Цифровая экономика

ТЕХНИЧЕСКИЕ И МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ СБОРА ДАННЫХ О ЦЕНАХ ОНЛАЙН-РИТЕЙЛЕРОВ

А. С. ЕВСЕЕВ
Р. Р. ЛАТЫПОВ
Е. А. ПОСТОЛИТ
Е. С. СИНЕЛЬНИКОВА-МУРЫЛЕВА

Данные о ценах онлайн-ритейлеров обладают огромной ценностью с точки зрения экономической науки: их использование позволяет уточнять прогнозы инфляции и предвосхищать будущие тенденции в моменте, корректировать оценки жесткости цен и выводы теоретических моделей ценообразования, проверять закон единой цены и т.д. Однако в процессе сбора данных возникают серьезные трудности, которые являются неочевидными и могут угрожать как качеству собираемых данных, так и устойчивости процесса их сбора во времени.

В статье впервые подробно обсуждаются технические и методологические проблемы, которые препятствуют непрерывному сбору данных в Сети, а также представлен опыт решения этих проблем; обсуждаются плюсы и минусы таких решений.

Статья подготовлена в рамках выполнения научно-исследовательской работы государственного задания РАНХиГС при Президенте Российской Федерации на 2022 год.

Ключевые слова: цены онлайн-ритейлеров, парсинг, инфляция, альтернативные данные, big data.
JEL: C81, C82, E31, C55.

Введение

В современных условиях растет количество данных, размещенных в Сети, в том числе тех, что представляют интерес для макроэкономистов. Тысячи ритейлеров, маркетплейсов, платформ с объявлениями о продаже предлагают свои товары и услуги и ежедневно обновляют информацию о ценах. Опыт зарубежных исследователей (например, [3; 4; 7; 10 и др.]) показал, что данные о ценах, собранные с онлайн-площадок, имеют высокую ценность, поскольку позволяют повышать точность прогнозов традиционных моделей прогнозирования инфляции (особенно на коротких горизонтах прогнозирования [3]), исследовать свойства ценообразования [6], проверять справедливость официальных оценок инфляции [4] и узнавать прочие факты как о поведении цен на онлайн-рынке, так и в целом в экономике.

Вместе с тем до сих пор доступ к текущим и историческим данным о ценах онлайн-ритейлеров затруднен — большинство таких данных либо закрыты, либо предлагаются платно (как, например, данные о ценах платформы Яндекс.Маркет), либо доступны для ограниченного числа стран и ритейлеров. Крупнейшим проектом по сбору цен на отдельные товары в мире, насколько нам известно, является проект The Billion Prices Project, данные которого охватывают цены на товары нескольких крупных ритейлеров для ряда развитых и развивающихся стран мира с 2007 г. по настоящее время [7]. Авторы проекта отмечают, что парсинг (извлечение) данных о ценах сопряжено с относительно низкими издержками (по сравнению с традиционным способом сбора данных национальными статистическими агентствами по физическим магазинам).

Евсеев Алексей Сергеевич, научный сотрудник РАНХиГС при Президенте Российской Федерации (Москва), e-mail: evseevals@ranepa.ru; *Латыпов Родион Ринатович*, главный экономист по России АО «Арована Капитал» (Москва), e-mail: Rodion.Latypov@mncap.ru; *Постолит Егор Анатольевич*, аналитик АО «Арована Капитал» (Москва), e-mail: postolitegor@gmail.ru; *Синельникова-Мурылева Елена Сергеевна*, старший научный сотрудник РАНХиГС при Президенте Российской Федерации, канд. экон. наук (Москва), e-mail: e.sinelnikova@ranepa.ru

Основываясь на преимуществе таких данных, особенно на низких издержках их сбора, в РАНХиГС (по 33 продовольственным категориям Росстата) [1] и в сотрудничестве с АО «Арована Капитал» (по всему индексу потребительских цен Росстата) [2], нами был запущен процесс сбора цен по российским онлайн-площадкам, и на момент написания настоящей статьи Росстат тестирует собранные нами данные в рамках своих экспериментов для построения индексов цен.

Наш опыт сбора таких данных показал, что его издержки действительно ниже, чем при традиционном сборе данных о ценах (как минимум, в денежном выражении), однако все же являются не такими уж незначительными. Наша цель — познакомить читателя с теми трудностями, с которыми сталкивается исследователь с экономическим образованием и не обладающий высокими техническими компетенциями в области программирования. Мы расскажем о ряде принципиальных проблем, которые, на наш взгляд, являются весьма острыми для исследователя как на начальных, так и на последующих этапах сбора таких данных. Кроме того, мы предложим оптимальные, на наш взгляд, решения этих проблем.

В статье будут подробно обсуждаться проблемы поддержания работоспособности парсеров, замещения выпавших (вследствие окончательного прекращения продажи) товаров на аналогичные и новые, экологичного (не нагружающего сервер ритейлера) для ритейлеров обхода блокировок, классифицирования товаров и услуг в систему товаров- и услуг-представителей Росстата, отбора ритейлеров, а также проблема определения конкретного набора товаров и услуг для разных исследовательских задач (например, для получения оперативных оценок текущей инфляции). Впервые акцентируется внимание на проблеме устойчивости проекта во времени.

Представленный в статье анализ возможных проблем сбора данных о ценах онлайн-ритейлеров в интернете и путей их решения является первой известной нам попыткой систематизировать опыт сбора таких данных и может представлять интерес не только для исследователей, занимающихся изучением поведения цен и прогнозированием инфляции, но также для тех, кто планирует заниматься автоматизированным сбором данных любого типа в Сети на регулярной основе.

В первом разделе мы коротко рассказываем о механизме автоматизированного сбора данных и о том, какие данные мы собрали за исследуемый период; во втором разделе описываем проблемы, с которыми мы столкнулись, а в заключении резюмируем содержание и выводы статьи.

Механизм автоматизированного сбора и описание собранных данных

В рамках настоящей статьи рассматриваются две базы данных: одна из них собирается в РАНХиГС с 1 февраля 2019 г. (далее — база № 1), а вторая — командой АО «Арована Капитал» — РАНХиГС с 20 июля 2020 г. (база № 2).

Механизм сбора заключается в следующем. Исследователи при поддержке технических специалистов написали коды программ на Python, которые запускаются на удаленном сервере ежедневно. В процессе работы этих программ собирается информация о текущей цене товара, зачеркнутой цене (если товар продается со скидкой), названии товара, ссылке на товар, а также в ряде случаев собирается информация о дополнительных характеристиках, таких как вес и цвет товара, размер упаковки и т.д.

Механизм обработки страниц различается в зависимости от базы данных. В рамках базы № 1 речь идет о сборе данных по отдельным продуктовым страницам, внутри которых содержится информация об одном конкретном товаре, а именно информация о названии товара, его актуальной стоимости, старой цене (в случае, если товар продается со скидкой), а также единице измерения — информация по последнему пункту доступна или из названия товара, или из его цены.

В этой базе данных текущий сбор осуществляется преимущественно среди продуктовых

онлайн-ритейлеров и охватывает конкретные товары, относящиеся к 33 продуктам условного (минимального) набора продуктов Росстата. Этот выбор был продиктован необходимостью сбора данных по узкому, но в то же время репрезентативному набору продуктов, составляющих значительную долю в расходах на все продовольствие, потребляемое населением. Помимо этих позиций в базе есть исторические данные о ценах на ряд непродовольственных товаров и услуг, большая часть из которых входит в фиксированный набор Росстата.

Главным преимуществом такого подхода является строгое отслеживание ценовой и неценовой информации по конкретным наименованиям товаров или услуг на протяжении всего периода сбора данных. Главный его недостаток – необходимость «ручного» отслеживания того, не исчез ли продукт из продажи, и замены навсегда выпавших товаров на новые продуктовые страницы.

В рамках базы № 2 охватывается большее число ритейлеров. Механизм сбора данных устроен иначе. Сначала программа собирает ссылки на все подкаталоги внутри корневого каталога сайта, затем «ходит» по всем страницам этих подкаталогов и собирает информацию о товарах на страницах этих каталогов, которые, как правило, находятся в карточках товаров. Поскольку иногда ритейлеры добавляют, убирают или изменяют структуру сайта и дерево подкаталогов, программа с периодичностью раз в неделю вновь собирает ссылки на все подкаталоги и цикл сбора продолжается по актуальным ссылкам.

У такого подхода есть большое преимущество: он позволяет аккумулировать цены буквально на все товары, представленные в подразделах каталога на сайте ритейлера. Таким образом, не нужно вручную отслеживать появление и исчезновение новых товаров. Вместе с тем данный подход требует обхода всех страниц с каталогом, что может существенно увеличивать временные затраты на парсинг и резко расширять базу данных «нерелевантными» товарами. Так, если исследователь ставит перед собой задачу сбора данных преимущественно по товарам–представителям Росстата, то при таком подходе ему придется совершить объемную работу по отделению нужных ему товаров от большой массы «нерелевантных».

У сбора данных по отдельным продуктовым страницам, как и у сбора по страницам каталога, есть общая проблема однозначной идентификации товара. Дело в том, что если наименование товара или ссылка на него, предположим «Арбуз Россия сочный 5–15 кг», изменится (по нашему опыту, к примеру, в названии может появиться название магазина), то будет трудно определить, что это тот же товар, а не исчезновение старого и появление нового. Следует сказать, что эта проблема в основном имеет значение для исследовательских задач, изучающих поведение цен на микроуровне.

Проблемы сбора данных

Наш опыт показал, что сбор данных с сайтов онлайн-ритейлеров сопряжен с рядом существенных проблем. Ниже будут представлены наиболее острые из них – на наш взгляд, это технические и методологические проблемы, с которыми может столкнуться исследователь, имеющий экономическое образование, но не обладающий соответствующими техническими навыками для написания сложных парсеров для сбора и обработки данных.

Методологические проблемы

Выбор ритейлеров

Первая проблема, с которой сталкивается исследователь при желании начать собирать данные из интернета, – это определение круга сайтов онлайн-ритейлеров для сбора соответствующей информации. Здесь выбор сильно привязан к тому, какую цель преследует формирование базы данных. Если речь идет о намерении формировать индекс цен, позволяющий как можно точнее делать выводы о поведении цен в обычных традиционных магазинах, аппроксимировать официальную инфля-

цию, то, на наш взгляд, имеет смысл придерживаться методологии, представленной в [7], а именно собирать данные о ценах с сайтов ритейлеров, которые реализуют одни и те же товары как онлайн, так и офлайн, но преимущественно вторым способом.

В [5] показано, что динамика цен на товары у таких ритейлеров в онлайн- и офлайн-сегментах во многом идентична, что позволяет с определенной долей уверенности судить по этим данным о поведении цен в традиционных точках продаж. На практике же при формировании нашей базы № 1 мы столкнулись с тем, что в условиях российской действительности доля мультиканальных ритейлеров в общем числе онлайн-ритейлеров не слишком высока, однако по возможности мы старались все же отбирать их.

Замещение наблюдений вследствие исчезновения товаров из продажи

Для расчета индекса потребительских цен во времени требуется непрерывный ряд для каждого товара- и услуги-представителя. Однако иногда в силу разных причин цены на наблюдаемые товары становятся недоступны. Поэтому требуется каким-то образом заменить пропущенные значения для обеспечения сопоставимости индекса цен во времени.

Для расчета индекса цен на самом низком уровне агрегирования (на уровне товаров- и услуг-представителей в конкретном городе N) Росстат собирает значения цен на 5–10 товаров и услуг с конкретными характеристиками. Так, например, для товара-представителя «Молоко питьевое цельное пастеризованное 2,5–3,2% жирности» он может собирать «Молоко ЭкоНива ультрапастеризованное 3,2%, 1 л БЗМЖ», «Село Зеленое Молоко питьевое ультрапастеризованное 3,2%, 950 мл» и еще 5–7 похожих видов конкретных марок молока с жирностью 3,2%. Если цена на товар отсутствует, Росстат рекомендует: «В ходе наблюдения… собирать данные о ценовых котировках на более широкий, чем включается в расчет ИПЦ, круг товаров (услуг) с конкретными потребительскими свойствами. Эта дополнительная совокупность товаров (услуг) может быть использована в качестве информационной базы при подборе замены в случае прекращения реализации наблюдаемого товара (услуги)»[1]. Кроме этого, Росстат в своей официальной методологии предлагает множество методов замены цен на отсутствующие товары в зависимости от природы товара (услуги) и причин его отсутствия.

Все эти методы позволяют Росстату обеспечивать высокую степень непрерывности рядов с индексом цен во времени, однако процедуры при этом сложны, поскольку требуют постоянного принятия все новых решений для каждой конкретной ситуации исчезновения товара. Вся эта сложность во многом объясняется ограничениями со стороны Росстата на количество наблюдаемых товаров с конкретными свойствами.

Между тем данные, собираемые с сайтов онлайн-ритейлеров, позволяют существенно снизить остроту проблемы с замещением пропущенных наблюдений, поскольку затраты на сбор дополнительных товаров с конкретными потребительскими свойствами при этом близки к нулю. Для многих продовольственных товаров собирается от 20 до 80 представителей, которые являются достаточно близкими заменителями. Наш опыт подтверждает результаты работы [7]: чем больше товаров (услуг) с конкретными свойствами собирается, тем ближе совокупный индекс цен по этой группе к динамике официального индекса.

Отнесение товара к категории

В настоящей статье мы исходим из той логики, что читатель/исследователь может интересоваться проблемой парсинга данных о ценах в первую очередь для расчета ИПЦ по этим

[1] Приказ Росстата от 30.12.2014 г. № 734 (ред. от 28.04.2021) «Об утверждении Официальной статистической методологии организации статистического наблюдения за потребительскими ценами на товары и услуги и расчета индексов потребительских цен» // Консультант Плюс. 2021. 28 апреля. URL: https://www.consultant.ru/document/cons_doc_LAW_174490

данным. В связи с этим важной задачей для него станет способ определения того, к какому товару- или услуге-представителю относится конкретный товар или услуга, данные по которому(ой) собираются им. В таблице ниже приведены примеры соотнесения конкретных товаров или услуг с соответствующими им товарами- или услугами-представителями.

В случае если мы собираем данные с отдельных продуктовых страниц, эта проблема не возникает, поскольку мы изначально вручную отбираем конкретные товары и услуги под группу нужных нам товаров- (услуг-) представителей. Такой подход весьма полезен для построения индекса цен, поскольку исследователь тщательно отбирает товары с подходящими для товара- (услуги-) представителя свойствами, что дает высокое приближение к официальной статистической методологии организации статистического наблюдения за потребительскими ценами на товары и услуги и расчета индексов потребительских цен. Однако этот подход обладает двумя большими недостатками. Во-первых, отбор продуктовых страниц сам по себе является затруднительным процессом, поскольку нужно найти товар, подходящий к товару- (услуге-) представителю, а во-вторых, такие страницы со временем будут «умирать» (либо вследствие изменения адреса страницы, либо вследствие исчезновения товара из продажи) и исследователю придется заново искать подходящие товары на замену, что тоже трудоемко.

Другим способом отнесения конкретных товаров и услуг к соответствующим им товарам- (услугам-) представителям является использование алгоритмов машинного обучения. Так, сотрудники АО «Арована Капитал» в качестве алгоритма используют логистическую регрессию [2] с кросс-валидацией. Главное достоинство такого подхода — возможность избежать высоких издержек от разметки сотен тысяч товаров вручную. Вместе с тем контроль качества разметки товаров на тестовой выборке все еще является отдельной и трудоемкой задачей. Однако работа над повышением качества разметки имеет смысл. На рисунке приведен пример того, насколько динамика онлайн-индекса ИПЦ становится ближе к динамике официального ИПЦ по мере повышения качества разметки и увеличения количества размеченных товаров- и услуг-представителей.

Географическая репрезентативность

Географическая репрезентативность — один из самых существенных недостатков сбора данных о ценах онлайн-ритейлеров по сравнению со сбором данных, осуществляемым официальными статистическими ведомствами. На сегодня сбор данных о ценах и для базы № 1, и для базы № 2 производится только по Московскому региону.

По нашим наблюдениям, онлайн-ритейлеры действуют в основном в городах-миллионниках. Однако интернет-торговля активно развивается. Так, в малых городах возникают пункты выдачи крупных онлайн-ритейлеров, данные по которым уже собираются нами. Последнее дает нам основания надеяться, что

Пример отнесения конкретных товаров (услуг) к соответствующим товарам- (услугам-) представителям

Товар- (услуга-) представитель	Конкретный товар (услуга), соответствующий этому товару- (услуге-) представителю
Молоко питьевое цельное пастеризованное 2,5-3,2% жирности	Молоко «Избенка» пастеризованное 3,2% 900 мл
	Молоко Valio 3,2% 1 л
Сметана	Сметана «Простоквашино» 20% 300 г
	Сметана «ЭкоНива» 20% 300 г

Источник: составлено авторами.

со временем проблема географической репрезентативности станет менее острой.

Технические проблемы

Обход блокировок

Большинство онлайн-ритейлеров стремятся защитить свои сайты от массированных DDoS-атак, для чего используют разные способы защиты: от простого ограничения на число запросов к серверу в секунду до полномасштабных проверок пользователя на то, является ли он роботом или нет. Такие защиты вполне разумны, поскольку магазинам необходимо предотвратить остановку сервера вследствие атак со стороны киберпреступников и недобросовестных конкурентов. Однако мы, как добросовестные исследователи, стремящиеся к невредоносному сбору нужных нам данных, также столкнулись с очень острой проблемой доступа к этим данным.

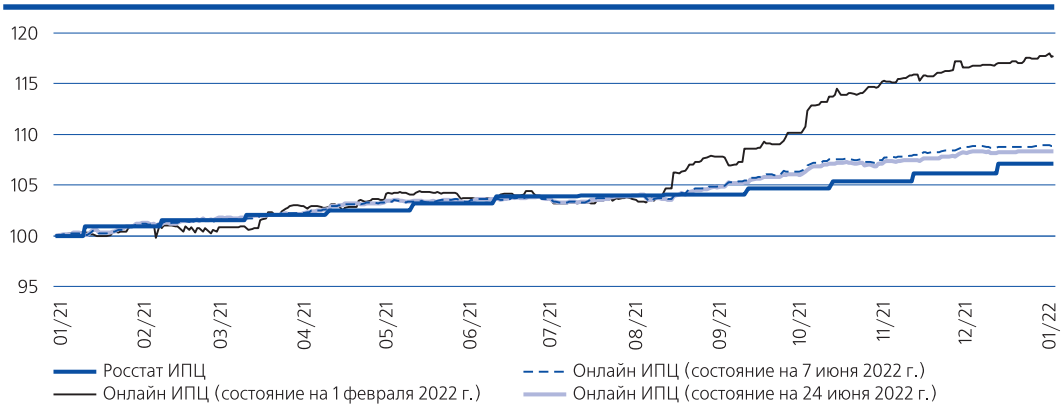
В самом начале наших экспериментов со сбором данных (февраль 2019 г.) проблема блокировок была в разы менее острой, чем на сегодня (октябрь 2022 г.), когда большинство крупных онлайн-ритейлеров стали устанавливать сильнейшие защиты для ограничения автоматизированного доступа к своим ресурсам.

Часть из используемых нами сайтов ритейлеров до сих пор не имеет серьезной защиты вовсе, но, как правило, это некрупные магазины, которые для нас не представляют большой исследовательской ценности в силу их малой репрезентативности. Поэтому для того, чтобы все-таки собирать данные от крупных ритейлеров с более серьезными защитами, нам пришлось изобретать пути обхода блокировок, которые, с одной стороны, не создавали бы для сайта ритейлера большой нагрузки и не приводили бы к остановке сервера, а с другой — позволяли бы продолжать непрерывный сбор данных без ручного вмешательства.

Проблема изменения разметки сайта

В момент написания программы для сбора данных специалист фиксирует, в каких местах сайта находится нужная ему информация, которая затем будет собираться на регулярной основе без его вмешательства. Эти места привязаны к HTML-разметке и названиям классов, id и прочих однозначных меток для извлечения этой информации в момент написания парсера. Так, например, цена на товар на момент написания парсера находится в элементе с названием класса «price». Со временем, однако, названия таких классов могут менять-

ИПЦ по г. Москве за период с 21 января 2021 г. по 21 января 2022 г. (официальный индекс vs онлайн-индекс, 21 января 2021=100)



Источник: данные базы № 2.

ся, чаще всего из-за оптимизации или изменения дизайна сайта. В таком случае информация о цене может находиться уже в другом классе, например с названием «item-price», в связи с чем жестко привязанный к названиям классов парсер перестает собирать информацию о цене и либо прерывает процесс сбора от всех ритейлеров, либо возвращает пустое значение.

Исходя из нашего опыта такая проблема возникает с разной периодичностью для разных сайтов. Так, например, для многих сайтов используемых нами продуктовых онлайн-ритейлеров разметка с 2019 г. не изменилась вообще, а для некоторых непродовольственных онлайн-ритейлеров она менялась буквально каждый день, так что приходилось вручную менять значения. На наш взгляд, изменение в разметке является одной из главных проблем обеспечения непрерывного и стабильного сбора данных.

Прочие технические проблемы

На практике мы сталкивались и с прочими, менее острыми проблемами. Так, например, один из продовольственных онлайн-ритейлеров изменил название сайта, что привело к необходимости коррекции всех ссылок на страницы с товарами. Кроме того, возникла проблема, связанная с добавлением новых маркировок молочных товаров (например, «Молоко 33 коровы 3,2% жирности» было заменено на «БЗМЖ Молоко 33 коровы 3,2% жирности», что временно привело к изменению ссылки и проблеме однозначной идентификации товара. Наконец, периодически возникают трудности из-за недоступности сайта в момент сбора данных — чаще всего это вызвано проведением на сайте технических работ. По нашему опыту, такого рода проблемы появляются редко и требуют внимательности специалиста, который занимается организацией сбора данных.

Отдельной проблемой являются мощности, необходимые для запуска парсеров. Поскольку собираются данные по значительному числу магазинов с тысячами страниц с каталогами, для покрытия масштабного сбора данных требуется развитая инфраструктура серверов, способных собирать данные без сбоев в течение адекватного времени.

Проблема устойчивости проекта во времени

В совокупности методологические и, в особенности, технические проблемы приводят к тому, что процесс сбора данных становится очень хрупким и трудозатратным. Издержки поддержания проекта особенно высоки с точки зрения привлечения сторонних специалистов, следящих, чтобы парсеры непрерывно и корректно собирали данные.

Вместе с тем наш опыт показывает, что выгоды от такого рода проектов тоже достаточно существенны. Накопленные данные позволяют тестировать гипотезы о различных макро- и микроэкономических тенденциях, связанных с ценами, что важно с точки зрения понимания взаимосвязей в экономике. Однако мы понимаем, что такие выгоды являются по большей части потенциальными, поскольку для проверки большинства гипотез и формирования серьезных научных выводов понадобятся годы наблюдений, что потребует вложения значительных ресурсов.

Большинство западных исследований, проверяющих гипотезы на данных онлайн-ритейлеров, основываются на данных научного проекта The Billion Prices Project Массачусетского университета, которые охватывают период с 2007 г. по настоящее время [7]. Столь продолжительный срок жизни этого проекта во многом объясняется его монетизацией: созданием коммерческого ответвления PriceStats. В остальных исследованиях, как правило, используются более короткие периоды наблюдений (до трех лет — например, в [8; 9; 12]).

Заключение

Результаты многих зарубежных исследований показывают, что данные о ценах онлайн-ритейлеров позволяют существенно улучшать

качество краткосрочных прогнозов по инфляции, отслеживать поведение цен на микроуровне, уточнять выводы теоретических моделей ценообразования и проч. Вместе с тем, как следует из нашего опыта, процесс сбора таких данных сопряжен с рядом методологических и технических проблем, которые требуют внимания и зачастую должны решаться самим экономистом-исследователем или при его участии, поскольку многие решения являются не только и не столько техническими, сколько содержательными.

Выбор источника данных, т.е. круга онлайн-ритейлеров, в значительной мере определяется задачей исследователя. Подход, представленный в [7] (отбор ритейлеров, реализующих товары и онлайн, и офлайн, но преимущественно офлайн), на наш взгляд, является оптимальным в случае, если исследователь стремится быть как можно ближе к официальной методологии расчета ИПЦ и распространять свои выводы на экономику в целом.

Еще одной значимой проблемой является замещение наблюдений вследствие исчезновения товаров из продажи. Наш опыт подтверждает выводы [7] о том, что эта проблема становится тем менее острой, чем больше однородных товаров (услуг) собрано в рамках товара- (услуги-) представителя.

Наконец, важно обозначить и проблему отнесения конкретных товаров и услуг к соответствующим им товарам- и услугам-представителям. Наш опыт показывает, что в случае ручного отбора отдельных товаров и услуг для сбора данных эта проблема не возникает, поскольку в момент отбора исследователь сам принимает решение, к какому товару- (услуге-) представителю отнести конкретный товар или услугу, однако такой подход для формирования выборки весьма трудоемок. В случае использования алгоритмов машинного обучения затраты ручного труда минимальны, однако качество разметки по таким алгоритмам требует постоянного контроля со стороны исследователя.

По нашему опыту, в процессе сбора данных возникает также ряд технических трудностей, среди которых нарушение механизма сбора из-за изменения разметки сайта, блокировка со стороны онлайн-ритейлера и прочие реже встречающиеся, но не менее значимые проблемы – к примеру, изменения названий товаров и изменение ссылок. Наш опыт свидетельствует о том, что среди всех технических проблем особенно острой является проблема блокировки, поскольку она требует все более технически сложных и одновременно не нагружающих сайт для парсинга способов обхода, и без специальных навыков с ней сложно справиться.

Несмотря на все перечисленные трудности, процесс сбора данных о ценах онлайн-ритейлеров обладает большим научно-исследовательским потенциалом. Главная трудность этого процесса – поддержание его устойчивости во времени с учетом всех сложностей, перечисленных выше, и того факта, что проведение многих научных исследований на основе полученных данных возможно лишь с накоплением достаточно длинных рядов наблюдений, а это, как правило, несколько лет. Наш опыт взаимодействия в рамках проекта РАНХиГС – «Арована Капитал» [2] показал, что поддержание такого процесса на приемлемом уровне возможно лишь в кооперации и при высокой степени заинтересованности в исследовательском результате со стороны всех участников проекта. ■

Литература
1. Евсеев А. Особенности ценообразования на рынках онлайн-торговли г. Москвы // Экономическое развитие России. 2019. Т. 26. № 10. С. 39–44.
2. Исаков А. и др. Твердые цифры: открытые микроданные о потребительских ценах // Деньги и кредит. 2021. Т. 80. № 1. С. 104–119.
3. Aparicio D., Bertolotto M. I. Forecasting inflation with online prices // International Journal of Forecasting. 2020. Vol. 36. No. 2. Pp. 232–247.
4. Cavallo A. Online and official price indexes: Measuring Argentina's inflation // Journal of Monetary Economics. 2013. Vol. 60. No. 2. Pp. 152–165.
5. Cavallo A. Are online and offline prices similar? Evidence from large multi-channel retailers // American Economic Review. 2017. Vol. 107. No. 1. Pp. 283–303.
6. Cavallo A. Scraped data and sticky prices // Review of Economics and Statistics. 2018. Pp. 105–119.
7. Cavallo A., Rigobon R. The billion prices project: Using online prices for measurement and research // Journal of Economic Perspectives. 2016. Vol. 30. No. 2. Pp. 151–178.
8. Haan J. de, Hendriks R. Online Data, Fixed Effects and the Construction of High-Frequency Price Indexes // Paper Presented at the Economic Measurement Group Workshop. 28–29 November 2013. Sydney, Australia.
9. Krsinich F. The FEWS Index: Fixed Effects with a Window Splice // Journal of Official Statistics (JOS). 2016. Vol. 32. No. 2.
10. Macias P., Stelmasiak D. Food inflation nowcasting with web scraped data // NBP Working Papers 302. 2019.
11. Nygaard R. The use of online prices in the Norwegian Consumer Price Index // Statistics Norway. 2015. Pp. 1–16.

References
1. Evseev A. Features of Price Setting Behavior in the Moscow e-commerce market // Russian Economic Development. 2019. Vol. 26. No. 10. Pp. 39–44.
2. Isakov A. et al. Hard Numbers: Open Consumer Price Database // Russian Journal of Money and Finance. 2021. Vol. 80. No. 1. Pp. 104–119.
3. Aparicio D., Bertolotto M. I. Forecasting inflation with online prices // International Journal of Forecasting. 2020. Vol. 36. No. 2. Pp. 232–247.
4. Cavallo A. Online and official price indexes: Measuring Argentina's inflation // Journal of Monetary Economics. 2013. Vol. 60. No. 2. Pp. 152–165.
5. Cavallo A. Are online and offline prices similar? Evidence from large multi-channel retailers // American Economic Review. 2017. Vol. 107. No. 1. Pp. 283–303.
6. Cavallo A. Scraped data and sticky prices // Review of Economics and Statistics. 2018. Pp. 105–119.
7. Cavallo A., Rigobon R. The billion prices project: Using online prices for measurement and research // Journal of Economic Perspectives. 2016. Vol. 30. No. 2. Pp. 151–178.
8. Haan J. de, Hendriks R. Online Data, Fixed Effects and the Construction of High-Frequency Price Indexes // Paper Presented at the Economic Measurement Group Workshop. 28–29 November 2013. Sydney, Australia.
9. Krsinich F. The FEWS Index: Fixed Effects with a Window Splice // Journal of Official Statistics (JOS). 2016. Vol. 32. No. 2.
10. Macias P., Stelmasiak D. Food inflation nowcasting with web scraped data // NBP Working Papers 302. 2019.
11. Nygaard R. The use of online prices in the Norwegian Consumer Price Index // Statistics Norway. 2015. Pp. 1–16.

Technical and Methodological Challenges of Collecting Price Data from Online Retailers

Alexey S. Evseev – Researcher of the Russian Presidential Academy of National Economy and Public Administration (Moscow, Russia). E-mail: evseev-als@ranepa.ru
Rodion R. Latypov – Head of Macroeconomic Research, Economist of the JSC Arowana Capital (Moscow, Russia). E-mail: Rodion.Latypov@mncap.ru

Egor A. Postolit – Analyst of the JSC Arowana Capital (Moscow, Russia). E-mail: postolitegor@gmail.ru
Elena S. Sinelnikova-Muryleva – Senior Researcher of the Russian Presidential Academy of National Economy and Public Administration, Candidate of Economic Sciences (Moscow, Russia). E-mail: e.sinelnikova@ranepa.ru

Price data from online retailers is a valuable source for economics. The use of these data makes it possible to refine inflation forecasts and anticipate future trends in the moment, refine estimates of price rigidity and the conclusions of theoretical pricing models, and test the law of one price. However, there are major difficulties in the data collection process that are not obvious and can threaten both the quality of the data collected and the sustainability of the collection process over time.

The article, for the first time in the literature, discusses in detail the technical and methodological problems that impede the continuous collection of data on the network and presents our experience in solving these problems. The pros and cons of solutions to emerging problems are discussed.

The article was written on the basis of the RANEPA state assignment research programme for 2022.

Key words: prices of online retailers, web-scrapping, inflation, alternative data, big data.
JEL-codes: C81, C82, E31, C55.