АНАЛИЗ СОЦИАЛЬНЫХ ДАННЫХ С ПОМОЩЬЮ ТЕХНОЛОГИЙ BIG DATA

Н.И. Журбич, П.А. Зяблецев
Томский политехнический университет
niz1@tpu.ru, paz4@tpu.ru

Введение

Социальные данные – самое ценное сырье XXI века, новая нефть, которая неиссякаема, считает исследователь, автор новой книги «Big Data. Вся технология в одной книге» Андреас Вайгенд [1]. Однако, как и нефть социальные данные требуют обработки до того, чтобы извлечь из них полезную информацию. Данный вопрос как никогда актуален в эпоху информационных технологий, так как объемы этих данных растут в геометрической прогрессии. Данная статья посвящена вопросу применения анализа социальных данных в разных сферах деятельности и способы анализа и обработки этих данных.

Области использования социальных данных (персональных данных)

Персональные данные (ПД) или личностные данные — любые сведения, относящиеся к прямо или косвенно определённому, или определяемому физическому лицу (субъекту персональных данных), которые предоставляются другому физическому или юридическому лицу либо лицам [2].

Под социальными данными в свою очередь понимается совокупность персональных данных большого количества людей. Специалистам в области Data Science важны не личные данные конкретного человека, а большие объемы этих данных для анализа и построения общей модели поведения группы людей.

Источниками социальных данных являются:

- Данные из социальных сетей.
- Данные интернет провайдеров (посещаемые сайты, покупки в интернете, скачиваемый контент и т.д.).
- Данные о физической активности пользователя (сбор информации с носимых устройств и устройств контроля жизнедеятельности).
- Данные из государственных систем.

Наличие большого количества источников позволяет обеспечивать поддержку информации в актуальном состоянии, а также накапливать набор данных о человеке или группе людей за определенный промежуток времени.

Одним из самых перспективных ресурсов для анализа персональных данных являются социальные сети. Во-первых, поведение людей в социальных сетях более открыто в сравнении с реальной жизни, где каждому необходимо подстраиваться под социальные нормы. Во-вторых, практически каждый человек зарегистрирован и пользуется той или иной социальной сетью. Многие крупные компании понимают важность информации в социальных сетях и требуют указать ссылку на них при приеме на работу.

Развитие технологий обработки больших данных привело к тому, что в настоящее время появился интерес к использованию различных данных пользователя из социальных сетей. Использование такого рода данных помогает решить следующие задачи:

- борьба с мошенничеством,
- целевой маркетинг (реклама товаров и услуг),
- управление брендом,
- формирование новых каналов сбыта (анализ покупательской способности отдельного региона).

Основная область применения социальных данных это маркетинг. В первую очередь, анализ данных покупателей/клиентов позволяет делать рекламу более направленной и соответственно более эффективной в соотношении цена/качество. Использование технологий больших данных позволяет предлагать каждому клиенту (пользователю) индивидуальный набор товаров или услуг, основанный на его предпочтениях. В начале прошлого века торговые компании тоже пытались выявить предпочтения людей в зависимости от их региона проживания, социального статуса и т.п., чтобы сделать предложения определенной группе людей, которая с большой вероятностью им заинтересуется. Тогда вычислительные мощности не позволяли производить такое количество операций, как сейчас. Более того, не существовало таких технологий, с помощью которых возможно было выявлять индивидуальные предпочтения каждого человека или делить людей на различные группы и кластеры. На сегодня данная область развивается настолько стремительно, что способствует улучшению технологий обработки данных.

Следующей сферой применения анализа пользовательских данных является борьба с мошенничеством. Во-первых, финансовые организации стремятся минимизировать свои убытки путем сокращения рискованных операций. Поэтому, выявление неблагонадежных заемщиков является одной из ключевых задач для кредитных организаций. Анализ данных о клиенте банка помогает формировать для него наиболее выгодные предложения исходя из индивидуальных потребностей, а также использовать информацию о клиенте для оценки платежеспособности человека. Это позволит уменьшать процентную ставку по кредиту надежным заемщикам.

Помимо прикладных отраслей анализ пользовательских данных может использоваться для различных научных исследований, связанных с психологией. С помощью анализа социальных сетей стало проще проводить эксперименты (например, определять зависимости между психологическим

портретом личности и его предпочтениями в музыке, литературе, кинематографе и т.д.). Это позволяет проводить психологические исследования с большей точностью, так как люди ведут себя подругому, если знают, что за ними наблюдают.

Методы интеллектуального анализа данных

Работа с большими наборами данных дает возможность создавать обобщенные результаты анализа данных по группам и сопоставления этих данных. В настоящее время существует много различных инструментов и методов для анализа и обработки больших данных.

Ниже приведены несколько методов, которые используются для интеллектуального анализа данных:

- Ассоциация;
- Классификация;
- Кластеризация;
- Прогнозирование;
- Последовательные модели;
- Деревья решений

Ассоциация – простое сопоставление двух или более элементов, в большинстве случаев одного и того же типа [3]. Данный метод позволяет создавать рекомендации пользователям на основе их поведения на сайте, благодаря тому, что выявлены связи различных товаров между собой. Например, при покупке нового телефона 90% клиентов приобретают различные аксессуары к нему.

Классификация используется для получения представления о типе клиентов, товаров или объектов, описывая несколько атрибутов для определения необходимого класса. Например, покупателей можно классифицировать по возрасту и социальной группе, что в свою очередь позволит более эффективно взаимодействовать с каждой группой людей.

Метод кластеризации используется для получения структурированного заключения, которое позволяет определить наиболее часто встречающиеся значения. Для этого необходимо сгруппировать отдельные элементы данных, исследуя один или более атрибутов класса. Кластеризация полезна при определении различной информации, так как она коррелируется с другими примерами, где можно увидеть, как диапазоны различных величин согласуются между собой.

Прогнозирование чаще всего применяется для выявления случаев мошенничества, прогнозирования прибыли компании и предсказания отказа компонентов той или иной системы. В комбинации с другими методами интеллектуального анализа данных прогнозирование помогает анализировать определенные тенденции, классифицировать различные параметры, сопоставлять их с другими моделями и отношениями.

Метод последовательных моделей является одним из самых популярных и применяется для анализа долгосрочных данных. Данные модели позволяют проследить определенные тенденции или регулярные повторения определённых действий в конкретный промежуток времени.

Дерево решений связано с большинством других методов (классификация и прогнозирование). Чаще всего деревья решений применяются для отбора данных, либо для поддержки выбора некоторых данных в рамках модели.

Все вышеперечисленные методы существуют только в комбинации между собой, так как на практике очень редко применяются по отдельности.

Заключение

В результате обзора областей применения социальных данных выявлено: интеллектуальный анализ данных активно развивается и постепенно входит в различные сферы повседневной жизни, такие как: системы поиска, облачные вычисления, социальные и информационные сети, биология и медицина, разработка ПО, мобильные и беспроводные технологии.

Также сделан вывод о том, что все рассматриваемые методы интеллектуального анализа данных существуют только в комбинации между собой, по отдельности данные методы не имеют практической пользы и некорректно отображают текущую ситуацию.

Список использованных источников

1. Андреас Вайгенд. Big Data. Вся технология в одной книге. «Издательство «Эксмо», 2018 – 480 с.
2. Федеральный закон РФ от 27 июля 2006 года № 152-ФЗ «О персональных данных»
3. Методы интеллектуального анализа данных [Электронный ресурс]. – URL: https://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques/index.html (дата обращения 05.11.2018).
4. Замятин А.В. Интеллектуальный анализ данных: учеб. пособие. – Томск: Издательский Дом Томского государственного университета, 2016. – 120 с.