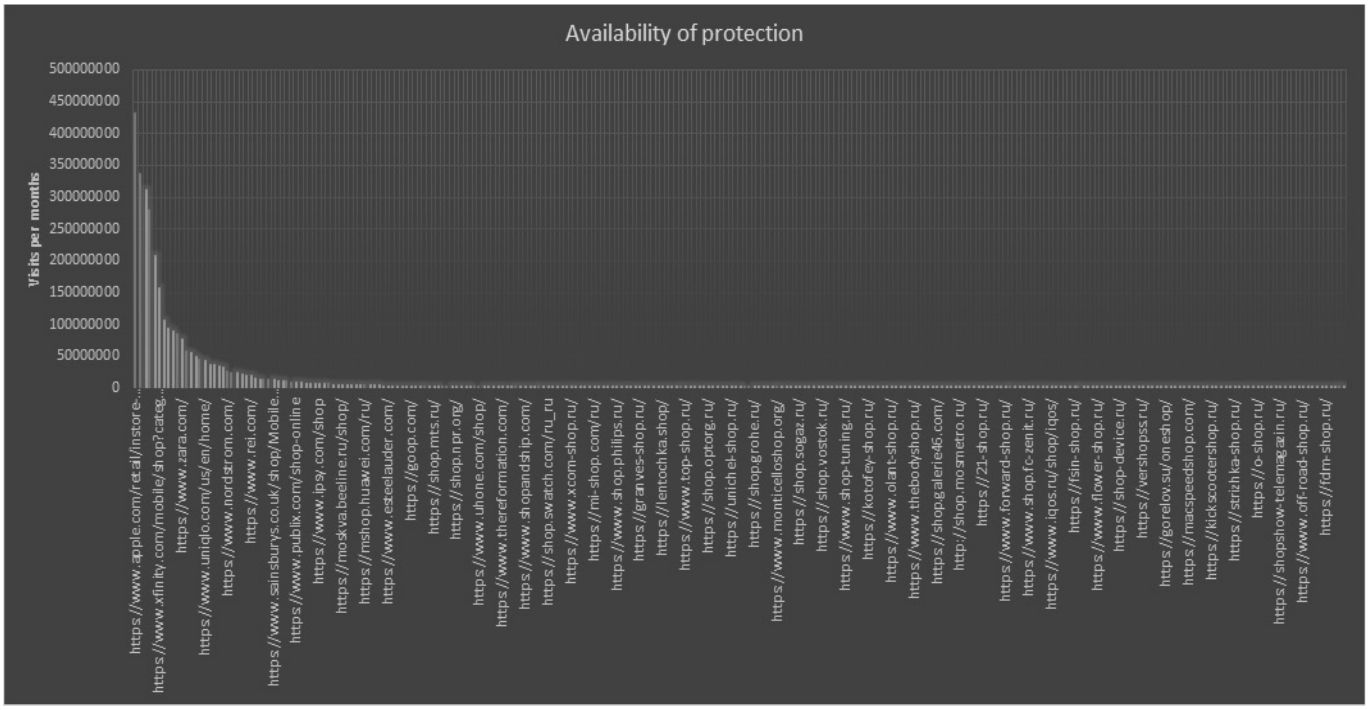

Рис. 8. Зависимость между посещениями и числом посещений в порядке убывания

На первом графике (Рис. 6) сделана выборка из менее популярных сайтов, а на втором (Рис. 7) из наиболее популярных. Третий график (Рис. 8) демонстрирует выборку лидеров по количеству посещений, от большего к меньшему, без учета номера релевантной выдачи. На основе этих выборок, были сформированы несколько гипотез, а именно:

1. $P(H|E)$ — гипотеза о том, что популярные торговые веб-сайты используют защиту от веб ботов;
2. $P(H|E)$ — гипотеза о том, что не популярные торговые веб-сайты используют защиту от веб ботов;
3. $P(H|E)$ — гипотеза о том, что релевантные торговые веб-сайты используют защиту от веб ботов;
4. $P(H|E)$ — гипотеза о том, что менее релевантные торговые веб-сайты используют защиту от посещения их страниц веб ботами.

Где H — случай, когда сайт обладает защитой, а E — соответствующее накладываемое ограничивающие условие (популярные, не популярные, релевантные, не релевантные сайты). Следовательно, $P(E)$ — будет вероятностью наступления такого условия, для каждого из четырех случаев оно отличается. За релевантных были выбраны первые 100 веб-страниц, а за популярные, те сайты, которые смогли превысить планку в 9500000 человек в месяц.

Расчеты, произведенные для гипотез, были занесены в таблицу 1.

$P(H)$ — априорная вероятность или доля случаев, когда эксперимент привел к результату H или первоначальный уровень доверия предположению H, когда сайт обладает защитой, следовательно для всех четырех случаев априорная вероятность равна: $P(H)$ = (число защищенных сайтов)/(общее число сайтов) = 0,10566038.

Согласно результатам по 1 и 2-й гипотезе $P(H|E)$ непопулярных значительно меньше, чем $P(H|E)$ для популярных сайтов.

Таблица 2. Рассчитанные вероятностные оценки для поисковой системы Yandex

	P(H)	P(H)	P(E|H)	P(E|H)	P(E)	P(H|E)
1	0,108247423	0,891753	0,52380952	0,13872832	0,18041237	0,31428571
2	0,10824742	0,891753	0,47619048	0,86127168	0,81958763	0,06289308
3	0,10824742	0,891753	0,476190476	0,554913295	0,360824742	0,142857143
4	0,10824742	0,891753	0,523809524	0,919354839	0,876538078	0,06468747

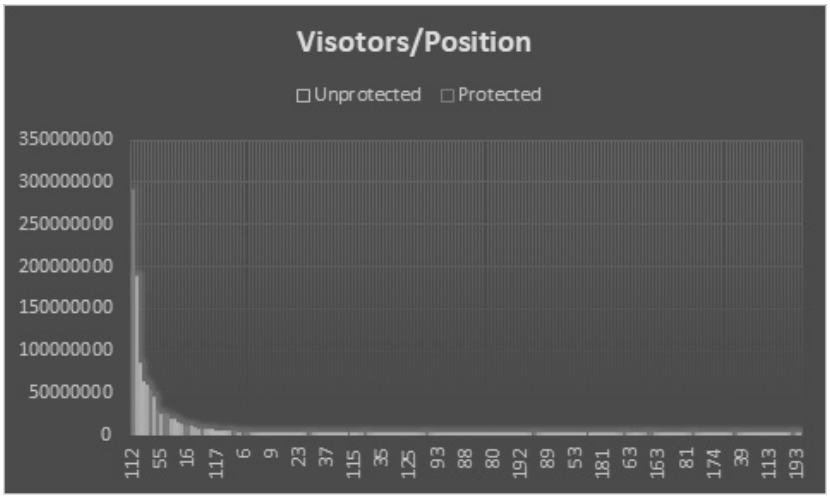

Рис. 9. Зависимость между посещениями и числом посещений в порядке убывания

Как показали результаты 3 и 4 гипотез: говорить о том, что менее или более релевантные сайты предпочитают использовать защиту невозможно в связи с тем, что постериорная вероятность мала в обоих случаях. Тем не менее вероятность встретить, менее релевантный сайт с защитой выше, что вероятно вызвано тем, что жесткая защита влияет на поисковых ботах.

Среди всех четырёх гипотез судить о возможной взаимосвязи между популярностью и наличием защиты позволяет только результат, полученный о вероятности того, что гипотеза верна в первом случае, где данная вероятность равна 0,34.

Для дополнительной проверки результатов и значимости экспериментов был произведен сбор метрик для альтернативной поисковой системы Yandex.

Для получения результатов использовались аналогичные программные способы рассмотренные для Google. За исключением SEO сайта, предоставляющего данные о месячном посещении: была произведена замена a.pr-cy.ru на be1.ru.

Результаты для поисковой системы Yandex и аналогичных гипотез представлены на графике (Рис. 9) и занесены в таблицу 2. Полученные результаты сопоставимы с результатами первого эксперимента.

Заключение

Проведенное исследование, показало о справедливости одной из выведенных гипотез, относительно связи о том, что популярные торговые веб-сайты используют защиту от веб ботов. Однако же остальные выведенные гипотезы, продемонстрировали малый коэффициент соответствия действительности, и не прошли проверку, а следовательно, не подтвердились. Так же, полученные цифры, позволяют предположить о работе различных алгоритмов и сценариев выдачи поисков систем, что могло повлиять на итоговые результаты исследования. Таким образом, для получения более обширного ответа и выведения более точных гипотез, следует использовать большее количество входных данных, таких как: системы поиска, различные ключевые слова для запросов, другие тематические ресурсы. А также, немаловажен тот факт, что для интернет-ресурсов, с большим трафиком жизненно необходимо наличие отсеивающих защит от интернет сборщиков информации, даже если это влияет на положение в выдаче поискового ресурса, так как это напрямую влияет на стоимость обслуживания данного ресурса и это

в свою очередь, может влиять на полученные в результате исследования цифры.

В соответствии с полученными результатами было принято решение о повторном исследовании и дополнении полученных результатов. Также разработанный веб-бот, предполагает усовершенствование с целью получения иных дополнительных сведений о торговых площадках для будущих исследований, которые позволят развивающих область SEO анализа.

ЛИТЕРАТУРА

1. Vargiu, E. Exploiting web scraping in a collaborative filteringbased approach to web advertising / E. Vargiu, M. Urru // Artificial Intelligence Research.— 2013.— Vol.2, № 1.— P. 44–54.
2. Drott, M.C. Indexing aids at corporate websites: the use of robots.txt and META tags / M.C. Drott // Artificial Intelligence Research.— 2002.— Vol.38, № 2.— P. 209–219.
3. CAPTCHA: Using Hard AI Problems for Security / L. von Ahn, M. Blum, N.J. Hopper, J. Langford // Advances in Cryptology — EUROCRYPT 2003.— Berlin: Springer, 2003.— P. 294–311.
4. Loesch, W., Fluker, D. Secure web site authentication using web site characteristics, secure user credentials and private browser // United States Patent № US8095967B2. 2012.
5. Spider Trap // techopedia, 2019 [Электронный ресурс].— Режим доступа: https://www.techopedia.com/definition/5197/spider-trap/, свободный. (дата обращения: 18.12.2019).
6. SEO: A unique approach to enhance the site rank by implementing Efficient Keywords Scheme / K. Rehman, A. Yasin, T. Mahmood, M. Azeem et al. // PeerJ Preprints, 2019.— P. 1–21.

Журнал «Современная наука: актуальные проблемы теории и практики»

НИУ ИТМО