Тема 2.2. Цифровой след потребителя

Автор: к.ф.м.н., зам. директора МАЦК «Политехник-SAP», Амбражей А.Н.

На протяжении всей жизни мы движемся через пространство, время, взаимодействуем с окружающим миром и конечно же оставляем следы. Следы в реальном мире подвержены изменениям, а следы в виртуальном потенциально могут храниться вечно.

Поэтому практически любой контакт с компьютерными системами и часто вне зависимости от нашего желания приводит к образованию следа. В современном мире практически нет возможности избежать этого, поэтому надо понимать, что оставляет след и как это могут использовать.

Источники цифрового следа. Фотографии.

Самое простое и самое распространенные средство оставить след – это разместить фотографию, на любом сайте, фотохостинге, социальной сети или просто отправить по почте.

Каждый файл с фото, в зависимости от камеры или действия пользователя, хранит в себе определенную метаинформацию или метаданные в стандарте EXIF.

Стандарт EXIF является чрезвычайно гибким (например, позволяет сохранить полученные с приёмника GPS координаты места съёмки) и допускает широкое развитие — как правило, фотоаппараты добавляют к файлу информацию, специфичную только для данной конкретной камеры. Правильно интерпретировать такую информацию могут только программы от изготовителя фотоаппарата.

В качестве примера информации, записываемой в EXIF, можно указать следующее:

- производитель камеры, модель,
- информация о правообладании,
- выдержка, диафрагма, ISO, использование вспышки, разрешение кадра, фокусное расстояние, размер матрицы, эквивалентное фокусное расстояние и т.д.
- дата и время съёмки,
- географические координаты и адрес места съёмки
- пользовательская информация.

Поэтому самое простое – начать с анализа метаинформации, используя один из многочисленных онлайн сервисов, например, http://exif.regex.info/exif.cgi

Метаинформацию можно удалять или изменять, некоторые сервисы удаляют метаинформацию из фотографий, некоторые ее оставляют, в частности данные о координатах точки съемки.

Если же метаинформации недостаточно, то есть еще много способов работы с картинками.

Прежде всего по картинкам можно искать, это делают все системы.

https://images.google.com — поиск по фотографиям в Google (можно загрузить картинку или указать на нее ссылку)

https://yandex.ru/images/search — поиск от Яндекса по картинкам.

Алгоритмы отличаются, поэтому можно попробовать обе системы, помимо похожих фото могут быть выданы предположения о том, что изображено на картинке.

Так что если Вы просто сделали фото в отпуске, то найти в каком отеле Вы остановились не большая проблема. Таким образом получается, что не только люди, но и предметы или здания оставляют свои следы.

Картинка может быть сделана не Вами, а другими людьми или автоматизированными системами.

Распознавание лиц в потоке, идентификация и поиск людей, изображенных на фото – эти технологии уже активно применяются не только спец. службами и компаниями, работающими с системами безопасности, но и социальными сетями, программами управления фотоальбомами и даже интернет магазинами. Для социальных сетей это используется, например, для автоматической отметки друзей на фото.

В целом найти нужного человека по фото не составляет большой сложности, если он хоть где-то оставил свой след.

Кроме идентификации человека, можно по фото определить его пол, возраст и даже эмоции.

Пример распознавания эмоций: https://azure.microsoft.com/ru-ru/services/cognitive-services/emotion/ и определение возраста https://how-old.net/

Источники цифрового следа. Торговые сети

Возьмем, например, чек. Это факт покупки на некоторую сумму. Сам по себе он не очень полезен, но если добавить данные покупателя, дату, информацию о купленном товаре, магазине, т.е. метаинформацию, и сделать это для всех покупок в сети, то мы получим данные, на основе которых мы можем сделать определенные выводы, построить модели поведения, предсказать действия покупателя. Это и будет то что мы называем Большими данными.

По прогнозам, количество данных на планете будет удваиваться каждые два года вплоть до 2020 года. В период между 2013 и 2020 годом количество информации увеличится десятикратно — с 4,4 трлн гигабайт до 44 трлн. При этом по оценкам

International Data Corporation (IDC), к 2020 году только 35% данных будет содержать ценную для анализа информацию. Поэтому сейчас говоря об умных больший данных (smart big data), которые представляют собой уже отобранную полезную информацию, которую можно эффективно использовать для решения различных задач.

Но чтобы эффективно использовать эти данные нам надо идентифицировать покупателя. Привязка к покупателю нужна прежде всего для анализа его потребительских предпочтений и особенностей поведения. Что, как часто, какие продукты вместе и даже то, как он перемещается по магазину – все это предмет пристального изучения.

Для идентификации нас, как покупателей, есть много способов. Самый простой – карта лояльности. Это обычно и идентификатор и канал коммуникации т.к. карта привязана к e-mail или номеру телефона, который сейчас по факту совмещает в себе все каналы.

Единственный минус карт лояльности – идентификация идет на кассе, поэтому нет возможности сделать мгновенное предложение.

Кстати все больше карт перемещается в специальные приложения, так что возможности идентификации повышаются.

В частности, есть проекты идентификации по Wi-Fi. Магазин или кафе предоставляют клиентам бесплатный Wi-Fi и уже после первого же подключения можно определять что клиент такой-то появился в сети определенного магазина, вернее появилось его устройство.

В любом случае, нам нужно чтобы покупатель как-то себя идентифицировал. Например, используя персональный ручной сканер для покупок, который выдается при предъявлении карты лояльности.

Этот проект стартовал в сети Лента (https://lenta.com/lenta-testiruet-novye-tehnologii-samoobsluzhivaniya).

Другой способ - распознавание лиц. Мы можем фотографировать человека в момент выдачи карты и дальше распознавать его в момент входа в магазин. Даже не персонифицированных покупателей мы можем распознавать под к-то условным идентификатором и дальше контролировать любое посещение им магазина. Точность таких системе примерно 85%, они универсально, но дороги.

Источники цифрового следа. Фитнес-браслеты

Еще пример, когда большие данные существенно меняют картину – геотрекеры. В ноябре 2017 разработчик приложения для фитнеса опубликовал большую карту данных с маршрутами пользователей по всему миру. Сами траектории обезличены, но оказалось, что с помощью этой информации легко можно отследить местоположение секретных военных баз. Действительно, почему так много людей бегают по кругу в такой пустынной местности?

Один трек не дал бы такой информации, а множество дали нам понимание как перемещаются люди и позволили сделать соответствующие выводы. Большие данные – это больше чем просто много данных собранных вместе. Они позволяют узнавать закономерности, о которых мы ранее и не подозревали и создавать на их основе новые ценности и даже спасать людские жизни.

Кстати, виртуальным следом можно даже рисовать.

Источники цифрового следа. Большие данные

Термин Большие данные ввёл Клиффорд Линч, редактор журнала Nature, 3 сентября 2008 года, когда опубликовал в специальном номере журнала свои размышлениях на тему «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объемами данных?»

Сам термин конечно же не относиться только к объему данных. Это общее обозначение структурированных и неструктурированных данных огромных объёмов и при этом сложно устроенных.

Их характеристики можно представить, как пять «V»:

- объем (volume), накопленная база данных представляет собой большой объем информации, который трудоемко обрабатывать и хранить традиционными способами, для них требуются новый подход и альтернативные инструменты;
- скорость (velocity), данный признак указывает как на увеличивающуюся скорость накопления данных (90% информации было собрано за последние 2 года), так и на скорость обработки данных, в последнее время стали более востребованы технологии обработки данных в реальном времени;
- многообразие (variety), т.е. возможность одновременной обработки, структурированной и неструктурированной разноформатной информации (на сегодняшний день 80% информации входит в группу неструктурированной);
- достоверность данных (veracity), все большее значение пользователи стали придавать значимость достоверности имеющихся данных;
- ценность накопленной информации (value). Большие Данные должны быть полезны компании и приносить определенную ценность для нее;

Источники цифрового следа. Социальные сети

Социальные сети – это, наверное, самый мощный источник информации о человеке. Анализу может подлежать все:

- информация из профиля
- анализ изображений
- анализ связей с другими людьми

- анализ текстов публикаций
- анализ лайков и т.д.

Например, тут можно посмотреть предпочтение Фейсбук по рекламе, связанной с Вашим профилем:

https://www.facebook.com/ads/preferences/?id=6095947696607&entry_product=ad_preferences_dialog_v2

Есть открытые системы, которые осуществляют поиск по социальным сетям, например, yandex.ru/people, которая в настоящий момент поиск осуществляется по 16 соцсетям или поиск по постам в соцсетях: https://www.social-searcher.com/ и другие сервисы.

Сейчас данные сетей используют даже банки, которые анализируют профили в соцсетях для определения надежности клиента, его психотипа, уровня дохода, образа жизни и т.д.

Не обходят своим вниманием страницы работников и работодатели.

Источники цифрового следа. Специализированные сервисы, государственные сервисы

Сейчас государственные органы стараются максимально консолидировать свои базы и переносят часть своих сервисов в виртуальное пространство. Конечно это позволяет иметь все более точный потрет гражданина.

Ниже перечень открытых сервисов, по которым можно найти информацию о человеке или объекте.

- База недействительных паспортов (http://services.fms.gov.ru/info-service.htm?sid=2000).
- База судебных приставов (http://fssprus.ru/iss/ip/).
- База дипломов (http://frdocheck.obrnadzor.gov.ru).
- Федеральная налоговая служба - можно узнать ИНН (https://service.nalog.ru/inn-my.do). Для этого необходимо знать ФИО, дату рождения, серию и номер паспорта, а также дату выдачи паспорта. Аналогично на сайте ФНС можно узнать, является ли физическое лицо ИП (индивидуальным предпринимателем) https://egrul.nalog.ru. Поиск можно осуществлять по ФИО и региону, ИНН или ОГРНИП.Там же можно получить данные об учредителях по наименованию юр лица.
- Наличие судебных решений : Поиск осуществляется в системе ГАС «Правосудие» (государственная автоматизированная система).
    - по делам и судебным актам — http://sudrf.ru/index.php?id=300#sp

- по текстам судебных решений — https://bsr.sudrf.ru/bigs/portal.html

- Проверяем находится ли человек в розыске:
  - Поиск лица в базе данных МВД «Розыск» — https://mvd.ru/wanted
  - Информирование о лицах, находящихся в розыске по подозрению в совершении преступлений – http://fssprus.ru/iss/suspect_info/
  - Реестр розыска по исполнительным производствам — http://fssprus.ru/iss/ip_search/
  - Публичная кадастровая карта https://pkk5.rosreestr.ru

Источники цифрового следа. Смартфон

Сматрфон сейчас это основное средство коммуникации, идентификации и платежа.

Источниками информации служат:

- приложения магазинов, сервисов, систем лояльности;
- банковские приложения;
- СМС
- навигационные системы
- история поиска, геопозиционирование
- операционная система
- почта
- мессенджеры
- прочие приложения, которые имеют доступ к персональной информации

По данным за 2017 доля мобильных операционных систем составляла: Android — 73.52%, iOS — 19.37%, Windows — 0.77%.

Среди персональных компьютеров в марте 2018 года Google Chrome установлен на 59.66% компьютеров.

Т.о. операционная система от Гугл является абсолютным лидером среди систем как для смартфонов, так и для ПК.

Отслеживание информации встроено на уровне самой системы, это сделано для предоставления контекстной рекламы, хотя может использоваться гораздо шире.

- Ниже ссылки на то что знает о нас Google
- Настройка рекламных предпочтений https://adssettings.google.com/authenticated
- Передвижения https://www.google.com/maps/timeline?pb

- История поиска https://myactivity.google.com/myactivity
- Доступ приложений https://myaccount.google.com/permissions
- История просмотра на youtube https://www.youtube.com/feed/history/search_history
- Сервис поиска телефона https://myaccount.google.com/find-your-phone?continue=https%3A%2F%2Fmyaccount.google.com%2F

Таким образом мы понимаем, что любой контакт с любым сервисом или использование электронного устройства, каким бы безобидным это нам не казалось, приводит к образованию цифрового следа.

Собранные вместе эти данные образуют нашу цифровую тень.

Цифровая тень и цифровой двойник

Большое количество данных позволяет узнавать закономерности, о которых мы ранее и не подозревали и создавать на их основе новые ценности, продукты или даже спасать людские жизни.

Из массы информации, которую собирают о нас, формируется наша цифровая тень, образ, каким видят нас различные системы: образ пациента, покупателя, гражданина.

Это позволяет строить модели, предсказывать поведение, предпочтении, разрабатывать индивидуальные продукты, диагностировать состояние здоровье и многое другое.

Давайте рассмотрим ряд сценариев использования Больших данных и цифровой тени потребителя.

Все мы знаем, как одежда на картинке в интернет магазине отличается от того как она выглядит на нас. Поэтому нам всегда нужна примерка. Теперь представьте если бы вы могли загрузить на сайт производителя одежды свою точную цифровую модель и примерить на нее одежду.

А производитель, в свою очередь, получил бы бесценную информацию о том какие размеры одежды у его клиентов. Такие проекты уже есть.

Как бы это изменило торговлю одеждой? Думаю, что кардинально. Пойдем дальше, как насчет производства одежды под вас, с учетом всех анатомических особенностей?

Рекомендательные системы

Большинство сайтов используют рекомендательные системы, самое простое – это рекомендовать понятные товары типа аксессуаров, как футляр к телефону или объектив к фотоаппарату, эти товары связаны напрямую. Но сейчас этого явно недостаточно, надо проанализировать действия человека на сайте, понять, что и как он ищет, что ему надо или может понадобиться в будущем.

Большинство систем использует алгоритмы коллаборативной фильтрации. В их основе лежит предположение что если мы сгруппируем похожих пользователей по каким-то признакам, то скорее всего им будут нравиться одни и те же товары т.к. их поведение, как покупателей, одинаково.

Так же можно воспользоваться фильтрацией на основе контента, т.е. проследить историю предпочтений пользователя и предложить ему новый товар из того же ряда. Например, если покупатель книжного магазина предпочитает отечественные детективы, то скорее всего его заинтересует новый детектив российского автора. Существуют и гибридные модели.

Анализу подвергается не только то, что смотрит человек, но и то, как он взаимодействует с сайтом или сервисом. Как долго он смотрел страницу, сколько пробыл на сайте и т.д.

На основе этого анализа можем отнести пользователя к тому или иному типу и подстроить контент сайта или тип коммутации под него. Сайт изменит дизайн так, чтобы предложить аналитику больше информации, а человеку склонному к импульсивному поведению – яркие впечатления. И эта технология уже работает.

Сайты анализируют поведение пользователей, накапливают информацию, и все точнее подстраиваются под них. Даже когда мы ушли с сайта с нами продолжают взаимодействовать - работают скрипты на основе поведения. Остались товары в корзине, можно напомнить, можно предложить скидку или товар заместитель.

Тут главное понять, что произошло, а это не всегда очевидно.

На помощь приходят алгоритмы машинного обучения.

Мы даем на вход модели все данные о пользователе и известные нам результаты, например, то что пользователь покинул сайт ничего не купив, или закрыл вклад в банке, или же у него наступил страховой случай. Система сама может выделить нужные факторы и научиться предсказывать уже известный результат с достаточной точностью. Так система обучается.

Если на вход подать данные для нового пользователя, то мы можем предсказать результат для него с определенной точностью. Так работаю предиктивные системы, они обучаются на исходных исторических данных и делают предсказание. Следующее поколение систем, основанное на прескриптивной аналитике, готовы дать конкретные указания что делать. Т.о. мы формируем автоматические сценарии реагирования.

Другая полезная для анализа следа потребителя технология - это сентиментный анализ. Он позволяет оценить тональность высказывания, например, понять хвалят или нет ваш продукт в комментариях и отреагировать соответствующим образом. Например, сегментировать клиентов на основе их отзывов о новом продукте в соц. сетях.

Следующий случай – это уже не продажа, а производство товаров под конкретного индивидуального потребителя при сохранении цены аналогичной серийному товару. Это новый тип производства и уже сейчас есть примеры его реализации.

Давайте рассмотрим этот сценарий на примере производства автомобиля.

Что нам нужно знать о пользователе и откуда мы можем получить эту информацию:

- Зачем ему нужен автомобиль?
- Какие требования к его оснащению и техническим характеристикам?
- Какой должен быть дизайн?
- Какому стилю вождения должен отвечать автомобиль и т.д?

Можно ли это сделать полностью в автоматическом режиме, без участия пользователя? Для этого придется собрать довольно много данных из различных источников.

Данные о стиле вождения может дать навигационная система, требования к эргономике водителя можно получить, если у нас будут его антропометрические данные, анализ социальных сетей может дать информацию о составе семьи, работе, предпочтениях.

Конечно нужна еще информация о социальном статусе, финансовых возможностях и т.д.

В целом чем полнее будет «цифровое досье» или цифровая тень, тем более точным будет попадание в цель. Примерно так и должны работать Фабрики будущего, собирая информацию о потребителе и производя индивидуальную продукцию.

Но все ли данные нам нужны? Сейчас многие компании собирают данные, но не знают, что с ними делать, поэтому появляется понятие Умных Больших Данных, когда мы проектируем те точки сбора данных, которые нам необходимы.

Правовые аспекты использования информации

Кто же контролирует эти потоки данных и где грань между преимуществами, которые дает сбор данных о нас и тотальным контролем?

Это очень сложный этический вопрос.

За последнее время было большое количество утечек данных пользователей различных систем. За нашими данными идет настоящая охота, их продают, покупают, используют без нашего ведома: для рекламы, политики, формирования поведения и т.д.

Большая часть государств стремится регламентировать эту сферу и ограничить использование персональных данных пользователей. Они создают законы о защите персональных данных, которые должны определить кто, как и зачем собирает наши данные, также большинство законодательных актов предусматривает хранение данных на территории проживания их владельцев, ведь данные становятся таким же богатством как нефть, золото или другие полезные ископаемые. Появилось даже понятие «цифровые недра» и принадлежать они должны тем, кто их породил.

Основные законы, регламентирующие работу с персональными данными.

В Российской Федерации:

- Федеральный закон Российской Федерации от 27 июля 2006 г. N 152-ФЗ «О персональных данных»
- Федеральный закон "О внесении изменений в Федеральный закон О противодействии терроризму и отдельные законодательные акты Российской Федерации в части установления дополнительных мер противодействия

терроризму и обеспечения общественной безопасности" от 06.07.2016 N 374-ФЗ (Закон Яровой)

- Право на забвение. Удаление информации о человеке из поисковиков. С 01.01.2016 г. вступил в силу Федеральный закон №264 от 13 июля 2015 г. «О внесении изменений в Федеральный закон «Об информации, информационных технологиях и о защите информации» и статьи 29 и 402 Гражданского процессуального кодекса Российской Федерации». Этот закон именуется также как «закон о праве на забвение». Суть данного закона заключается в следующем: по требованию гражданина поисковая система должна удалить из поиска информацию о заявителе, распространяемую с нарушением Российского законодательства, недостоверную информацию, неактуальную, утратившую значение для заявителя. Исключения: информацию об уголовно наказуемых деяниях, если сроки привлечения к ответственности не истекли и информацию о совершенном преступлении, если не истекла судимость удалить нельзя. Оператор поисковой системы обязан в течении 10 дней удалить информацию. Отказ в удалении информации можно обжаловать в суд. Также необходимо понимать, что сама информация из Интернета физически не удаляется. Удаляется из поиска выдача ссылок на эту информацию. И если кто знает эту ссылку, он может зайти по ней напрямую и увидеть эту информацию.

В ЕС:

- GDPR.General Data Protection Regulation. Генеральный регламент о защите персональных данных.