DOI: 10.15514/ISPRAS–2020–33(2)–3

Регулярные выражения для обнаружения Web-рекламы на основе автоматического скользящего алгоритма

Д. Рианьо, ORCID: 0000-0001-8998-0129 <donovan20@comunidad.unam.mx>
Р. Пиньон, ORCID: 0000-0002-4819-5703 <rodrigo_pinon@comunidad.unam.mx>
Г. Молеро-Кастильо, ORCID: 0000-0002-6330-6408 <gmoleroca@fi-b.unam.mx>
Э. Барсенас, ORCID: 0000-0002-1523-1579 <ebarcenas@unam.mx>
А. Веласкес-Мена, ORCID: 0000-0003-3509-6236 <mena@fi-b.unam.mx>

Национальный автономный университет Мексики,
Мексика, 04510, Мехико, Мэрия Койоакана

Аннотация. Представлена реализация алгоритма распознавания Web-рекламы с использованием регулярных выражений. Сегодня при разработке программного обеспечения важную роль играет использование регулярных выражений, оптического распознавания символов, баз данных и автоматизированного тестирования. Тесты проводились в трех веб-браузерах. Результатом явилось выявление рекламы на испанском языке, которая отвлекает внимание пользователей, а прежде всего, позволяет получать информацию о них. Основная особенность алгоритма заключается в том, что его автоматическое и настраиваемое выполнение не требует доступа к коду рассматриваемой страницы, и будущем может появиться приложение, работающее в фоновом режиме. Кроме того, при поддержке оптического распознавания символов мы получаем приемлемую эффективность при выявлении рекламы.

Ключевые слова: цифровой маркетинг; оптическое распознавание символов; регулярные выражения; интернет-реклама

Для цитирования: Рианьо Д., Пиньон Р., Молеро-Кастильо Г., Барсенас, Э., Веласкес-Мена А. Регулярные выражения для обнаружения Web-рекламы на основе автоматического скользящего алгоритма. Труды ИСП РАН, том 33, вып. 2, 2021 г., стр. 65-76. DOI: 10.15514/ISPRAS–2021–33(2)–3

Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm

D. Riaño, ORCID: 0000-0001-8998-0129 <donovan20@comunidad.unam.mx>
R. Piñon, ORCID: 0000-0002-4819-5703 <rodrigo_pinon@comunidad.unam.mx>
G. Molero-Castillo, ORCID: 0000-0002-6330-6408 <gmoleroca@fi-b.unam.mx>
E. Bárcenas, ORCID: 0000-0002-1523-1579 <ebarcenas@unam.mx >
A. Velázquez-Mena, ORCID: 0000-0003-3509-6236 <mena@fi-b.unam.mx>
National Autonomous University of Mexico,
Coyoacan Mayor's Office, Mexico City, CDMX, C.P. 04510, Mexico

Abstract. This paper presents the automation of a Web advertising recognition algorithm, using regular expressions. Currently, the use of regular expressions, optical character recognition, Databases, and automation tests have been critical for multiple Software implementations. The tests were carried out in three Web browsers. As a result, the detection of advertisements in Spanish, that distract attention and that above all extract

information from users was achieved. The main feature of the algorithm is that automatic and versatile execution does not require access to the code of the page in question and that in the future it can be an application with background operation. In addition, being supported by optical character recognition gives us acceptable efficiency in detecting advertising.

Keywords: Digital Marketing; Optical Character Recognition; Regular Expressions; Web Advertising

For citation: Riaño D., Piñon R., Molero-Castillo G., Bárcenas E., Velázquez-Mena A. Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 2, 2021, pp. 65-76 (in Russian). DOI: 10.15514/ISPRAS–2021–33(2)–3.

1. Введение

В прошлом маркетинг не существовал в режиме онлайн, и его главная цель заключалась в привлечении средств массовой информации, чтобы сформировать у людей или даже компаний положительное мнение о рекламируемой продукции или о планируемых к продажам идеях. Но теперь этому наступил конец; имея под рукой все инструменты, которые дала нам современная технология, люди используют поисковые системы, чтобы найти все необходимое и более того, они могут получить доступ также и к критике и комментариям сообщества.

С появлением цифрового маркетинга основной целью становится пользователь [1], поэтому изменилась парадигма методов маркетинга. Сегодня цифровая стратегия должна включать все подходящие пространства для взаимодействия с «целью», выискивая людей, могущих повлиять на мнение других пользователей, чтобы привлечь их в свою сеть и усилить позиции идей или продуктов. Эта стратегия также может быть направлена на совершенствование поисковых систем, которые, как показывает опыт, становятся все более назойливыми в своих попытках проникновения в умы пользователей.

Для лучшего понимания текущих тенденций цифрового маркетинга на основе, динамических Web-страниц, была разработана семантическая сеть. Важно отметить, что семантическая сеть – это форма представления знаний через взаимосвязи в виде графа [2]. На рис. 1 показаны взаимосвязи семантической сети, в которой рекламные объявления на веб-страницах можно разделить на три группы: a) сайты, которые существуют за счет рекламы; b) информационные, корпоративные сайты и интернет-магазины; c) социальные сети или нечто подобное.

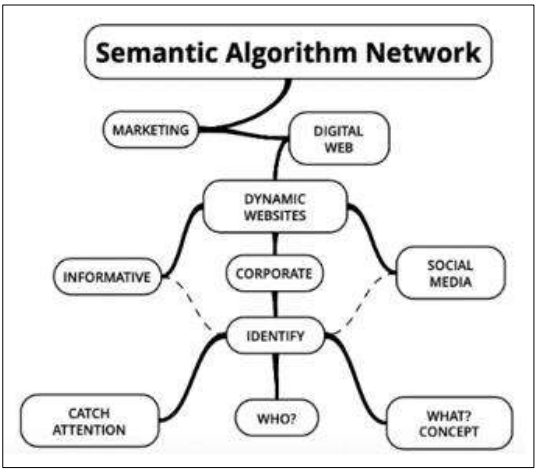

Рис. 1. Семантическая сеть современной Web-рекламы
Fig. 1. Semantic network of current Web advertising

Эта семантическая сеть связывает следующее: i) кто предлагает, ii) какой продукт или услуга предлагается, iii) каким образом привлекается внимание (предложение или раскрутка). Эти

типы рекламы распознаются при просмотре Интернета, некоторые из них были предметом анализа в настоящей исследовательской работе.

Веб-сайты, существующие за счет рекламы, предлагают краткие новостные, образовательные или развлекательные материалы, и их рекламный контент предоставляется Google. В них реклама постоянно меняется. Информационные, корпоративные сайты и интернет-магазины представлены такими сайтами, как MSN, Amazon, Sanborns, Adidas и др., целью которых является предложение текущих трендов продуктов или услуг. Социальные сети и их вариации, такие как Facebook, Instagram, Linkedin, Twitter и пр. являются сайтами, которые предлагают взаимодействие с пользователями и рекламу, привязанную к профилям пользоватей.

В настоящей статье представлена реализация алгоритма распознавания Web-рекламы с использованием регулярных выражений. Тесты проводились в трёх браузерах: Chrome, Firefox и Safari. Особенностью алгоритма является его автоматическое и настраиваемое выполнение, поскольку он не требует доступа к коду просматриваемой Web-странице и может считаться приложением, работающим в фоновом режиме.

2. Предварительная информация

При наличии высокого потенциала и развитых стратегий современного цифрового маркетинга эта деятельность используется все чаще как важная часть деятельности кампаний по обеспечению приверженности потребителей к торговым маркам, поскольку обеспечивается коммуникационный канал для взаимодействия между потенциальных клиентов и торговой маркой. К числу действий, ориентированных на интернет-маркетинг, относятся [3]: обеспечение потребительской лояльности; повышение имиджа торговой марки и продаж продуктов; проведение акций по раскрутке и тестированию продуктов; поощрение повторной покупки продукта той же торговой марки; проведение прямых и персонализированных коммуникационных кампаний.

Одним из способов борьбы с рекламой и обеспечения конфиденциальности в Web-браузерах является использование регулярных выражений (regular expression, RegExp, Regex, или RE). Эти выражения представляют собой способ представления символьных строк, которые соответствуют некоторому шаблону [4] [5]. Приложения RegExp разнообразны, например, валидация полей форм, идентификация текстовых строк в социальных сетях, поисковые команды и т.д. [6] [7].

2.1 Управление конфиденциальностью пользователей

В настоящее время управление информацией и конфиденциальностью находится в критической ситуации [8]. Такие компании, как Facebook, Twitter, Google, Amazon и др. включают в свои веб-сайты, приложения или поисковые системы средства накопления информации. Эти компании предоставляют услуги «бесплатно», но используют информацию о пользователях по своему усмотрению, в рекламных целях.

Увеличение количества рекламы в Web-браузерах ставит под угрозу конфиденциальность пользователей. Web-сайты, предлагаемые по результату поиска определенных видов товаров и услуг, перенасыщены информацией о клиентах. Другой важный аспект заключается в том, что сами поисковые запросы пользователей позволяют собрать большой объем данных. Как следствие, эти данные при определенной обработке могут также предоставить яполезную информацию для более эффективного управления целевой рекламой [9] [10].

Для примера на рис. 2 показано, как поисковая система Google автоматически создает профили пользователей на основе просмотра ими веб-страниц. Например, релевантным является пользователь мужского рода, возраст которого колеблется от 18 до 24 лет, заинтересованный в покупке онлайн, проведении открытых мероприятий, разработке мобильных приложений, просмотре фильмов и прочее. Например, может быть актуален

пользователь мужского пола в возрасте от 18 до 24 лет, который заинтересован в совершении покупок в Интернете, занятиях активным отдыхом, разработке мобильных приложений, просмотре фильмов и т.д. Данные об таких пользователях собираются через файлы cookie, трекеры и информацию из профиля Gmail, и эти данные могут быть использованы для группировки профилей с общими характеристиками с целью рекламы различных предложений и промоакций.

Рис. 2. Поисковая система Google для автоматического создания профилей пользователей на основе просмотра Web-страниц

Fig. 2. Google search engine for automatic creation of user profiles based on Web browsing

Поэтому при постоянном взаимодействии со службами Google стоит почаще знакомиться с изменениями в политике конфиденциальности в Маунтин-Вью[1], чтобы иметь лучшее представление о безопасности и использовании нашей информации.

2.2 Блокировщики рекламы веб-браузеров

В последние годы были увеличены усилия по внедрению блокировщиков веб-рекламы [11]. Первоначально они были реализованы для браузера Firefox, который годами совершенствовался, чтобы получился браузер без рекламы. Но компании искали способы продолжить рассылку рекламы. В скором времени возникла полемика [1] в связи со значительными потерями возможной прибыли [12].

Еще один термин, получивший распространение в настоящее время, – это трекеры в маркетинге, которые являются средствами отслеживания и индикаторами эффективности таргетированных рекламных кампаний. Эти инструменты сохраняют информацию с помощью файлов cookie, а те, в свою очередь, предоставляют местоположение выполняемых поисковых запросов. По мере роста спроса и расширения возможностей блокировки рекламы в большинстве Web-браузеров были реализованы новые блокировщики рекламы.

Необходимо учитывать тот важный факт, что в последние годы Google контролирует 85% мирового рекламного бизнеса в поисковых системах и около 50% всей онлайн-рекламы [13]. Люди и общество, как правило, считают Google сервисом [14], но за этим стоит технология, которая поддерживает специальные функции и включает уникальные и ограничительные расширения.

[1] В Маунти-Вью (Mountain View), Калифорния находится штаб-квартира компании Google.

2.3 Управление Web-браузерами на основе Selenium

Selenium, называемый также Selenium Webdriver, – это инструмент для автоматизации процессов в различных Web-браузерах [15]. Его цель – улучшить поддержку обнаружения проблем в любом Web-браузере [16]. Этот инструмент позволяет тестировать любой веб-браузер с получением HTML-кода, изменять и открывать вкладки окон браузера, а также и перемещаться между вкладками, изменять размеры окон, создавать скриншоты, заполнять поля форм многое другое.

Selenium поддерживает языки программирования Java, Python, C#, Ruby, Perl и JavaScript. Операционные системы –Windows, Mac OS и Linux, каждая со своими соответствующими пакетами и интегрированными средами разработки (Integrated Development Environment, IDE) [17].

С другой стороны, в настоящее время существует интерес к использованию оптического распознавания символов (OCR) при обнаружении Web-рекламы [14], однако все еще требуются дополнительные усилия для охвата всех стратегий цифрового маркетинга.

2.4 Родственные работы

Одной из работ, посвященных обнаружению Web-сайтов, пригодных для таргетированной рекламы, была статья [18], где представлен алгоритм, выполняющий Web-краулинг. Метод состоит в получении информации с Web-сайтов путем обучения системы на заранее размеченном наборе страниц (для этого использовались страницы MSN). Классификация производилась по предполагаемым ключевым словам, встречающимся в начале фразы, в середине фразы, в конце фразы и вне фразы.

Работа [19] описывает анализ контекстной рекламы с помощью PageSense, который направлен на ассоциирование рекламы со стилем Web-страниц. С помощью этой платформы выявляются пустые области Web-страниц и выбирается не назойливое место для размещения рекламы, не нарушая оригинальный стиль Web-страницы. Для анализа использовались байесовские комбинации и вероятности, которые отражают процентное соотношение рекламных объявлений для различных видов товаров или услуг.

В работе [20] анализ производился на основе евклидовых расстояний. Эти расстояния позволяют оценить, какой интерес представляет реклама для пользователей, насколько она помогает при поиске продуктов, а также способствует адаптации профилей пользователей, позволяя разбить его на разделы, такие как здоровье, спорт, бизнес-общество, образование, искусство, наука, компьютер и т.д.

Наконец, в работе [21] описан анализ, проведенный примерно на 500 веб-страницах и направленный на выявление типов (не содержания) рекламы. Среди проанализированных типов рекламы выделяются всплывающие окна, карусели, видео, GIF-файлы, игры, стикеры или текст. Также были проанализированы страны, откуда поступают рекламные объявления, их частота, размер и происхождение URL.

3. Методика

Алгоритм был реализован на языке Python с применением библиотек Tesseract [22], Pillow и OpenCV [23], а также пакета Tesseract-OCR. Реализация начинается с точки, предшествующей процессу OCR. Кроме того, мы использовали библиотеку Selenium для выполнения автоматического скольжения в Web-браузере.

При реализации также учитывалась разница в высоте мониторов различных размеров, представленных на рынке, поэтому алгоритм динамически определяет высоту окон, подстраиваясь под любой размер экрана. По этой причине перемещение информации в настоящей работе является вертикальным.

3.1 Скольжение по Web-сайту

Через Selenium открывается новое окно для браузера, совместимого с этим инструментом. Терминал запрашивает у пользователя поле URL. Далее тестируемый браузер расширяется на весь экран для быстрого и полного сканирования Web-сайта.

С целью приостановки процесса на несколько мгновений используются программные потоки управления (thread). Первая пауза делается для того, чтобы дать странице загрузиться, так как в зависимости от скорости Интернета и даже от состояния страницы требуется время для полной ее загрузки, чтобы можно было сделать правильный снимок экрана для его последующего анализа с помощью OCR.

На следующем этапе окно разворачивается до максимума, а затем скользит для захвата и сохранения экранов. Это вызывает вторую паузу, которая составляет 0,5 секунды. Последовательно запускается процесс захвата экрана, пока не будет обработано все вертикальное содержимое Web-страницы. После этого окно Web-браузера автоматически закрывается.

3.2 Оптическое распознавание символов

OCR, позволяет с высокой эффективностью извлекать буквенный текст из изображений с независимо от языка, размера или цвета текста. Эффективность OCR колеблется от 71% до 98%. Такая система способна достигать средних значений 85,1% для рукописного текста и 90,93% для печатного текста [24].

Была определена специальная функция для поиска всех скриншотов, сохраненных в файловой системе с заданной частью путевого имени. Затем был реализован цикл, в котором все найденные изображения анализировались в том порядке, в котором они были захвачены.

После обработки нужного изображения результаты OCR сохраняются в текстовой переменной. Затем текст форматируется: все буквы заменяются на прописные, слова разделяются, устраняются пробелы и символы, не поддерживаемые в символьных строках в языке SQL, чтобы иметь возможность сопоставления с регулярными выражениями.

3.3 Регулярные выражения

После того, как информация о содержимом Web-страницы получена и преобразована в текстовые строки, производится доступ к локальному серверу для проверки содержимого Web-сайта с помощью базы данных, в трех таблицах которой хранится около 600 различных слов на основе регулярных выражений: i) слова, наиболее часто используемые в цифровом маркетинге; ii) торговые марки, учитывая их соответствующие суб-бренды в продуктах или услугах, предлагаемых компаниями; iii) тип продукта или сервиса, или из чего состоит продукт или услуга.

В табл. 1, 2 и 3 показаны фрагменты регулярных выражений, соответственно относящиеся к наиболее часто используемым словам в цифровом маркетинге, некоторым узнаваемым брендам и некоторым продуктам, наиболее часто упоминаемым в Мексике.

Табл. 1. Фрагмент наиболее часто используемых слов в цифровом маркетинге на испанском языке
Table 1. Fragment of the most used words in digital marketing in Spanish

Слово	**Множественное число**	**Акцент**	**Символ**
Ahorro	Ahorros	Null	Null
Bajo	Bajos	Null	Null
Comprar	Null	Null	Null
Cotiza	Null	Null	Cotizar
Descuento	Descuentos	Null	%
Dinero	Dineros	Null	Null

Especial	Especiales	Null	Null
Gratuito	Gratuitos	Null	Gratis
Hasta	Null	Null	Null
Ilimitado	Ilimitados	Null	Null
Interes	Intereses	Interés	Null
Internet	Null	Null	Web
Oferta	Ofertas	Null	Null

Табл. 2. Фрагмент некоторых из самых продаваемых и самых дорогих брендов Мексики
Table 2. Fragment of some of the best-selling and highest-paid brands in Mexico

Бренд	**Суб-бренд**	**Продукт**	**Акроним**
Adidas	Boost	Boost	Null
Adidas	NMD	NMD	Null
Apple	iMac	iMac	Null
Apple	iPhone	iPhone	Null
Bancomer	BBVA	BBVA	BBVA
Banorte	Banorte	Banorte	Null
Levis	Trucker	Trucker	Null
Levis	Western	Western	Null
Mazda	Mazda2	Mazda2	Null
Mazda	Mazda3	Mazda3	Null
Microsoft	Azure	Azure	Null
Microsoft	Office	Office	Null
Nike	Jordan	Jordan	Null

Табл. 3. Фрагмент некоторой продукции, рекламируемой в Мексике
Table 3. Fragment of some products advertised in Mexico

Ключ	**Тип**	**Категория**
5G	Internet	Internet
Americano	Deportes	Entretenimiento
Basquetbol	Deportes	Entretenimiento
Béisbol	Deportes	Entretenimiento
Chamarra	Ropa	Vestimenta
Chico	Talla	Vestimenta
Compacto	Autos	Automóvil
Ellos	Género	Social
Familia	Género	Social
Grande	Talla	Vestimenta
Hatchback	Autos	Automóvil
Jeans	Ropa	Vestimenta
Laptops	Electrónicos	Electrónica
Licuadora	Electrónicos	Electrodomésticos
Smartphones	Electrónicos	Electrónica

Эти слова и символы, составляющие регулярные выражения, обычно используются в рекламе в Интернете [25].

Кроме того, поскольку реклама играет не только с визуальными эффектами, но и с буквами, размерами и стилями, диапазон поиска был расширен, с тем чтобы охватить множественные числа слов, акценты и символы, относящиеся к некоторым ключевым словам.

Важно отметить, что эти регулярные выражения брендов и продуктов связаны с исследованиями брендов в Мексике и некоторыми исследованиями в Латинской Америке. Для Мексики статистические данные за 2019 год были найдены в базах данных Национального института статистики и географии (Instituto Nacional de Estadística y Geografía, INEGI) – мексиканской правительственной организации, осуществляющей сбор и

систематизацию статистической, демографической, географической и экономической информации о стране. Другим источником данных была Экономическая комиссия для Латинской Америки и Карибского бассейна (Economic Commission for Latin America and the Caribbean, ECLAC/ЭКЛАК), которая является органом Организации Объединенных Наций, предоставляющим доступ к информации в некоторых странах Латинской Америки.

3.4 Псевдокод

Итак, основная идея алгоритма заключается в использовании автоматической прокрутки для захвата информации, содержащейся на Web-страницах. Затем эти захваченные изображения обрабатываются и преобразуются в текстовый формат, чтобы впоследствии идентифицировать существующие рекламные объявления на основе фильтров и сопоставлений с задаваемыми в виде регулярных выражений семействами слов, используемыми в цифровом маркетинге. Наконец, выявляются наиболее часто рекламируемые бренды, которые лидируют в поисковых запросах через Web-браузеры. Это полезно для их отслеживания. На рис. 3 представлен псевдокод разработанного алгоритма.

```
Open a browser with selenium
Enter the URL of the desired page
iteration = 1

while true:

    Page Height = Height of the web page in pixels
    according to selenium function
        Height = High computer screen size in pixels
            Slip = Height * iteration
                Capture with selenium tools
                    Save image with the capture number name
                        Slide page according to the number
                        of pixels indicating that Slide
    if Slip> = Page Height:
            break
                iteration +=1
Close selenium browser
Search file where captures were saved
Add in a list all the paths of the elements of this file
Connection is made to the database

for i in range (capture path list):
    image = capture path list [i]
        list = pytesseract.image_to_string (
        img).upper().split ()
            #Separate the text string returned by
            pytesseract into many smaller strings
                #with atomic elements, that are words,
                because the segmentation process was
                carried out
                    #when a space is found. These words
                    are now capitalized.
    WordsFound = []
        for j in range (list):
                #We proceed to make queries to find each
                word in 'list' in the 3 tables of the base
                        Find = Result of queries looking
                        for list [j]
        if length (Find)! = 0:
                    WordsFound.append (list [j])
                        Delete Capture already analyzed
    print (Capture number analyzed)
        print (Found Words without repeating, their number
        of occurrences and precedence table)
```

Рис. 3. Псевдокод реализованного алгоритма
Fig. 3. Pseudocode of the implemented algorithm

В целом, как уже было описано в предыдущих пунктах, алгоритм состоит из трех основных этапов: i) необходимо получить URL-адрес страницы, а затем сделать снимки экрана с помощью автоматической прокрутки; ii) изображения скриншотов затем обрабатываются в том порядке, в котором они были сделаны, для преобразования их в текст с помощью OCR; и iii) текст анализируется, в результате получается список совпадений с регулярными выражениями, хранящимися в таблицах, характеризующий компании, которые размещают рекламу, продукты и стратегии.

Одним из ограничений работы является то, что в ней не используется Web-скрейпинг (Web Scraping), представляющий собой процесс автоматического сбора данных и информации из Интернета, которые затем анализируются для определенных нужд и целей [25]. Кроме того, не учитывались ни персональные расширения браузеров, ни связанные учетные записи для синхронизации с устройствами. Кроме того, не анализировалась реклама с боковым выдвижением, потому что слайдинг в нашем случае вертикальный, сверху вниз.

4. Результаты

Для анализа веб-рекламы были рассмотрены три типа динамических Web-страниц, протестированных в трех разных браузерах: Google Chrome, Mozilla Firefox и Safari, разработанный компанией Apple. Были проанализированы следующие сайты:

- MSN: www.msn.com/es-mx;
- Sanborns: www.sanborns.com.mx;
- AhorraSeguros: https://ahorraseguros.mx.

В Табл. 4 обобщаются результаты, полученные для каждого URL-адреса в каждом Web-браузере, общее количество слов (регулярных выражений), которые появляются в рекламе на каждом оцениваемом Web-сайте, количество снимков экрана и время выполнения с момента открытия Web-браузера до завершения сравнения.

Табл. 4. Результаты выполнения программы
Табл. 4. Results of the algorithm evaluation

Web-сайт	**Web-браузер**	**Реклама**	**Количество скриншотов**	**Время (сек)**
URL 1, MSN	Chrome	106	9	11.636
	Firefox	103	9	12.144
	Safari	118	9	14.539
URL 2, Sanborns	Chrome	56	4	4.449
	Firefox	62	4	4.036
	Safari	68	4	4.209
URL 3, Ahorra Seguros	Chrome	133	9	13.547
	Firefox	149	9	14.547
	Safari	153	9	14.556

Как показывает Табл. 4, наибольшее количество рекламных объявлений алгоритм обнаружил при использовании браузера Safari. Такая лучшая идентификация рекламы обусловлена тем, что в этом случае алгоритм лучше выравнивает содержимое Web-страницы, и на обработку уходит больше времени. В Chrome и Firefox также обнаружено значительное количество рекламных объявлений, но некоторая информация была потеряна при прокрутке Web-страницы.

Для определения эффективности алгоритма был выполнен визуальный анализ информации, содержащейся на оцениваемых Web-страницах. Этот анализ заключался в подсчете RegEx в Web-рекламе, результаты которого должны соответствовать общему количеству слов,

обнаруженных алгоритмом. В табл. 5 приведены результаты, полученные вручную и автоматически при выполнении нашего алгоритма.

Табл. 5. Слова, идентифицированные алгоритмом по отношению к общему количеству слов с рекламным содержанием

Table. 5. Words identified by the algorithm with respect to the total of words with advertising content

Web-сайт	Раздел сайта	Визуальный подсчет	Chrome	Firefox	Safari
URL 1, MSN	Microsoft	22	22	14	22
	Новости	9	9	9	9
	IOS	10	10	0	10
	Android	10	10	0	10
	MSN	12	3	10	2
	Rebaja	15	13	12	14
	Всего	**124**	**106**	**103**	**118**
URL 2, Sanborns	$	27	18	23	24
	Libros	3	2	1	3
	Perfumes	2	2	2	2
	Tecnología	3	3	3	3
	Videojuegos	2	2	2	2
	Всего	**75**	**56**	**62**	**68**
URL 3, Ahorra Seguros	Seguros	57	48	54	56
	Seguro	22	22	22	21
	Beneficios	7	5	7	7
	Precios	5	4	5	4
	Servicios	5	2	1	4
	Всего	**172**	**133**	**149**	**153**

В случае www.msn.com/es-mx, URL 1 была достигнута отличная производительность алгоритма обнаружения Web-рекламы через браузер Safari с достоверностью 95,16%. В то же время браузеры Chrome и Firefox также достигли значительной достоверности в 85,48 и 83,06% соответственно.

Для URL 2, www.sanborns.com.mx при работе через Safari получена достоверность 90,66%, для Chrome – 74,66% и для Firefox – 82,66%. Эти результаты обусловлены удивительным визуальным дизайном сайта, который понятен людям, но сложен для анализа из-за наложения некоторых слов, а также наличия логотипов, а также слов разного размера и с разными шрифтами, связанных друг с другом.

Для случая https://ahorraseguros.mx, URL 3 также были достигнуты значительные результаты с достоверностью 88,95% в Safari, 86,62% в Firefox и 77,32% в Chrome. Более высокий уровень достоверности не был достигнут из-за большого количества логотипов разных брендов с различными шрифтами и фоном. Это явилось основной проблемой при оптическом распознавании символов.

Специфика результатов тестов обусловлена обрезкой и выравниванием экрана при автоматическом скольжении, то есть в каждом Web-браузере конфигурация варьируется, изменяя способ организации информации на Web-сайте, что и вызывает потери содержимого.

6. Заключение

Достижения современных технологий имеют также и отрицательные стороны для пользователя, такие как наплыв рекламы в Интернете – персонализированные рекламные объявления, в которых учитываются потребности и интересы пользователей.

Интернет-реклама состоит не только из слов или предложений, привлекающих внимание пользователя к рекламным акциям или предложениям. Для рекламодателей важно, чтобы пользователь знал, кто занимается продвижением товаров и услуг, независимо от того,

действительно ли клиент планирует купить рекламируемый продукт. Для рекламодателей самое главное – привлечь внимание пользователя, чтобы он запомнил бренд для будущих покупок.

Использование регулярных выражений в нашей работе было полезным. Кроме того, реализация базы данных облегчила организацию обнаружения Web-рекламы, позволяя охватить больше случаев использования рекламных объявлений.

Для проведенных тестов были получены ожидаемые результаты с достоверностью от приемлемого до высокого, в диапазоне от 74,66% до 95,16%, а самый высокий уровень достоверности был получен при использовании MSN благодаря простому дизайну, использованию распространенных шрифтов постоянного размера, отсутствию составных слов, логотипов и т.д.

В будущем предполагается включить в базу данных больше регулярных выражений и произвести расширение алгоритма, то есть включить алгоритмы искусственного интеллекта, способные распознавать рекламу на основе цветовых узоров, размера и расположения баннеров и других возможностей современных рекламодателей.

Список литературы / References

[1]. Marketing Digital. What is digital marketing? URL: https://www.mdmarketingdigital.com/en/what-is-digital-marketing, accessed 12/01/2020.

[2]. Redes Semánticas. URL: http://tesis.uson.mx/digital/tesis/docs/9049/Capitulo1.pdf, accessed 12/01/2020 (in Spanish).

[3]. Marketing Online: Potencial y Estrategias, 2019. URL: https://www.cecarm.com/Guia_Marketing_Online_Potencial_y_Estrategias_-_CECARM.pdf-6120, accessed 12/01/2020 (in Spanish).

[4]. Pomol R., González C., González S. Una herramienta didáctica para el aprendizaje interactivo de expresiones regulares. 2013. URL: http://repositorio.uigv.edu.pe/handle/20.500.11818/804, accessed 12/01/2020 (in Spanish).

[5]. Beltrán R. El uso de expresiones regulares en la detección de errores escritos: implicaciones para el diseño de un corrector gramatical, 2008. URL: https://dialnet.unirioja.es/servlet/articulo?codigo=4007478, accessed 12/01/2020 (in Spanish).

[6]. Gallego A. La jerarquía de Chomsky y la facultad del lenguaje: consecuencias para la variación y la evolución. Teorema: Revista internacional de filosofía, vol. 27, no. 2, 2008, pp. 47-60 (in Spanish).

[7]. García Campos I., Herramienta para la corrección automática de autómatas finitos, 2017. URL: https://riull.ull.es/xmlui/handle/915/5846, accessed 12/01/2020 (in Spanish).

[8]. Sánchez J., López L., Martínez J. Solución para garantizar la privacidad en el Internet de las Cosas. El profesional de la información, vol. 24, no 1, 2015, pp. 62-70 (in Spanish).

[9]. Ortiz M., Aguilar L., Marín L. Los desafíos del marketing en la era del big data. e-Ciencias de la Información, vol. 6, no. 1, 2016, pp. 1-30 (in Spanish).

[10]. Riaño D., Molero-Castillo G., Velázquez-Mena A., Bárcenas E. Expresiones regulares para el tratamiento de privacidad de navegadores Web. Abstraction and Application, vol. 25, 2019, pp.121-130 (in Spanish).

[11]. Cerezo, P., Ad blocking: el modelo publicitario digital, a revisión, Cuadernos de periodistas: revista de la Asociación de la Prensa de Madrid, 2016, pp. 81-89 (in Spanish).

[12]. Londaitz A., Publicidad en los celulares: Publicidad invasiva vs. derecho a la privacidad. Thesis, Universidad del Salvador, 2011.

[13]. Bienvenido a Google, la mejor empresa para trabajar, 2013. URL: www.expansion.com/2013/08/23/directivos/1377273795.html, accessed 12/01/2020 (in Spanish).

[14]. Jarvis J. Y Google, ¿cómo lo haría?, 2000. URL: https://narrativabreve.com/2013/10/libro-google-jeff-harvis.html, accessed 12/01/2020 (in Spanish).

[15]. Leotta M., Clerissi D., Ricca F., Spadaro C. Comparing the maintainability of selenium webdriver test suites employing different locators: A case study. In Proc. of 1st International Workshop on Joining AcadeMiA and Industry Contributions to Testing Automation, 2013, pp. 53-58

[16]. Gojare S., Joshi R., Gaigaware D., Analysis and Design of Selenium WebDriver Automation Testing Framework, Procedia Computer Science, vol. 50, 2015, pp. 341-346.

[17]. Selenium Webdriver, 2017. URL: www.tutorialspoint.com/selenium/pdf/selenium_webdriver.pdf, accessed 12/01/2020.

[18]. Yih W., Goodman J., Carvalho V. Finding Advertising Keywords on Web Pages, In Proceedings of the 15th International Conference on World Wide Web, 2006, pp. 213-222.
[19]. Mei T., Li L., Tian X., Tao D., Ngo C. PageSense: Toward Stylewise Contextual Advertising via Visual Analysis of Web Pages. IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 1, 2018, pp. 254-266.
[20]. Sánchez D., Viejo A. Privacy-preserving and advertising-friendly web surfing. *Computer Communications*, vol. 130, 2018, pp. 113-123.
[21]. Krammer V. An Effective Defense against Intrusive Web Advertising. In Proc. of the Sixth Annual Conference on Privacy, Security and Trust, 2008, pp. 3-14.
[22]. Sajjad K. Automatic license plate recognition using Python and Opencv. College of Engineering, 2010. URL: https://pdfs.semanticscholar.org/bddf/1200eb17f239e4dce2a9cec938eb8cf305f5.pdf, accessed 12/01/2020.
[23]. Patel C., Patel A., Patel D. Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications, vol. 55, no. 10, 2012, pp. 50-56.
[24]. Vallez M. Keyword Research: métodos y herramientas para identificar palabras clave. BiD: textos universitaris de biblioteconomia i documentació, vol. 27, 2011, pp. 1-14 (in Spanish).
[25]. Slamet C., Andrian R., Maylawati D. et al. Web Scraping and Naïve Bayes Classification for Job Search Engine. In Proc. of the 2nd Annual Applied Science and Engineering Conference, 2018, pp. 1-7.

Информация об авторах / Information about authors

Донован РИАНЬО-ЭНРИКЕС, студент.

Donovan RIAÑO-ENRIQUEZ, Student.

Родриго ПИНОН-АЯЛА, студент.

Rodrigo PINON-AYALA, Student.

Гильермо МОЛЕРО-КАСТИЛЬО, кандидат наук, доцент кафедры вычислительной техники. Область научных интересов: искусственный интеллект, наука о данных, интеллектуальный анализ данных, машинное обучение.

Guillermo MOLERO-CASTILLO, Ph.D. in Information Technologies, Associate Professor, Computing Engineering Department. Research interests: Artificial Intelligence, Data Science, Data Mining, Machine Learning.

Эверардо БАРСЕНАС, кандидат наук, доцент. Область научных интересов: модальная логика, теория доказательств, автоматизированные рассуждения, логика описания, model checking, представление знаний, планирование, компьютерное зрение.

Everardo BARCENAS, Ph.D., Assistant Professor. Research interests: modal logics, proof theory, automated reasoning, description logics, model checking, knowledge representation, planning, computer vision.

Алехандро ВЕЛАСКЕС-МЕНА, магистр, доцент. Область научных интересов: сетевой компьютер, туманные вычисления, распределенные вычисления.

Alejandro VELAZQUEZ-MENA, Master of Science, Assistant Professor. Research interests: Network Computer, Fog Computing, Distributed computing.