Московский государственный университет имени М. В. Ломоносова

Факультет журналистики

Кафедра новых медиа и теории коммуникации

BIG DATA КАК ИНСТРУМЕНТ МАРКЕТИНГА: ОТЕЧЕСТВЕННЫЙ ОПЫТ

Исследовательская работа

студентки 408 группы

дневного отделения

Долголаптевой К.О.

Москва 2018

Содержание

Введение

Из-за постоянного роста объема данных технологии сбора и анализа Big Data применяются в самых разных сферах: финансовой, здравоохранения, сайтостроения, медиа и других. Что касается сферы маркетинга, то будет большим заблуждением считать, что Big Data служит лишь своеобразным сырьем для RTB (рекламной технологии, которая позволяет организовать аукцион между продавцами и покупателями рекламы в реальном времени, максимально точно отбирая контент для каждого целевого посетителя[1]). Анализ больших данных позволяют сегментировать клиентов в Email-маркетинге (увеличение ROMI до 760%[2]), повышать лояльность клиентов и снижать коэффициент оттока, прогнозировать продажи, а также разрабатывать коммерчески успешные продукты или услуги.

Важнейшим толчком для развития больших данных послужила цифровизация, начало которой было положено еще в 1980-е. Переломным стал 2007 год: развивающиеся технологии упростили и удешевили процесс анализа данных[3]. Брайан Уилкоу в своей статье «The Best Things Come In Threes»[4] предлагает разделить историю «эры аналитики» на 3 составляющие: первый этап подразумевает хранение информации для составления отчетов, для этого нужно бесчисленное количество специалистов и миллионы долларов; второй этап характеризуется появлением облачных технологий, демократизации технологических инноваций; третий этап — анализ больших данных в реальном времени с помощью машинного обучения. Сегодня интерес к большим данным проявляют только крупные компании, поскольку недостаточно купить массив данных, его нужно правильно обрабатывать с помощью искусственного интеллекта, а также хранить на больших по объему серверах.

Четкого определения термина Big Data не существует: «Большие данные — это способность обрабатывать огромные массивы информации, мгновенно их анализировать и получать порой совершенно неожиданные выводы […] Это масса новых задач, касающихся общественной безопасности, глобальных экономических моделей, неприкосновенности частной жизни, устоявшихся моральных правил, правовых отношений человека, бизнеса и государства»[5], — такое определение дает книга Виктора Майера-Штенбергера и Кеннета Кукьера. «Большие данные (безотносительно их качества и происхождения) — неструктурированные объемы информации, которые выходят за рамки старого представления о данных. Не существует четкого определения их количества или веса, но их объем как минимум

[1] Словарь маркетинговых терминов: https://bit.ly/2QEwnGJ (дата обращения от 29.11.2018)
[2] Big Data в маркетинге: https://bit.ly/2zMgqF2 (дата обращения от 29.11.2018)
[3] The Best Things Come In Threes: https://tcrn.ch/2ISsgUM (дата обращения от 29.11.2018)
[4] Там же.
[5] Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. М., 2014. С. 4–8

больше объема одного жесткого диска»[6], — заявляет Data Scientist компании uKit Роман Штейнберг. «Термин "большие данные" […] определяет не только размер наборов данных, который превосходит возможности обычных баз данных (БД) по занесению, хранению, управлению и анализу, но и неструктурированную информацию, перед обработкой и анализом которой бессильны традиционные алгоритмы»[7], — такое определение Big Data дано в научной работе О. Ю. Денисовой и Э. А. Мухутдинова. «Big Data — это не какой-либо определенный массив данных, а совокупность методов их обработки. Определяющей характеристикой для больших данных является не только их объем, но также и другие категории, характеризующие трудоемкие процессы обработки и анализа данных»[8], — пишут журналисты издания ForkLog. «Само по себе понятие Big Data в разных источниках определяется по-разному, однако все определения согласованы с концепцией трех V: *volume*, *velocity*, *variety* (объем, скорость, разнообразие). В отдельных источниках к ним добавляется и четвертая «*V*» − *veracity* (достоверность). В широком смысле Big data – это разнородные неструктурированные данные крайне большого объема, увеличение которого происходит ежедневно с большой скоростью»[9], — такую характеристику большим данным дает С.А. Вартанов.

Новизна нашего исследования связана, во-первых, с тем, что анализу рынка больших данных как инструмента маркетинга в России посвящено относительно небольшое количество работ, а во-вторых, с тем, что отечественными компаниями достоинства Big Data еще не оценены в полной мере из-за отсутствия знаний, компетентных сотрудников, способных грамотно работать с ними.

Актуальность нашей работы связана, во-первых, с перспективностью использования больших данных для оптимизации работы многих корпораций (в том числе медиакомпаний) и, как следствие, улучшения качества обслуживания клиентов (на примере проекта uKit AI мы рассмотрим, как бизнес сможет адаптировать свои сайты индивидуально под каждого посетителя), а во-вторых, с повышенным интересом мировой общественности к возможностям, которые открывает анализ больших данных.

Цель работы — проанализировать опыт компании uKit и рассмотреть, как развивался отечественный рынок Big Data, выяснить, для чего используются большие данные и как маркетинговые решения на их основе меняют деятельность компании, изучить методы

---

[6] См. приложение.

[7] Денисова, О. Ю., Мухутдинов Э. А. Большие данные — это не только размер данных // Вестник технологического университета. 2015. Т.18, №4. С. 1

[8] Big Data и блокчейн — прорыв в области анализа данных: http://bit.ly/2s1hUan (дата обращения от 29.11.2018)

[9] Вартанов С.А. Большие данные в онлайн-СМИ: подходы и стратегии использования // Медиаскоп. 2017. Вып. 4: http://bit.ly/2IDupDo (дата обращения от 29.11.2018)

работы с большими данными, поразмышлять над проблемами безопасности хранения данных, тенденциями и перспективами.

Общая схема работы с Big Data такова: обработка большого объема данных, накопление и анализ, обучение машинной модели. Однако каждая стадия этого процесса у каждой компании протекает по-разному, здесь важна как специфика предприятия, так и параметры данных, с которыми оно работает.

Основные проблемы использования больших данных связаны с обеспечением безопасности, целостности данных, а также этичности их получения и применения. В связи с этим **методология** нашего исследования сводится к анализу опыта компании uKit, а также моделированию ситуации использования больших данных в российских медиа. Мы взяли интервью с представителями описываемых компаний; с расшифровками можно ознакомиться в приложении.

СПЕЦИФИКА РЫНКА BIG DATA В РОССИИ

В сравнении с Западом российский рынок больших данных относительно мал[10]: это связано с тем, что анализ больших массивов данных могут позволить себе только зрелые компании, которые заинтересованы в поиске новых решений для повышения своей эффективности. Как правило, с Big Data работают банки, мобильные операторы, маркетинговые компании: «Решения по анализу больших данных внедрены в Сбербанке, Газпромбанке, ВТБ24, "Альфа-Банке", ФК "Открытие", "Райффайзенбанке", […] а также у главных телеком-операторов. Из крупных ритейлеров этими технологиями пользуются X5 Retail Group, "Глория Джинс", "Юлмарт", сеть гипермаркетов "Лента", "М.Видео", Wikimart, Ozon, "Азбука вкуса", из нефтяных компаний — "Транснефть", "Роснефть" и "Сургутнефтегаз"»[11].

Основная проблема работы с Big Data заключается не в их извлечении (если компания не обладает собственными данными, она может их купить или воспользоваться открытыми источниками), а в анализе: «Пять-семь лет назад многие компании начали использовать большие данные, но сейчас, поработав с ними более плотно, поняли, что результаты не такие высокие, как хотелось бы. Использование больших данных не всегда оправдано и не всегда вообще нужно идти в эту сферу. В большинстве случаев непонятно, как настраивать модель. Данные есть, но что с ними делать — большой вопрос»[12], — заявляет Data Scientist компании uKit Роман Штейнберг. Есть и другая точка зрения: «Граница, разделяющая в цифровом мире проигравших и победителей, проходит на уровне данных […] Чтобы оставаться конкурентоспособными, компании должны научиться мыслить глобально и быстро масштабироваться — технологии больших данных, интернета вещей и машинного обучения являются лучшим ресурсом для этого»[13], — считает генеральный директор Hitachi Vantara в России и СНГ Юрий Скачков.

Работа с Big Data стоит недешево, однако эти затраты будут бессмысленны, если компания не знает, что она хочет получить в результате применения больших данных. Кроме того, для их анализа нужен специалист, который не только настроит машинную модель, но и будет регулярно перенастраивать ее, чтобы достичь точного результата с минимальной погрешностью. Многие компании не могут позволить себе такие в своем роде неоправданные финансовые затраты.

Государственный сектор экономики также работает с большими данными: «Среди госструктур обработку Big Data внедрили Федеральная налоговая служба, аналитический

[10] Как устроен рынок Big Data в России: http://bit.ly/2x7lXXT (дата обращения от 29.11.2018)
[11] Там же.
[12] См. приложение.
[13] Традиционные методы анализа данных больше не работают: http://bit.ly/2Lse6ao (дата обращения от 29.11.2018)

центр правительства России, Пенсионный фонд, правительство Москвы, Фонд обязательного медицинского страхования, Федеральная служба безопасности, Следственный комитет и Служба внешней разведки»[14]. Однако наиболее перспективные области для применения больших данных — медицина, медиасфера — как правило, используют традиционные способы обработки данных.

Сегодня в России технология наиболее эффективно используется в сфере маркетинга: «Благодаря развитию интернета и распространению всевозможных коммуникационных устройств поведенческие данные (такие как число звонков, покупательские привычки и покупки) становятся доступными в режиме реального времени»[15]. Большие данные также используют в своей работе медиамаркетологи: СМИ не только продают аудитории контент, но и зарабатывают на предоставлении рекламных площадей, им необходимо знать свою целевую аудиторию как можно точнее для эффективной коммуникации как с самой аудиторией, так и с рекламодателем: «Возможности измерять и оценивать аудиторию в реальном времени и по множеству параметров также выгодно отличает новые медиа от традиционных»[16]. Кроме всего прочего, сетевые медиа отдают предпочтение анализу редакционной метрики (доскроллы и время вовлечения, коэффициент виральности)[17]. Можно ли назвать подобный анализ работой с большими данными? Если это данные небольшого сетевого издания, и их можно проанализировать с помощью баз данных, электронных таблиц, регрессионным методом — то нет. Если перед нами новостной агрегатор вроде «Яндекс.Новости» или «Новости Mail.Ru» — однозначно да, потому что для анализа огромного массива неструктурированных данных, с которыми вынуждены работать эти площадки, не подходят традиционные инструменты анализа.

С телевидением вопрос более сложный. Основными потребителями телевизионного контента (среднесуточная доля) является аудитория 55 лет и старше[18]. Однако покупательная способность начала смещаться в сторону так называемых цифровых аборигенов, которые привыкли получать интересующий их контент в удобное время. Из-за стремления телевизионных компаний привлечь молодую аудиторию, на Западе возник феномен адресного телевидения[19]. Однако Россия пока что не может позволить себе сегментировать аудиторию и более эффективно таргетировать рекламу посредством анализа больших данных,

---

[14] Как устроен рынок Big Data в России: http://bit.ly/2x7lXXT (дата обращения от 29.11.2018)
[15] Big Data и блокчейн — прорыв в области анализа данных: http://bit.ly/2s1hUan (дата обращения от 29.11.2018)
[16] Модели монетизации в медиа: как и на чем заработать изданию и автору? http://bit.ly/2s9DfOb (дата обращения от 29.11.2018)
[17] Редакционные метрики Mail.Ru: как мы оцениваем работу редакции: http://bit.ly/2IK2fmn (дата обращения от 29.11.2018)
[18] Исследования TV Index, Mediascope: http://bit.ly/2GKexZS (дата обращения от 29.11.2018)
[19] What marketers need to know about addressable TV and OLV: https://selnd.com/2x3OWM2 (дата обращения от 29.11.2018)

потому что переход на цифровое телевидение еще не осуществлен до конца: «Так, около 70% телевизоров россиян способны поддерживать цифровой прием, хотя охват составляет в среднем 98%. Адресное планирование ТВ-рекламы в РФ — скорее, вопрос отдаленного будущего, чем настоящего»[20].

Применение Big Data в медиасфере позволяет лучше узнавать постоянного читателя и находить нового, оценивать степень привлекательности СМИ для аудитории и понимать, соответствует ли контент ее потребностям, а также создавать потенциально интересные читателю проекты. Сегментация аудитории с помощью анализа больших данных дает возможность максимально персонализировать медиа, что положительно оценит как читатель, так и рекламодатель.

Издание Rusbase разделяет участников отечественного рынка больших данных на следующие категории[21]:

1. Поставщики Big Data — хранят обрабатывают данные (Sap, Oracle, IBM, EMC, Microsoft и др.);
2. Датамайнеры — разрабатывают инструменты для анализа данных (Yandex Data Factory, «Алгомост», Glowbyte Consulting, CleverData и др.);
3. Системные интеграторы — составляют системы анализа данных клиента («Форс», «Крок» и др.);
4. Потребители инструментов работы с большими данными (мобильные операторы, банки и др.);
5. Разработчики готовых сервисов на базе больших данных — такие решения позволяют малому и среднему бизнесу обрабатывать Big Data с минимальными затратами и усилиями.

В рамках нашей работы мы рассмотрели опыт компании uKit, поскольку ее решения на основе Big Data могут перевернуть сферу сайтостроения, усовершенствовать маркетинговые механизмы, которые позволят бороться за внимание пользователей в автоматическом режиме.

Разработчики uKit научили искусственный интеллект делать «умные» сайты с персонализированным дизайном (пользователь видит такое оформление страницы, которое соответствует его предпочтениям и ожиданиям): «Все люди разные: кто-то любит heavy metal и ему будет приятно увидеть черное оформление сайта, кто-то же любит маленьких пудельков и, наверное, розовый сайт ему будет интереснее. Это возможно благодаря оцениванию

[20] Как медиакорпорации используют Big Data: http://bit.ly/2IFfush (дата обращения от 29.11.2018)
[21] Как устроен рынок Big Data в России: http://bit.ly/2x7lXXT (дата обращения от 29.11.2018)

цифрового следа, анализа cookies»[22], — объяснил концепцию проекта ведущий Data Scientist компании uKit Роман Штейнберг.

Потенциально сайты СМИ могут использовать такого рода наработки не только для потворствования вкусовым пристрастиям аудитории в оформлении контента, но и для выдачи релевантных интересам пользователя материалов. Это повысит лояльность аудитории, увеличит количество времени, которое читатель проводит на сайте издания, улучшит эффективность работы с рекламодателем.

Однако есть и другая сторона медали. Известный социолог, технопессимист Евгений Морозов в одной из своих работ говорит о том, что уже сегодня персонализация касается не только дизайна сайтов, но и вербальной части — текстов. Алгоритмы учитывают информационные потребности и образованность пользователя, чтоб выдавать ему адаптированные с точки зрения тематики, лексики, композиционных особенностей материалы: «"Машинная журналистика" мгновенно предлагает такие публикации, которые подстраиваются под каждого отдельного читателя в соответствии с его интересами и интеллектуальным уровнем. И это основание для серьезного беспокойства. Рекламодатели и издатели обожают точность в маркетинговой "стрельбе" по потребителю, с помощью которой можно подольше удерживать его внимание. Но чем это оборачивается для общества, не вполне понятно. В конце концов, есть угроза того, что люди, поглощая одни информационные объедки и не догадываясь об интеллектуальном разнообразии, попадут в ловушку однородных новостей»[23].

[22] См. приложение.

[23] Морозов Е. Робот отобрал у меня «Пулитцера» // Техноненависть: как интернет отучил нас думать / Пер. с англ. — В. Гончарук [сб. статей]. М, 2014. С.31

ИННОВАЦИИ В ОБЛАСТИ САЙТОСТРОЕНИЯ: ОПЫТ UKIT

Конструктор uKit появился в 2015 году как новая версия uCoz — инструмента для создания сайтов: «Мы помогаем пользователям, не имеющим навыков программирования, делать сайты. Главный принцип: простота и удобство. […] Преимущество uKit [состоит] в том, что его сайты сделаны на современном стеке технологий»[24], — поделился Data Scientist компании uKit Роман Штейнберг.

В рамках данной работы мы рассмотрим не собственно конструктор сайтов, а новые проекты uKit Group, работающие с большими данными: WebScore AI[25] и uKit AI[26].

Компания начала использовать большие данные относительно недавно — со стартом проекта WebScore AI. UKit Group не покупает и продает данные («покупать их не пришлось, потому что датасетов, связанных с сайтами, очень мало; да и те, которые есть в сети, не подходят под наши цели»[27]): за основу берутся данные владельцев сайтов на платформе uKit, данные асессоров, а также данные из открытых источников — Интернета. Разработчики также утверждают, что действующее российское законодательство не осложняет работу с Big Data.

WebScore AI: искусственный интеллект оценивает привлекательность сайтов

WebScore AI является составляющей проекта uKit AI — системы редизайна сайтов с помощью искусственного интеллекта. WebScore AI оценивает привлекательность страницы по 10-балльной шкале. При выставлении баллов за оформление (цветовая гамма, шрифты) искусственный интеллект моделирует впечатление простого интернет-пользователя. При этом система учитывает формальные параметры, вроде адаптивности под разные экраны (мобильная и десктопная версии) и структурированности информации.

Компания запустила WebScore AI в конце 2017 года. На начальном этапе разработчики собрали данные при помощи асессоров — фокус-группы пользователей (непрофессионалов в области сайтостроения), которая оценивала привлекательность подборки сайтов тремя способами: попарное сравнение (сопоставить скриншоты двух сайтов — какой из двух лучше), поочередное (выставление оценки по 10-балльной системе), списком. «Мы показали асессорам около 1 500 сайтов — это, конечно, не так много, но само понятие больших данных очень спорно: например, ложка — это много или мало? Если ложка рыбьего жира, то, наверное, это много. А если ложка сладкого сиропа, то можно еще взять. Полутора

[24] См. приложение.
[25] Режим доступа: https://webscore.ai/
[26] Режим доступа: https://ukit.ai/
[27] См. приложение.

тысяч сайтов было достаточно для того, чтоб принять решение о том, какой из способов разметки выбирать»[28], — прокомментировал Роман Штейнберг. Для дальнейшего обучения системы разработчики увеличили количество сайтов до 12 000. Нужно было получить градацию сайтов от плохих до очень хороших, чтобы система была готова к тому, с чем она столкнется в Сети. WebScore AI на удивление точен: «Модель оценивает сайты в каком-то смысле лучше, чем наши асессоры. Мы подсчитали для каждого асессора, насколько он отклоняется от средней оценки, и для WebScore AI. И модель ставит баллы более точно: она ошибается в среднем на 0.5 балла, а пользователи — на 1, 1.2 балла»[29].

Единственная проблема WebScore AI — инструмент не анализирует анимацию, динамические элементы. Но, по словам разработчиков, она решаема в ближайшем будущем. Чтобы система могла оценить, например, видео, его нужно описать набором признаков (сегодня сайты разбиваются на 325 признаков, видео значительно расширит этот набор).

WebScore AI будет одной из составляющих uKit AI, но сегодня его основная роль состоит в демонстрации компетенций команды и в привлечении новых инвесторов: «Во-первых, это полезный инструмент, который показывает клиентам, что мы умеем работать с искусственным интеллектом, в области машинного обучения и анализа данных. […] Во-вторых, мы хотим использовать WebScore AI для дальнейшего обучения uKit AI тому, что такое хорошо и что такое плохо»[30].

Важно отметить, что WebScore AI обучается только под руководством разработчиков, оценки пользователей учитываются лишь после модерации: «Известен случай в компании Facebook, когда разработчики отпустили чат-бота, чтобы его учили рядовые пользователи, и этот чат-бот приобрел черты грубияна-фашиста с расистскими наклонностями. Такие вещи лучше держать под контролем»[31].

uKit AI: персонализированный и генеративный дизайн для любого сайта в Сети

uKit AI — система редизайна сайтов, основанная на генеративных алгоритмах и нейросетях. Это молодой амбициозный проект, который завершил финальный этап ICO в апреле 2018 года. Сегодня система работает в beta-режиме. Релиз uKit AI 1.0 (улучшение сайтов с помощью искусственного интеллекта) состоится в июле 2018. Запуск версии uKit AI 2.0, способной осуществлять генеративный дизайн, дополненной системой вознаграждений токенами, пройдет в сентябре 2018 – январе 2019 года. Параллельно с этим в мае 2018

[28] См. приложение.
[29] Там же.
[30] См. приложение
[31] Там же.

года компания занялась разработкой проекта uData — блокчейн-хранилища больших данных; релиз намечен на август 2018 года[32].

Прототипом uKit AI послужил западный проект The Grid. Павел Кудинов и Александр Пезиков усовершенствовали идею при разработке отечественного аналога. Сегодня искусственный интеллект в области сайтостроения — рынок, на котором действует относительно небольшое количество игроков. Можно сказать, что компания нашла свой так называемый голубой океан.

Проект рассчитан в большей степени на сегмент B2B, его целевая аудитория — владельцы малого и среднего бизнеса: теперь те услуги, которые были доступны только крупным предприятиям (например, нанимать Data Scientists, чтобы оптимизировать собственные сайты и повышать конверсию с помощью искусственного интеллекта), заменяет технология uKit AI: «Малому бизнесу постоянно поддерживать сайты в современном состоянии дорого, да и нет времени заниматься этим. Мы как раз предоставляем услуги для этого круга клиентов. Они очень легко могут перенести сайт на нашу платформу, причем мы предлагаем эту услугу бесплатно. Клиент платит только за то, что сайт "живет" на нашей платформе. Наш искусственный интеллект будет поддерживать сайт в актуальном состоянии»[33].

UKit AI работает следующим образом: пользователь дает ссылку на свой сайт (независимо от платформы размещения), uKit AI анализирует структуру и распознает отдельные элементы, затем система формирует контентное дерево и перестраивает его таким образом, чтобы важная информация оказалась ближе к корню: «Если мы можем выделить на сайте объекты, потенциально интересные среднестатистическому пользователю, то эти компоненты нужно постараться вытащить наверх, показать в первом экране. Например, контакты не должны быть написаны мелким шрифтом в уголке»[34]. Затем вышеописанная нами система WebScore AI оценивает привлекательность полученного сайта. Пользователь может вручную отредактировать составляющие любой страницы с помощью конструктора uKit.

Отдельного внимания заслуживает экономика проекта. Любую услугу можно оплатить не только рублями, долларами или евро, но и токенами — внутренней криптовалютой. Ее можно «намайнить», поделившись с компанией обезличенными данными.

Когда система перестраивает сайт, она учитывает такой параметр, как фокус внимания посетителей. Для этого uKit AI использовали разработки проекта Visimportance, главной составляющей которого также является искусственный интеллект, эмулирующий вни-

[32] Искусственный интеллект для сайтов // uKit ICO: http://bit.ly/2LyoOfq (дата обращения от 29.11.2018)
[33] См. приложение.
[34] Там же.

мание человека и составляющей специальные карты: «Карты внимания — это прогноз искусственным интеллектом того, куда посмотрит пользователь первым делом, как зайдет на сайт»[35].

Еще одна положительная новость для тех, кто стремится подчеркнуть индивидуальность своего сайта и максимально уйти от шаблонности: разработчики подчеркивают, что уникальность не пострадает после редизайна. Они приводят аналогии с рекламной кампанией Nutella в 2017 — дизайн 7 млн. банок шоколадной пасты генерировался нейросетью — покупателю нравилось приобретать Nutella с уникальной этикеткой. Такой шаг не мог не повысить продажи.

Что касается персонализированного дизайна (для каждого отдельного пользователя сайт выглядит по-разному) — это ближайшее направление разработок компании. Задумка состоит в том, что искусственный интеллект в реальном времени будет подстраивать вебстраницы под посетителей, анализируя цифровой след каждого пользователя. Чтобы это стало возможным, необходима совокупность трех составляющих: база данных об интересах интернет-пользователей (для этого компания uKit планирует закупать данные у рекламных платформ, а также вознаграждать собственных клиентов за то, что они делятся данными), система персонализации сайтов на основе искусственного интеллекта, система постоянного анализа данных для улучшения конверсии такого сайта. Искусственный интеллект будет обучаться в реальном времени, отслеживая результаты персонализации и запоминая реакции пользователей на разные версии сайта.

[35] См. приложение.

Заключение

Большие данные имеют большой потенциал. Однако сегодня не каждая компания может позволить себе обрабатывать Big Data — существует ряд проблем, которые не так просто решить.

Во-первых, работа с большими данными требует существенных финансовых расходов для сбора датасета, настройки машинной модели, а также покупки мощного аппаратного обеспечения, оплаты труда квалифицированного специалиста. От последнего зависит то, насколько полно окупятся затраты: большие данные являются своеобразным топливом для искусственного интеллекта — чем лучше качество датасета, чем правильнее настроена модель, тем точнее результат. Основной сложностью (равно как и главным преимуществом) работы с большими массивами данных является их неупорядоченность[36], из-за которой специалисту важно правильно на начальном этапе определить цели предстоящей работы, а также определить, какие данные нужны для извлечения и анализа. Перед Data Scientists стоит сложнейшая задача понять, как применить методы точных наук к предмету изучения общественных наук[37]. Профессия аналитика Big Data востребована, однако сегодня в вузах нет специальных направлений подготовки. Крупные компании вынуждены открывать собственные программы обучения. Кроме того, малый и средний бизнес не готов идти на риски, учитывая тот факт, что опыт внедрения Big Data в России не так велик и готовых кейсов, на которые можно опереться при разработке собственных проектов[38], практически нет.

Во-вторых, этический вопрос, а также проблема безопасности хранения данных становятся все более острыми и заслуживающими отдельного обсуждения: «Как защититься от машин, которые собирают данные? Недавний скандал с Facebook и Mail.ru лишнее тому подтверждение. Больше половины населения не понимает, о чем речь, другая же половина начинает задумываться о том, как свои данные обезопасить. Этические нормы в работе с Big Data очень важны»[39]. Безопасность хранения данных может обеспечить технология распределенного реестра: «Одна из главных проблем больших данных, которую призван решить блокчейн, лежит в сфере информационной безопасности. Технология […] может гарантировать целостность и достоверность данных, а благодаря отсутствию единой точки отказа, блокчейн делает стабильной работу информационных систем»[40]. К тому же,

[36] Чехарин Е.Е. Большие данные: большие проблемы // Перспективы Науки и Образования. 2016. С. 21
[37] Человеческая физика: как точные науки изучали общество до эпохи Big Data: http://bit.ly/2ILA4r4 (дата обращения от 29.11.2018)
[38] Большие данные (Big Data) в России: http://bit.ly/2L0N2hi (дата обращения от 29.11.2018)
[39] См. приложение.
[40] Big Data и блокчейн — прорыв в области анализа данных: http://bit.ly/2s1hUan (дата обращения от 29.11.2018)

интеграция блокчейна в экономическую сферу поможет не только упорядочить данные о многочисленных финансовых операциях, повысить эффективность борьбы с мошенничеством, но и «сократить время обработки транзакций от нескольких дней до нескольких минут»[41].

В-третьих, на отечественном рынке Big Data мы можем отметить негибкость многих структур: например, потенциал обработки больших данных до сих пор недооценивается российскими медиаменеджерами. Это хорошо заметно, если мы сопоставим любую социальную сеть, размещающую медиаконтент, и, собственно, СМИ. Первые борются за внимание аудитории, внедряя новые алгоритмы формирования новостной ленты, вторые продолжают неструктурированно и без учета пользовательских предпочтений размещать материалы на неадаптированных под мобильные экраны сайтах. В условиях, которые диктует современность, в эпоху, когда главным ресурсом становится не информация, а внимание, медиа должны стать более гибкими, научиться предугадывать поведение своей аудитории посредством анализа больших данных и модернизации своих веб-сайтов, используя технологии персонализированного дизайна.

[41] Там же.

Библиография

1. Вартанов С.А. Большие данные в онлайн-СМИ: подходы и стратегии использования // Медиаскоп. 2017. Вып. 4
2. Денисова, О. Ю., Мухутдинов Э. А. Большие данные — это не только размер данных // Вестник технологического университета. 2015. Т.18, №4
3. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. М., 2014
4. Морозов Е. Техноненависть: как интернет отучил нас думать / Пер. с англ. — В. Гончарук [сб. статей]. М, 2014
5. Чехарин Е.Е. Большие данные: большие проблемы // Перспективы Науки и Образования. 2016
6. Big Data в маркетинге: https://bit.ly/2zMgqF2 (дата обращения от 29.11.2018)
7. Большие данные (Big Data) в России: http://bit.ly/2L0N2hi (дата обращения от 29.11.2018)
8. Big Data и блокчейн — прорыв в области анализа данных: http://bit.ly/2s1hUan (дата обращения от 29.11.2018)
9. Искусственный интеллект для сайтов // uKit ICO: http://bit.ly/2LyoOfq (дата обращения от 29.11.2018)
10. Исследования TV Index, Mediascope: http://bit.ly/2GKexZS (дата обращения от 29.11.2018)
11. Как медиакорпорации используют Big Data: http://bit.ly/2IFfush (дата обращения от 29.11.2018)
12. Как устроен рынок Big Data в России: http://bit.ly/2x7lXXT (дата обращения от 29.11.2018)
13. Кредит временно недоступен: http://bit.ly/2sjLMhL (дата обращения от 29.11.2018)
14. Модели монетизации в медиа: как и на чем заработать изданию и автору? http://bit.ly/2s9DfOb (дата обращения от 29.11.2018)
15. Редакционные метрики Mail.Ru: как мы оцениваем работу редакции: http://bit.ly/2IK2fmn (дата обращения от 29.11.2018)
16. Словарь маркетинговых терминов: https://bit.ly/2QEwnGJ (дата обращения от 29.11.2018)
17. Традиционные методы анализа данных больше не работают: http://bit.ly/2Lse6ao (дата обращения от 29.11.2018)
18. Человеческая физика: как точные науки изучали общество до эпохи Big Data: http://bit.ly/2ILA4r4 (дата обращения от 29.11.2018)
19. Google Trends. Big Data: http://bit.ly/2k5bjY6 (дата обращения от 29.11.2018)
20. The Best Things Come In Threes: https://tcrn.ch/2ISsgUM (дата обращения от 29.11.2018)
21. What marketers need to know about addressable TV and OLV: https://selnd.com/2x3OWM2 (дата обращения от 29.11.2018)

Приложение

Расшифровка интервью с ведущим Data Scientist компании uKit Романом Штейнбергом

— Конструктор uKit появился на рынке в 2015 году. Когда вы начали использовать Big Data?

— Конструктор uKit появился как новая версия инструмента для создания сайтов. 13 лет назад мы запустили uCoz, а uKit действительно появился в 2015 году как новая версия конструкторов сайтов. Мы помогаем пользователям, не имеющим навыков программирования, делать сайты. Главный принцип: простота и удобство. Backend мы разрабатываем с помощью инструмента Node.js, который позволяет сайту быть современным. 15 лет назад сайты делали на Perl или PHP, а сегодня появились более новые технологии. То есть главное преимущество uKit в том, что его сайты сделаны на современном стеке технологий. А что касается привлекательности сайтов для пользователей, здесь основной акцент сделан на работу дизайнеров. Последние отвечают за то, чтоб сайты, сконструированные пользователем, выглядели привлекательно.

С большими данными работает проект uKit AI. Он стартовал не сегодня, у него было три волны развития. В 2016 году мы начали делать лабораторию uKit AI, благодаря которой смогли визуализировать процесс конвертирования сайта. Это внутренний продукт, который, скорее всего, не будет опубликован — он помогает нам делать uKit AI и наблюдать результаты в более удобной форме. Это не нейросеть, а инструмент визуализации: он показывает оригинальный сайт и результирующий.

Пять-семь лет назад многие компании начали использовать большие данные, но сейчас, поработав с ними более плотно, поняли, что результаты не такие высокие, как хотелось бы. Использование больших данных не всегда оправдано и не всегда вообще нужно идти в эту сферу. В большинстве случаев непонятно, как настраивать модель. Данные есть, но что с ними делать — большой вопрос

Сегодня uKit AI работает в beta-режиме и представляет сайты пользователя в том виде, который подходит для нашей платформы. Это нужно для того, чтобы унифицировать страницы. Во-первых, дизайнеры и программисты делают сайты по-разному, во-вторых, встречаются сайты с ошибками. И задача нашего конвертера — привести сайт пользователя в тот вид, с которым мы можем работать.

Мы только начинаем использовать большие данные. Пока покупать их не пришлось, потому что датасетов, связанных с сайтами, очень мало. Да и те, которые есть в Сети, не подходят под наши цели. Я так понимаю, что все те люди, которые делают искусственный интеллект в области сайтостроения, создают свои подборки.

У нас есть данные сайтов наших пользователей, но они нерепрезентативны — это данные только тех страниц, которые расположены на uKit.com. Наша задача — собрать такую подборку сайтов, которые будут представлять Интернет наиболее полно.

Чтобы настроить модель машинного обучения, нужны данные. Благодаря ним существует проект WebScore AI. Он разработан следующим образом: мы брали подборку сайтов (около 15 000), делали скриншоты этих сайтов и показывали нашим асессорам. Нам нужно было получить оценку для каждого сайта. Чем больше оценок, тем лучше — это дает возможность представить отношение большей части пользователей Интернета. Полученные данные мы, естественно, усредняли: сайт имеет оценку 5, если пользователи поставили разные оценки, но среднее арифметическое — пятерка. Оценки собирались по десятибалльной шкале.

— Обучается ли нейросеть самостоятельно, когда посетители WebScore AI ставят оценки сайтам?

— На сегодняшний день мы не хотим отпускать ее в свободное плаванье. Известен случай в компании Facebook, когда разработчики отпустили чат-бота, чтобы его учили рядовые пользователи, и этот чат-бот приобрел черты грубияна-фашиста с расистскими наклонностями. Такие вещи лучше держать под контролем. WebScore AI обучается только тогда, когда мы его хотим обучить: получаем новую модель, оцениваем, насколько она лучше, чем предыдущая, и только потом выкладываем ее в продакшн. Это стандартный способ: те же прогнозы погоды работают по такой же схеме, они раз в неделю обучают свои модели и сравнивают их с предыдущими. В конечном итоге предсказание прогноза погоды осуществляется лучшей моделью.

— Но вы наверняка рассматриваете пользовательские оценки для калибровки нейросети.

— Да, группа экспертов рассматривает каждую пользовательскую оценку, после чего решает, стоит ли ее вносить в систему или нет. Бывает, что пользователи не всегда честны: если 10 человек оценили сайт на 9 баллов, а один оценил на 2 балла, то, наверное, этот пользователь негодует по какому-то поводу и решил занизить оценку. Такие случаи мы называем аномалиями. Поиск аномалий — это еще одна из задач машинного обучения.

— Почему для оценки сайта система делает скриншот? Не проще ли анализировать код сайта?

— Мы делаем и то, и другое, оцениваем две составляющие: визуальную (нейронная сеть Resnet50 оценивает скриншот) и структурную (html-код). Полученная информация проходит проверку на соответствие 325-и признакам, после чего можно сказать, насколько хорош сайт.

— Почему WebScore AI не умеет оценивать анимацию и планируете ли вы решать эту проблему?

— Каждый объект, который используется для машинного обучения, должен быть описан каким-то набором признаков (в нашем случае сайты описываются 325 признаками). Как только мы попытаемся расширить количество задач и описать видео, нам нужно будет понять, как описать видео с помощью чисел. Как только мы поймем, как это делать, то к 325 мы добавим еще набор признаков и сможем анализировать сайты более полно.

— И все же какую цель вы преследовали, когда запускали WebScore AI?

— Во-первых, это полезный инструмент, который показывает клиентам, что мы умеем работать с искусственным интеллектом, в области машинного обучения и анализа данных. У нашей компании прошел заключительный этап ICO[42] (*прим*.: проект uKit AI), и наиболее частыми вопросами у потенциальных инвесторов были «что у вас за команда?», «что вы умеете?» — так что чем больше мы можем продемонстрировать разработок в этой сфере, тем лучше.

Во-вторых, мы хотим использовать WebScore AI для дальнейшего обучения uKit AI тому, что такое хорошо и что такое плохо. Когда искусственный интеллект будет редизайнить сайты, его составляющей, позволяющей оценить результат, будет WebScore AI.

— В анонсе uKit AI сказано, что при редизайне сайта система перемещает важную информацию к корню контентного дерева. А как нейросеть понимает, какая информация важная?

— Есть проекты, которые по изображению определяют наиболее важные, с точки зрения пользователя, места — куда в первую очередь он посмотрит. Эти проекты тоже сделаны на базе искусственного интеллекта. Один из них — Visimportance (Visual Importance). Этот проект работает не только с сайтами, но и с документами — такой искусственный интеллект позволяет эмулировать внимание человека.

[42] Режим доступа: https://ico.ukit.com/ru

— Это связано с тепловыми картами сайтов?

— Тепловая карта не совсем связана с вниманием или искусственным интеллектом, тепловая карта — это способ визуализации. Карты внимания — это прогноз искусственным интеллектом того, куда посмотрит пользователь первым делом, как зайдет на сайт. Но опять же, он использует ту информацию, которую задали после того, как провели опыт с асессорами: пользователям надевали приборы наблюдения за взглядом (eye-tracker) и показывали изображения. Мы используем возможности Visimportance для своего проекта.

— Давайте вернемся к контентному дереву.

— Если мы можем выделить на сайте объекты, потенциально интересные среднестатистическому пользователю, то эти компоненты нужно постараться вытащить наверх, показать в первом экране. Например, контакты не должны быть написаны мелким шрифтом в уголке.

— Есть ли на Западе аналог uKit AI?

— Да, это проект The Grid. Когда я пришел в компанию, процесс разработки uKit AI уже шел, у его истоков стояли Павел Кудинов и Александр Пезиков. Вообще в области сайтов машинное обучение пока еще редко применяется. Идеи есть, но готовых проектов, которые могут сделать сайт, еще нет. Тот же The Grid не на все сто справляется со своей задачей.

— На сайте uKit AI подчеркивается, что проект разработан для массового рынка. Значит ли это, что любой владелец мелкого бизнеса может не нанимать отдельного специалиста и иметь сайт, соответствующий требованиям времени?

— Действительно, малому бизнесу постоянно поддерживать сайты в современном состоянии дорого, да и нет времени заниматься этим. Мы как раз предоставляем услуги для этого круга клиентов. Они очень легко могут перенести сайт на uKit, причем мы предлагаем эту услугу бесплатно. Клиент платит только за то, что сайт «живет» на нашей платформе. Наш искусственный интеллект будет поддерживать сайт в актуальном состоянии.

— Вы работаете над персонализированным дизайном (пользователь видит тот вариант сайта, который соответствует его предпочтениям и ожиданиям). Это возможно благодаря анализу цифрового следа?

— Это еще одно направление, в котором нам предстоит работать. На мировом рынке каких-то больших достижений в этой области нет, но мы собираемся заняться разработкой этого направления. Все люди разные: кто-то любит heavy metal и ему будет приятно увидеть черное оформление сайта, кто-то же любит маленьких пудельков и, наверное, розовый сайт ему будет интереснее. Это возможно благодаря оцениванию цифрового следа, анализа cookies (естественно, если пользователь разрешает такую опцию).

— Не потеряют ли сайты индивидуальность после генеративного дизайна?

— Интересный вопрос! Мы не думали об этом пока. Скорее всего нет. Одно из знаковых событий в истории нейросетей в сфере рекламы — когда Nutella в 2017 году выпустила около 7 млн. банок шоколадной пасты с уникальными этикетками, дизайн которых генерировался нейросетью. Это способствовало повышению продаж, потому что их аудитории было интересно получить баночку с уникальным рисунком. Я думаю, что мы тоже можем сделать так, чтоб у нас были уникальные сайты. Если мы загоним искусственный интеллект в узкие рамки, то да, он будет генерировать похожие друг на друга сайты, но мы знаем, как этого избежать.

— Чем компьютерное зрение отличается от машинного? Это синонимы?

— Есть понятие machine learning и есть термин computer vision. Последнее — это как раз про то, как из картинки извлечь информацию. А машинное обучение направлено на обучение программы. Если вы пишете программу, которая умножает два числа, то ее особо ничему учить не надо, но, если вы делаете программу со множеством параметров, вам понадобятся ее обучить. И научить систему подбирать параметры так, чтоб в конечном итоге сайт выглядел привлекательным для пользователя — это и есть машинное обучение.

— Продаете ли вы большие данные?

— Сегодня мы только храним их и используем для развития своих проектов. Насчет того, чтоб продавать — к нам еще никто не обращался. И это будет в любом случае решать руководство.

— Действующее российское законодательство не осложняет работу с Big Data?

— Нам пока не очень осложняет. Мы берем только обезличенные данные из открытого доступа, а также данные сайтов наших пользователей (uKit.com). Конечно, мы их не распространяем, но доступ имеем. Все, что опубликовано в Интернете — это открытые данные.

— Какие перспективы у больших данных в сайтостроении?

— У больших данных всегда большие перспективы. Сегодня данные — это та ценность, которой владеют далеко не все (поэтому набор данных стоит денег). Некоторые даже называют их нефтью современного мира. Собрать данные — большой труд, сами источники данных не всегда открыты. Большие данные (безотносительно их качества и происхождения) — неструктурированные объемы информации, которые выходят за рамки старого представления о данных. Не существует четкого определения их количества или веса, но их объем как минимум больше объема одного жесткого диска. Почему сегодня не может

появиться еще один мощный поисковый движок? Потому что любой новый поисковик будет начинать с нуля, совершать много ошибок, и пользователи будут уходить к Yandex, Google и т.д.

Если говорить про наши разработки, то примерно 80% времени уходит на работу с данными и 20% — на программирование процесса обучения. Летом 2017 года мы решили делать WebScore AI. Первое, что нам нужно было сделать, это собрать набор данных — датасет. Далее мы определились, откуда все это возьмем, написали соответствующий инструмент, который собирает данные (то есть это все протекает в автоматическом режиме). Следующей задачей было найти людей, которые поставят оценки сайтам — не специалистов в сайтостроении, а простых пользователей. Затем мы выработали систему, по которой асессоры должны были оценивать сайты (попарное сравнение, поочередное, списком). Мы показали асессорам около 1 500 сайтов — это, конечно, не так много, но само понятие больших данных очень спорно: например, ложка — это много или мало? Если ложка рыбьего жира, то, наверное, это много. А если ложка сладкого сиропа, то можно еще взять. Полутора тысяч сайтов было достаточно для того, чтоб принять решение о том, какой из способов разметки выбирать. Затем мы добавляли все большее количество сайтов в систему. Кстати, асессоров мы тоже отбирали, потому что некоторые из них недобросовестно относились к своей работе, в то время как для нас было важно качество данных. После того как данные были собраны и почищены, мы подали их инструменту Gradient Boosting для обучения искусственного интеллекта. Нейросеть мы тоже использовали, но как подзадачу. Итоговую оценку сайта определяет именно градиентный бустинг.

Самое главное, что есть результат — наша модель оценивает сайты в каком-то смысле лучше, чем наши асессоры. Мы подсчитали для каждого асессора, насколько он отклоняется от средней оценки, и для WebScore AI. И модель оценивает точнее: она ошибается в среднем на 0.5 балла, а пользователи на 1, 1.2 балла.