# ВЕБ-СКРАПИНГ И ЕГО ЗНАЧЕНИЕ ДЛЯ ЭКОНОМИКИ

**Ашурметова Н.А**
*к.э.н., доцент*
**Абдукаримов К.Ш**
*студент*
**Мирхосилов У.А**
*студент*
*Ташкентский государственный аграрный университет*

**Аннотация:** *Сегодня предприятия, фирмы, компании по всему миру собирают информацию с Веб-сайтов для анализа и принятия эффективных решений для своего бизнеса. Веб-скрапинг позволяет компаниям выгружать огромные массивы неструктурированных данных и структурировать их для дальнейшего анализа и извлечения ценной информации. В статье рассмотрены вопросы использования веб-скрапинга для экономики, в частности, для сбора необходимой информации субъектами бизнеса, которые можно использовать для сравнения, проверки и анализа на основе потребностей и целей данного бизнеса.*

**Ключевые слова:** *Веб-сканирование, Веб-скрапинг, цифровая экономика, технологии, экономика, эффективность.*

**Abstract:** *Today, enterprises, firms, companies around the world collect information from Websites to analyze and make effective decisions for their business. Web scraping allows companies to download huge amounts of unstructured data and structure it for further analysis and extraction of valuable information. The article discusses the use of web scraping for the economy, in particular, for collecting the necessary information by business entities, which can be used for comparison, verification and analysis based on the needs and goals of a given business.*

**Keywords:** *Web crawling, Web scraping, digital economy, technology, economics, efficiency.*

Сегодня в век цифровых технологий, весь мир уверенно шагает в сторону цифровизации, которая охватывает все сферы и области, не остается в стороне и экономика. Цифровая экономика - это, в первую очередь, работа с байтами информации, огромным количеством данных, которые загружаются и хранятся в интернет-пространстве. Заслуживает внимания тот факт, что государство, фирмы и индивиды выгружают и используют открытые данные, ведут «прозрачную» деятельность, что несколько десятилетий назад просто казалось чем-то невозможным. Именно благодаря непрерывной цифровой революции стоимость хранения и передачи данных снизились до такой степени, что маржинальные (предельные, дополнительные) издержки практически равны нулю. По данным IDC

(International Data Corporation - компания, предоставляющая технологические медиа данные и маркетинговые услуги), общий объем созданных и хранимых совокупных данных возрос с 0,8 Зеттабайтов (ЗБ) или триллион гигабайтов в 2009 до 33 ЗБ в 2018 и ожидается что этот показатель возрастет до 175 ЗБ в 2025 году. [3]

В результате высоких темпов роста объемов информационных ресурсов, в частности научной информации, которая в 2020 году превысила 35 триллионов гигабайтов данных, создано множество новых данных, которые позволяют не только проверять устоявшиеся экономические гипотезы, но и решать вопросы человеческого взаимодействия, не опробированные за пределами лаборатории. Эти новые источники данных включают социальные сети, различные краудсорсинговые проекты (например, Википедию), приложения GPS-слежения, статические данные о местоположении или спутниковые снимки. А появление Интернета вещей объем данных, доступных для парсинга, еще больше увеличился (CTA).

Рост доступных данных совпадает с огромными преобразованиями в технологиях и программном обеспечении, используемых для их анализа. Методы искусственного интеллекта (ИИ) позволяют исследователям находить существенные закономерности в больших объемах данных любого типа, полезную информацию не только в таблицах данных, но также в неструктурированном тексте или даже в изображениях, голосовых записях и видео. Использование этих новых данных, ранее недоступных для количественного анализа, позволяет исследователям задавать новые вопросы и помогает избежать систематической ошибки из-за упущенных переменных, включая информацию, которая была известна другим, но не была включена в наборы данных количественного исследования.

Большая часть полученных данных хранится на частных серверах, но значительная часть становится общедоступной на 1,83 миллиарда веб-сайтов, доступных также исследователям, обладающим базовыми навыками веб-скрапинга. Веб-скрейпинг - это технология получения веб-данных путём извлечения их со страниц веб-ресурсов. [2]

Многие онлайн-сервисы, используемые в повседневной жизни, включая поисковые системы, агрегаторы цен и новостей, были бы невозможны без парсинга веб-страниц. Фактически, даже Google, самая популярная поисковая система по состоянию на 2021 год, представляет собой всего лишь крупномасштабный веб-сканер. Автоматизированный сбор данных также используется в бизнесе, например, для исследования рынка и привлечения потенциальных клиентов. Поэтому неудивительно, что этот метод сбора данных также привлекает все большее внимание в социальных исследованиях.

Если ai (artificial intelligence) или ml (machine learning) помогают конечному пользователю обрабатывать и генерировать данные видео и аудио формата, Веб-скрапинг используется преимущественно для сбора текста с веб-сайтов. Используя веб-скрапинг, исследователи могут извлекать данные из различных источников для

создания индивидуальных наборов данных, соответствующих индивидуальным исследовательским потребностям. [1]

Примером таких данных может служить информация, находящаяся в открытом доступе, к примеру Сайт Агентства статистики при Президенте Республики Узбекистан, где можно найти многие статистические данные по экономике, такие как ВВП страны, информация об инфляции, безработице и других важных макроэкономических показателей.

Благодаря веб-скрапингу, данные, которые до недавнего времени были доступны в агрегированной форме и с большой задержкой, теперь доступны в режиме реального времени и со значительно большей детализацией, чем то, что традиционно предлагалось статистическими управлениями или поставщиками коммерческих данных. Например, при изучении цен парсинг веб-страниц, выполняемый в течение длительных периодов времени, позволяет собирать данные о ценах от всех поставщиков в данной области с подробной информацией о продукте (включая идентификатор продукта) с желаемой степенью детализации. Исследования рынка труда или рынков недвижимости выигрывают от извлечения информации из подробных описаний объявлений.

Помимо этого, многие статистические управления начали использовать данные, полученные из Интернета, например, для расчета индексов цен. Однако качество данных, выборка и репрезентативность являются серьезными проблемами, как и правовая неопределенность в отношении конфиденциальности. Хотя это справедливо для всех типов данных, большие данные усугубляют проблему: большая часть создаваемых больших данных не является конечным продуктом, а скорее побочным продуктом других приложений.

Что делает парсинг веб-страниц возможным и относительно простым, так это регулярная структура кода, используемого для веб-сайтов, предназначенных для отображения в веб-браузерах. Веб-сайты, созданные с использованием HTML, можно очищать с помощью стандартных инструментов анализа текста: либо сценариев на популярных (статистических) языках программирования, таких как Python или R, либо автономных специализированных инструментов веб-анализа. Некоторые из этих инструментов даже не требуют каких-либо предварительных навыков программирования.

Для того, чтобы понять, как работает web scrapping и изучить методы применения данной технологии в экономике, нужно разобраться, что такое html (hyper text markup language) - стандартизированный язык гипертекстовой разметки документов для просмотра веб страниц в браузере. Документ формата html - это текстовый файл, который содержит определенный синтаксис, показывающий компьютеру и серверу что находится внутри этого файла и как нужно читать данный файл.

Веб страница состоит из html элементов (рис.1). С помощью этих элементов задается структурная семантика текста (заголовки, абзацы, списки, ссылки, цитаты и другие элементы). Помимо этого, используя конструкции html изображения, интерактивные формы и другие объекты могут быть встроены в отображаемую страницу.

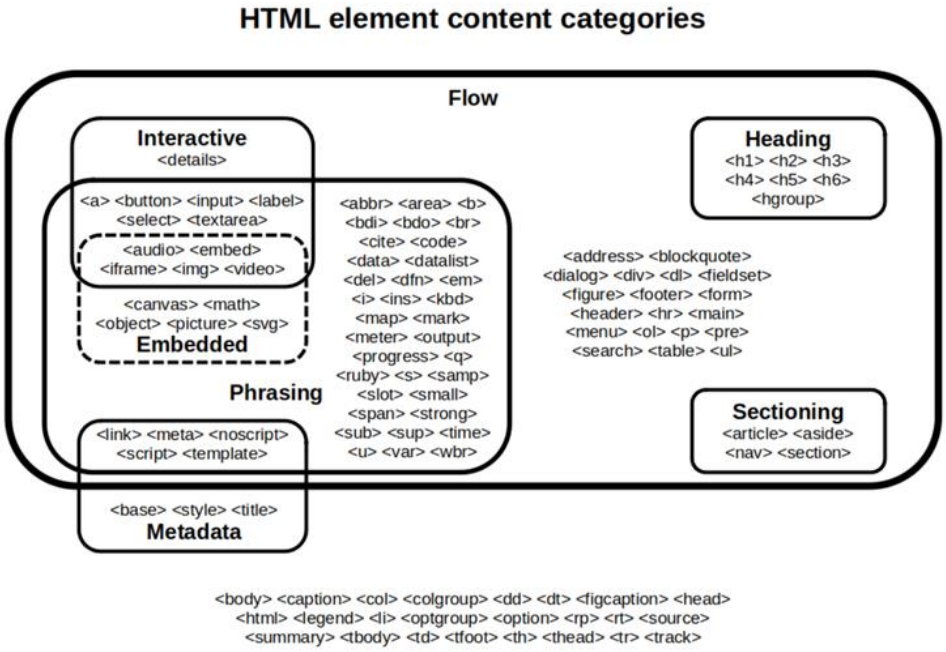


**Рис. 1. Элементы html документа**

Таким образом, Веб-скрапинг - это автоматический метод получения больших объемов данных с веб-сайтов. Большая часть этих данных представляет собой неструктурированные данные в формате HTML, которые затем преобразуются в структурированные данные в электронной таблице или базе данных, чтобы их можно было использовать в различных целях. Существует множество различных способов парсинга веб-страниц для получения данных с веб-сайтов. К ним относятся использование онлайн-сервисов, определенных API или даже создание кода для парсинга веб-страниц с нуля. Многие крупные веб-сайты, такие как Google, Twitter, Facebook, StackOverflow и др., имеют API, которые позволяют нам получать доступ к их данным в структурированном формате. Рассмотрим некоторые варианты использования веб скрапинга в экономической деятельности компаний:

1. Мониторинг цен. Веб-скрапинг может использоваться компаниями и фирмами для сбора данных о своих и конкурирующих продуктах, а также для того, чтобы увидеть, как это влияет на их ценовую стратегию. Компании могут использовать эти данные для установления оптимальных цен на свою продукцию и получения максимального дохода.

2. Исследование рынка. Веб-скрапинг может использоваться компаниями для исследования рынка. Высококачественные веб-данные, полученные в больших объемах, могут быть очень полезны компаниям для анализа потребительских

тенденций и понимания, в каком направлении компания должна двигаться в будущем.

3. Поиск трудового капитала. Веб-скрапинг может использоваться компаниями и фирмами для поиска трудового капитала. Примерно с 2010 года, когда все большее число работодателей и соискателей работы полагаются на онлайн-порталы вакансий для рекламы и поиска работы, исследователи все чаще идентифицируют онлайн-рынок труда как новый источник данных для анализа динамики и тенденций рынка труда.

Огромное количество доступных в Интернете данных практически по каждой теме должно привлечь интерес всех исследователей. Поскольку все большая часть нашей повседневной деятельности перемещается в Интернет, поиск информации в Интернете станет единственным способом найти информацию о значительной части человеческой деятельности.

Данные, собранные с помощью веб-скрапинга, использовались в тысячах проектов и привели к лучшему пониманию ценообразования, механизмов аукционов, рынков труда, социальных взаимодействий и многих других важных тем. Поскольку новые данные регулярно загружаются на различные веб-сайты, старые ответы можно проверить в различных условиях и поставить новые исследовательские вопросы.

Интегрирование веб скрапинга в экономическую деятельность компаний Узбекистана откроет огромные возможности для повышения эффективности персонала и увеличения прибыли как больших компаний, так и средних и малых предприятий.

**ЛИТЕРАТУРА:**

1. Шаг в будущее: искусственный интеллект и цифровая экономика. Революция в управлении: новая цифровая экономика или новый мир машин [Текст] : материалы II Международного научного форума. Вып. 4 / Государственный университет управления. – М.: Издательский дом ГУУ, 2018. – 478 с.

2. Boeing, G.; Waddell, P. New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings // Journal of Planning Education and Research. — 2016. — doi:10.1177/0739456X16664789. — arXiv:1605.05397.

3. https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html