Формализация информации и Big Data

«02.03.03 - Математическое обеспечение и администрирование информационных систем направленность разработка и администрирование информационных систем»

http//vikchas.ru

Лекция 9 «Обработка Big Data»

Часовских Виктор Петрович
д-р техн. наук, профессор кафедры ШИиКМ

ФГБОУ ВО «Уральский государственный экономический университет»

Екатеринбург 2023

Краудсорсинг (Crowdsourcing) инструмент Big Data

Обычно анализом Big Data занимаются компьютеры, но иногда его поручают и людям.

Для этих целей существует **краудсорсинг** — привлечение к решению какой-либо проблемы большой группы людей.

Краудсорсинг — это практика использования знаний группы ради общей цели.

Больше всего она полезна при решении сложных проблем инновационным образом или для упрощения сложных процессов.

Начиная с 2010-х гг. краудсорсинг стал удобным инструментом поиска отклика от множества лиц, заинтересованных в решении той или иной проблемы, получив широкое распространение как в научной среде, так и в бизнес-сообществе.

Основная предпосылка исследования с использованием данного метода базируется на многообразии общественных интересов и потребностей, их сходстве и различии. Это осуществляется путем зачастую удаленного взаимодействия, при высказывании мнений как можно большего числа респондентов, для которых решение поставленной задачи нередко является областью специализации.

Для демонстрации возможных сфер применения краудсорсинга целесообразно привести классификатор Г. Сакстона, основанный на четырех критериях: вид услуги или продукта, который отдается на аутсорсинг; функции краудсорсеров; степень сотрудничества краудсорсеров и организация проекта.

Приведенная классификация Сакстона группирует задачи, решаемые в рамках типовых реализуемых проектов. Как мы видим, во всех случаях реализации краудсорсинговых проектов имеет место запрос задачи от организаторов проекта, а также открытый сбор мнений от лиц, вовлеченных в решение поставленной задачи.

Классификация проектов краудсорсинга согласно Сакстону

Модель	Описание
Модель посредника	Организаторы проекта выступают в роли посредника между заказчиком и исполнителями-краудсорсерами
Производство медиа гражданами	Создание коллективного медиапродукта
Совместная разработка ПО	Адаптация модели «сайта-посредника» для создания компьютерных программ
Продажа цифровых товаров	Электронная торговая площадка, где краудсорсеры размещают цифровой контент — текст, фотографии, видео, графику и т. д.
Разработка дизайна продуктов	Торговая площадка, где краудсорсеры выставляют свои варианты дизайна; создатели проекта налаживают производство и доставку вещей с выбранным дизайном
Децентрализованное финансирование	Площадка для организации взаимодействия заемщиков и инвесторов
Потребительский отчет	Частный случай модели «производства медиа гражданами»: краудсорсеры оставляют только отзывы о товарах или работниках
База знаний	Ответы на вопросы пользователей, аналитические статьи, инструкции, бизнес-идеи и т. д.
Научно-технический проект	Краудсорсерам предлагается решить головоломку, подобрать подпись к изображениям

С развитием взаимодействия между пользователями посредством коммуникаций в рамках глобальной сети Интернет в краудсорсинговые проекты вовлекается все больше пользователей прикладного программного обеспечения. Данная вовлеченность реализуется как в игровой форме — в виде целевых игр (порталы типа Fold.it и EteRNA), так и в форме проектов, интегрирующих координаты краудсорсеров и внешние базы данных, например Яндекс.Пробки. Преимущество краудсорсинга при этом заключается в том, что отбор респондентов не является случайным: как правило, в сборе мнений по тому или иному вопросу участвуют подготовленные и заинтересованные лица, а их отклик в рамках обсуждения либо в ответ на событие является логически аргументированной реакцией.

Отечественные исследователи предлагают классификацию проектов краудсорсинга исходя из уровня участия различных ресурсов (человеческого капитала либо использования электронных ресурсов при принятии решений). В частности, возможна следующая классификация краудсорсинговых проектов.

Первая группа предполагает использовать на входе только «контент человека.

К данной группе отнесены проекты, где краудсорсеры поставляют продукт своего труда: текст, фотографию, рисунок, графику, видео и т. д., при этом внешние ресурсы не импортируются.

Такие проекты охватывают широкий круг задач, связанных с разработкой новых продуктов, анализом разрозненной и не структурированной информации и приданием ей систематичности и искомых свойств.

К кругу данных проектов можно отнести контент-проекты, направленные на сбор и представление информации путем аккумулирования ресурсов так называемого «коллективного разума» участников различных площадок (порталы для генерации идей и разработки программного обеспечения, базы знаний типа Википедии, базы резюме и вакансий: wikipedia.org, Ответы@mail.ru (otvet.mail,ru) и ряд других.

Вторая группа объединяет на входе «контент человека» и внешние ресурсы. Эти проекты затрагивают привлечение человеческого ресурса для решения поставленной задачи с учетом уже используемых данных — определенных потребительских предпочтений, сформированных моделей, шаблонов и форм.

Роль участников проекта сводится к приданию конечному продукту специфических черт и отличительных особенностей.

Внешними ресурсами могут быть не только материалы, но и базы данных.

Фактически участники краудсорсиногового проекта занимаются доработкой уже готового продукта и адаптацией его для конечного потребителя на конкретных специфических рынках.

В **третью группу** попадают проекты, в которых мобильные устройства краудсорсеров поставляют автоматически генерируемый «технический контент» — координаты или показания датчиков, который может дополнять традиционный «контент человека».

Проекты позволяют принимать решения с учетом изменившейся обстановки при условии использования технологий массовой обработки данных.

В качестве примера можно привести сервис «Делимобиль», использующий определение местоположения краудсорсера. Водители, желающие подработать, посылают в центр обработки данных предполагаемый маршрут, а пассажиры — свое местоположение и место назначения.

Четвертая группа представляет собой на входе «технический контент» и внешние ресурсы и (иногда) «контент человека» В эту группу включены проекты с импортом внешних ресурсов, в которых краудсорсеры поставляют «технический контент», а иногда и «цифровой контент». Типичный пример — российский проект Яндекс.Пробки, который интегрирует GPS-координаты краудсорсеров-автомобилистов, данные транспортных организаций и наземных видеокамер, а также карты, полученные путем спутникового мониторинга. Сервис позволяет отслеживать загруженность автодорог, прокладывать оптимальный маршрут с учетом дорожных пробок и осуществлять краткосрочный прогноз дорожного трафика.

Оценка результатов, полученных методом краутсорсинга, находится в родстве с контент-анализом, т. е. предполагает, во-первых, выделение ключевых признаков (или ключевых слов), характеризующих данный объект, во-вторых, определение градаций (состояний), в которых может пребывать объект, в-третьих, создание какого-либо конечного продукта или получение иного результата, характеризующего данный объект в целом.

Возможный результат применения краудсорсинга многообразен, как и количество областей его применения.

Преимущества краудсорсинга заключаются в возможности создания с его помощью собирательного образа исследуемого объекта, применяя знания и опыт множества людей, вовлеченных в данный процесс.

Нередко краудсорсинговые площадки становятся центром притяжения экспертного сообщества, способного генерировать множество вариантов решения той или иной проблемы с позиции специфики деятельности. В качестве экспертов могут быть привлечены лидеры мнений в регионе, в области специализации, а также из-за рубежа.

A/B-тестирование

A/B тестирование — это метод исследования, при котором разным группам посетителей сайта одновременно показаны две версии одной и той же веб-страницы для определения, какая из них работает эффективнее.

Если вкратце, то исследование показывает, какая из **версий (A) или (B)** лучше?

A/B тестирование, также известное как **сплит-тестирование (split testing)** или **групповое тестирование (bucket testing)**, по сути, представляет собой эксперимент, в котором пользователям случайным образом показываются два или более варианта **рекламы, маркетингового емэйла или веб-страницы**, а затем используются различные методы статистического анализа для определения какой вариант дает больше конверсий.

А/В-тестирование, представляющее собой набор из двух альтернатив и их демонстрацию большому количеству пользователей.

Цель А/В-тестирования заключается в поиске наиболее предпочтительного варианта из двух альтернативных (взаимоисключающих). А/В-тестирование представляет собой метод, позволяющий найти более эффективное решение за счет сбора обратной связи одновременно от большого числа респондентов (пользователей тестируемой системы). В качестве материала для применения А/В-тестирования можно использовать различные существующие IT-инструменты, а именно:

Особенности применения А/В-тестирования при использовании различных IT-инструментов

Инструмент	Оцениваемый параметр
Лендинги и отдельные посадочные страницы, веб-сайты	Формат посадочной страницы (удобство использования одного из двух вариантов лендингов)
Письма в email-рассылках, посты в социальных сетях	Оценка отклика на один из двух форматов информационного сообщения

Исходным материалом для проведения A/B-тестирования является некая альтернатива, предусматривающая наилучший отклик со стороны пользователей.

В ходе проведения A/B-тестирования разработчик решает ряд задач, состоящих в получении ответов на некоторые вопросы:

1. Какое визуальное оформление контента наиболее привлекательно для пользователей?

2. Каким наилучшим образом можно разместить контент на лендинге (одностраничный сайт с краткой информацией) для наибольшего отклика пользователей?

3. Какой текст поста получает наибольший отклик от пользователей?

Чтобы получить от A/B-тестирования положительный эффект, необходимо выполнить ряд требований:

1) тестирование альтернатив только в рамках одного изменения. Подача контента двум различным тестовым группам (одной группе — текущий вариант, другой — планируемый вариант) позволит получить отклик об удобстве предлагаемых изменений, а также ускорить адаптацию к данным изменениям;

2) одновременность проведения тестирования в различных аудиториях. Это позволит получать отклик от двух тестовых групп одновременно, без поправки на временной интервал, а также оценить реакцию на точечное изменение одного тестируемого свойства.

К преимуществам данного метода можно отнести достаточно высокую степень объективности оценок пользователей, построенную на визуальных предпочтениях наибольшей доли аудитории. В этой связи А/В-тест представляет собой аналог опроса пользователей посредством информационного контента, только вместо прямых вопросов, предполагающих альтернативные варианты ответа, пользователю предлагается отклик на тот или иной формат контента.

Недостатком метода является длительность разработки инструментов оценки - внедрение тех или иных точечных изменений предполагает трудозатраты команды разработчиков на внедрение инструментов и запуск их в действующей информационной системе.

В целом же А/Б-тест представляет собой эффективный инструмент интернет-маркетинга для повышения лояльности аудитории.

Прогнозная аналитика

Предсказательная (прогнозная) аналитика (предиктивная аналитика) представляет собой метод анализа данных, позволяющий экстраполировать ретроспективу развития той или иной системы на будущие периоды.

Под прогнозированием следует понимать научное выявление вероятных путей и результатов предстоящего развития явлений и процессов, оценку показателей, характеризующих эти явления и процессы для более или менее отдаленного уровня.

Прогнозирование — это научная деятельность, направленная на выявление и изучение возможных альтернатив будущего развития и структуры его вероятных траекторий. Каждая альтернативная траектория развития связывается с наличием комплекса внешних относительно исследуемой системы или явления условий.

К преимуществам прогнозной аналитики относятся:

1) доступность и распространенность инструментария проведения исследований. Используемый статистический инструментарий (скользящие средние, взвешенные скользящие средние, темпы прироста — среднегодовые, среднемесячные) позволяет провести экспресс-прогнозрование в случае выделения относительно стабильной тенденции;

2) возможность осуществления прогнозирования без применения специального программного обеспечения.

Недостатки прогнозной аналитики:

1) в ряде случаев невозможность учесть множественные параметры и предпосылки для нелинейного тренда;

2) слабый учет качественных сдвигов, возникающих после достижения системой точек бифуркации, так как они построены на количественных, вероятностных методах;

3) исходя из двух предыдущих особенностей целесообразно использовать методы прогнозной аналитики только на коротких временных интервалах.

Самообучающиеся системы (искусственный интеллект)

В последние годы набирает влияние новый тренд — внедрение в производственные и информационные системы искусственного интеллекта (ИИ).

Данный тренд задан в связи с бурным развитием визуального контента и средств обработки массивов данных.

Актуальность метода заключается в необходимости экономии ресурсов при организации деятельности крупных промышленных систем за счет апробации результатов на математической модели, с учетом вероятных направлений развития системы и ее реакции на различные внешние воздействия.

Для построения систем искусственного интеллекта необходима разработка и внедрение таких алгоритмов обработки информации, которые позволяют не только накапливать, хранить и извлекать данные из многообразия доступной информации, но и, адаптируясь к ее качеству, самостоятельно решать вопросы совершенствования технологий ее обработки.

Такие алгоритмы называются сложными когнитивными структурами или нейросетевыми структурами.

Одной из задач искусственного интеллекта является обработка образов. Данная деятельность предполагает установление визуально-логической связи между имеющимся в наличии электронным образом некоего объекта и оценкой свойств данного объекта. При этом широко используется способность искусственного интеллекта к обучению, к накоплению, хранению и извлечению больших массивов информации, поиску закономерностей в них. Если между входными и выходными данными существует какая-то связь, даже не обнаруживаемая традиционными корреляционными методами, то искусственный интеллект способен автоматически настроиться на нее с заданной степенью точности.

Основной принцип ИИ — машинное обучение, т. е. создание такого алгоритма, который способен проанализировать большой объем данных, найти взаимосвязь полученных результатов, построить предиктивные и регрессионные модели.

Сетевой анализ

Еще одним методом, применяемым при обработке больших массивов данных, является сетевой анализ. Анализ сетей позволяет использовать различные подходы к исследованию данных и выявлению взаимозависимостей: статистические, кибернетические, системные, имитационные.

Применение сетевого анализа позволяет структурировать связи, установленные взаимодействия между различными социальными единицами: людьми и организациями. Для описания взаимодействий между элементами системы посредством установления социальных связей между ними используется понятие «социального графа», т. е. способа визуализации многообразия исследуемых взаимозависимостей между множеством элементов в пространстве, в том числе в виртуальной среде.

При взаимодействии в рамках построенной сети все ее элементы обмениваются ресурсами и информацией в пределах выстроенных взаимосвязей.

Результатом применения сетевого моделирования является модель распределения ресурсов некой системы: общества, государства, компании.

Модель позволяет оперировать множеством взаимосвязей между отдельными объектами и динамически оценивать значимость каждого из факторов сетевого взаимодействия.

Среди преимуществ данного метода следует указать способность к визуализации взаимосвязей между объектами в рамках одной системы.

Факторы, входящие в систему, обладая различными параметрами, характеризующими их центральное либо отдаленное положение, обеспечивают влияние на смежные элементы. Данное влияние необходимо оценивать динамически, по мере изменения расстановки сил в рамках системы.

К недостатком метода относится трудоемкость установления взаимосвязей между элементами сложной системы.

РЕЗЮМЕ.

Краудсорсинг является удобным инструментом поиска отклика от множества лиц, заинтересованных в решении той или иной проблемы.

A/B-тестирование представляет собой набор из двух альтернатив и их демонстрацию большому количеству пользователей с целью сбора обратной связи от пользователей, отдающих предпочтение той или иной альтернативе.

Прогнозная аналитика — метод анализа данных, позволяющий экстраполировать ретроспективу развития на будущие периоды.

Системы ИИ способны вырабатывать решения исходя из результатов анализа эффективности решений, принятых в прошлом.

Сетевой анализ с целью выявления взаимозависимостей в исследуемых данных комбинирует системные, статистические, кибернетические, имитационные и прочие методы.

Благодарю за внимание!