Оглавление

Предисловие от издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв прямо на нашем сайте **www.dmkpress.com**, зайдя на страницу книги, и оставить комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу **dmkpress@gmail.com**, при этом напишите название книги в теме письма.

Если есть тема, в которой вы квалифицированы, и вы заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу **http://dmkpress.com/authors/publish_book/** или напишите в издательство по адресу **dmkpress@gmail.com**.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы удостовериться в качестве наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в тексте или в коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от расстройств и поможете нам улучшить последующие версии этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу **dmkpress@gmail.com**, и мы исправим это в следующих тиражах.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконно выполненной копией любой нашей книги, пожалуйста, сообщите нам адрес копии или веб-сайта, чтобы мы могли применить санкции.

Пожалуйста, свяжитесь с нами по адресу электронной почты **dmkpress@gmail.com** со ссылкой на подозрительные материалы.

Мы высоко ценим любую помощь по защите наших авторов, помогающую нам предоставлять вам качественные материалы.

Предисловие

Когда в 2003 году я окончил университет, были опасения, что в Европе программисты будут не востребованы, поскольку все разработки будут проводиться в странах с гораздо более низким уровнем зарплат. К счастью, по множеству причин эти опасения не сбылись. Как я предполагаю, одной из главных проблем было то, что компании недооценивали следующую проблему: разработчики не понимали культурных реалий тех мест, где будут применяться их разработки. Разрабатывалось необходимое программное обеспечение, но функциональные возможности не соответствовали ожиданиям.

В настоящее время аналогичная проблема возникла в сфере машинного обучения и анализа данных. Разница лишь в том, что у ее истоков не низкие зарплаты, а программное обеспечение как услуга (SaaS), система, в которую вы загружаете данные и которая делает за вас всю остальную работу.

Меня, как и всех остальных, беспокоит тот факт, что машины не понимают домены и людей. Машины еще не настолько умны, чтобы можно было исключить человека из задачи. Развитие идет быстрыми темпами, но я рискну предположить, что любой, кто читает эту книгу, сможет работать с рекомендательными системами вплоть до завершения своей трудовой деятельности.

Как я попал в эту сферу? Я работал разработчиком ПО в Италии и собирался переезжать в Англию, где мне хотелось получить работу поинтереснее простого управления данными в базе. К счастью, на меня вышел замечательный сотрудник кадрового агентства RedRock Consulting Ltd. Агентство связало меня с компанией, занимающейся разработкой рекомендательных систем, куда меня взяли для работы над базовым программным обеспечением. И все – я с головой ушел в машинное обучение (мне было невероятно интересно). Помимо работы над рекомендательными системами, я начал изучать информацию об интернете и читал множество книг по этой и схожим тематикам.

Сейчас и чихнуть нельзя без того, чтобы как минимум 10 человек не попытались поведать вам какие-то знания из области машинного обучения. Мне очень удивительно видеть одностраничные или одночасовые учебные курсы, авторы которых заявляют, что научат вас всему, что нужно знать о машинном обучении. С таким же успехом я могу написать руководство, как стать летчиком-истребителем:

«Взлетайте и управляйте полетом с помощью штурвала. Если вам нужно стрелять, нажмите кнопку. Прежде чем у вас закончится топливо, необходимо совершить посадку…»

Подобное руководство, возможно, станет для вас неплохой отправной точкой – я начинал точно так же. Но не пытайтесь себя обмануть: для понимания машинного обучения необходимы комплексные знания. Добавьте сюда человеческий фактор, который всегда еще больше все усложняет.

Возвращаясь к моей истории: я работал над рекомендательными системами и получал от этого удовольствие, а затем сменил работу. На новой должно-

сти я должен был продолжить работу над рекомендательными системами, но этот проект был отложен. Тогда я стал переживать, что больше мне не представится возможность работать с рекомендательными системами, и именно в этот момент издательство Manning предложило мне написать о них книгу. Конечно же, я с радостью взялся за это задание. Как только контракт был подписан, проект по рекомендательным системам все-таки начался. Работая над этой книгой, я существенно обогатил свои знания и надеюсь, что и вам она принесет пользу.

Задача этой книги – познакомить вас с рекомендательными системами: не только с алгоритмами, но и со всей экосистемой. Алгоритмы не отличаются особой сложностью, но, чтобы их понимать и применять, необходимо понимать пользователей, для которых система будет предлагать рекомендации. Содержание книги менялось в процессе ее написания, поскольку я пытался включить в нее как можно больше информации. Надеюсь, что, прочитав эту книгу, вы получите все необходимые знания, чтобы начать работать с рекомендательными системами, и у вас появится база для дальнейшего обучения.

Благодарности

Я хочу отметить и поблагодарить две группы людей: тех, кто активно работал над книгой, и тех, кому приходилось терпеть и поддерживать меня в течение последних трех лет, пока я работал, не обращая внимания ни на что вокруг.

Хотя на обложке «Рекомендательных систем на практике» стоит мое имя, эта книга не появилась бы на свет без активного участия замечательных специалистов из издательства Manning. Я хочу отдельно поблагодарить редактора-консультанта Хелен Стергиус (Helen Stergius) за постоянную помощь и содействие. Она и другие сотрудники превратили мой несколько несвязный текст в руководство, обучающее людей создавать рекомендательные системы.

Также я хочу поблагодарить технических редакторов Фуркана Камаси (Furkan Kamaci) и Валентина Креттаза (Valentin Crettaz), а также всех рецензентов, которые нашли время прочитать первые варианты книги и помогли мне отточить текст, в их числе: Адхир Рамджиаван (Adhir Ramjiavan), Александр Мильцев (Alexander Myltsev), Элвин Радж (Alvin Raj), Амит Ламба (Amit Lamba), Эндрю Колльер (Andrew Collier), Фазел Кештkар (Fazel Keshtkar), Джаред Дункан (Jared Duncan), Яромир Немец (Jaromir Nemec), Мартин Бир (Martin Beer), Маюр Патил (Mayur Patil), Майк Далримпл (Mike Dalrymple), Норин Дертинджер (Noreen Dertinger), Оливье Дукаттеу (Olivier Ducatteeuw), Питер Хэмптон (Peter Hampton), Симеон Лейзерзон (Simeon Leyzerzon), Серен Линд Кристиансен (Søren Lind Kristiansen), Стивен Парр (Steven Parr), Тобиас Бергер (Tobias Bürger), Тобиас Гетрост (Tobias Getrost) и Випул Гупта (Vipul Gupta).

Работая над книгой, я обращался ко множеству различных библиотек, систем и баз данных, и очень благодарен всем сообществам, которые мне помогали. Также я благодарен сообществам разработчиков ПО с открытым исходным кодом, которые создали ряд инструментов, избавив нас от необходимости делать все с нуля.

И в первую очередь я благодарю свою жену, сына и тещу, остальных родственников, а также близких друзей за поддержку, любовь и, самое главное, терпение. Им было нелегко жить с членом семьи, который все время норовит ускользнуть и засесть за написание книги, пока вся семья переезжает в новый дом и лицезреет, как наши дома в Италии разрушаются до основания в результате землетрясений. Не говоря уже о том, что при этом новоиспеченный писатель взялся одновременно за две новые работы. Спасибо вам, и я обещаю, что хотя бы пару лет не буду начинать новые проекты. Люблю вас всех!

О книге

Испытываете ли вы зависть, когда видите рекомендуемые товары Amazon или когда Netflix попадает в точку с рекомендациями подписчикам? Тогда у вас появился шанс пополнить свой арсенал подобными умениями. Читая эту книгу, вы получите представление о том, что такое рекомендательные системы и каково их практическое применение. Чтобы рекомендательная система работала, необходимо, чтобы одновременно выполнялось несколько процессов. Нужно понимать, как собирать данные о пользователях и как их интерпретировать, а также нужно владеть различными алгоритмами рекомендательных систем, чтобы для каждой ситуации можно было выбрать наиболее подходящий из них. И самое главное, нужно уметь чувствовать, хорошо ли функционирует рекомендательная система. Все это и многое другое вы найдете в данной книге.

Кто должен прочитать эту книгу

Книга «Рекомендательные системы на практике» в первую очередь предназначена для разработчиков, которые хотели бы создать рекомендательную систему. В книге собраны практические советы, и материал изложен простым, доступным языком. Есть и вычисления, и статистика, но всегда обязательно присутствуют цифры и программный код. Новоиспеченные специалисты по обработке и анализу данных также почерпнут много полезного из данной книги: получат базовые представления об алгоритмах рекомендательных систем и инфраструктуре, необходимой для их запуска и функционирования. Менеджерам эта книга будет интересна в качестве общего пособия по рекомендательным системам: что это такое и как это можно применить на практике.

Чтобы получить от книги максимум пользы, нужно разбираться в языках программирования, таких как Python и Java, понимать SQL-запросы и обладать базовыми знаниями высшей математики и статистики. Цифры и листинги кодов, которые приводятся для наглядности, дают лишь самое общее представление о теме.

Структура книги

Книга делится на две части: первая посвящена инфраструктуре рекомендательных систем а вторая – алгоритмам.

Из части I вы узнаете, как после добавления рекомендательной системы в приложение получать данные и применять их:

- глава 1 представляет собой общий обзор рекомендательных систем и их ключевых элементов. Она знакомит читателя с рекомендательными системами в целом и базовыми принципами их работы;

- глава 2 говорит о том, как научиться понимать пользователей и их поведение, а также перечисляет способы сбора данных о пользователях;
- глава 3 посвящена веб-аналитике и рассказывает, как можно создать сводную информационную панель для отслеживания данных о своих рекомендательных системах;
- глава 4 говорит о том, как на основе данных о поведении пользователей составлять рейтинги;
- глава 5 рассматривает общие рекомендации (без учета индивидуальных предпочтений пользователя);
- глава 6 описывает проблемы, связанные с новыми пользователями и товарами, и предлагает простые решения.

Часть II рассказывает об алгоритмах рекомендательных систем, а также о том, как на основе собранных системой данных выстроить рекомендации для пользователя:

- глава 7 говорит о формулах для определения степени сходства между различными пользователями или контентом, например фильмами;
- глава 8 рассказывает о составлении персональных рекомендаций путем совместной фильтрации;
- глава 9 приводит методы оценки рекомендательной системы вне интернета и описывает способы составления рекомендаций онлайн;
- глава 10 знакомит вас с фильтрацией по контенту, которая позволяет найти сходные черты в контенте с помощью различных типов алгоритмов, таких как латентное размещение Дирихле и TF-IDF;
- глава 11 возвращается к обсуждению совместной фильтрации, о которой говорилось в главе 8, но сосредотачивается на методах снижения размерности;
- глава 12 рассказывает, как можно совместить различные типы рекомендательных систем;
- глава 13 говорит об алгоритмах ранжирования и способах оценивания рекомендаций;
- глава 14 подводит общий итог, дает задел на будущее, рассказывает, какие вопросы необходимо изучить и какие книги нужно еще прочитать, а также рассуждает об алгоритмах и контексте.

Книга построена таким образом, что лучше ее читать от начала до конца, поскольку во многих местах есть отсылки к предыдущим главам, но вполне допустимо читать и отдельные главы.

Загрузки

Программный код для запуска демонстрационного сайта под названием MovieGEEKs можно загрузить с сайта издательства по ссылке **www.manning.com/books/practical-recommendersystems**, а также с сервиса Github.com по ссылке **mng.bz/04K5**. Сайт создан с помощью фреймворка Django. У нас будет два набора данных: один генерируется автоматически, а второй нужно загрузить из базы MovieTweetings. Все инструкции по установке можно найти на сайте GitHub.

Формат кода

В книге содержится много примеров программного кода – как в виде пронумерованных листингов, так и в составе обычного текста. И в том, и в другом случае исходный код выделен шрифтом определенной ширины, чтобы его было легко отличить от обычного текста. Иногда код также выделен жирным шрифтом, чтобы акцентировать внимание читателя на измененных фрагментах, например при добавлении новой функции в строку кода, приведенную ранее в той же главе.

Во многих случаях мы внесли коррективы в первоначальный формат исходного кода, например добавили переносы строк и изменили отступы, чтобы максимально продуктивно использовать пространство книжной страницы. В редких случаях даже этого было недостаточно, и тогда мы вынуждены были применять символы, указывающие на продолжение строки (➥). Кроме того, из листингов часто приходилось убирать комментарии к исходному коду (когда код описывается в тексте). Многие листинги сопровождаются аннотациями к коду, в которых отражаются важные моменты.

Форум для читателей книги

При покупке «Рекомендательных систем на практике» читатель получает доступ к закрытому интернет-форуму издательства Manning Publications, где можно оставить комментарии о книге, задать технические вопросы и получить помощь от автора и других пользователей. Для входа на форум пройдите по ссылке **forums.manning.com/forums/practical-recommendersystems**. Подробнее познакомиться с форумами издательства Manning и правилами поведения на них можно на странице **forums.manning.com/forums/about**.

Издательство Manning стремится предоставить читателям площадку, где читатели могут общаться друг с другом или с автором книги. Издательство не берет на себя ответственности за обязательное привлечение на форум автора, чье участие в обсуждениях на форуме исключительно добровольное (и неоплачиваемое). Советуем задавать автору интересные вопросы, чтобы вовлечь его в общение! Форум и архивы старых обсуждений будут доступны на сайте издательства, пока книгу не снимут с печати.

Об авторе

Ким Фальк – специалист по обработке и анализу данных, обладающий большим опытом создания приложений, управляемых данными. Рекомендательные системы и машинное обучение в целом вызывают его живейший интерес. Он разрабатывал рекомендательные системы, предлагающие конечным пользователям фильмы и рекламу, и даже помогал юристам находить материалы по судебным прецедентам. Он занимается большими данными и машинным обучением с 2010 года. Ким часто говорит и пишет о рекомендательных системах. Его веб-страница: **kimfalk.org**.

Когда Ким не занят тем, что обучает машины следить за людьми, он отличный муж, отец и владелец курцхаара.

Об обложке

Иллюстрация на обложке «Рекомендательных систем на практике» – это работа под названием «Amazone d'Afrique», т. е. «Африканская амазонка». Этот рисунок Жака Грассе де Сен-Совера (1757–1810) является частью серии изображений национальных костюмов различных стран под названием «Costumes de Différents», опубликованной во Франции в 1797 году. Каждая работа тщательно прорисована и раскрашена вручную.

Богатое разнообразие рисунков в серии Жака Грассе де Сен-Совера напоминает нам о том, насколько сильно разнилась культура городов и регионов мира всего 200 лет назад. Люди, жившие изолированно друг от друга, разговаривали на разных диалектах и языках. На городских улицах или в сельской местности по одежде человека можно было легко определить, чем он занимается и где живет. С тех пор наша манера одеваться сильно изменилась, а региональное разнообразие, столь богатое в те времена, стерлось. Теперь трудно различить жителей разных континентов, а тем более разных стран, регионов и городов. Похоже, что мы пожертвовали культурным многообразием ради того, чтобы достичь большего разнообразия в личной жизни каждого отдельного человека, которая, безусловно, теперь более многогранна, динамична и технологична.

Во времена, когда трудно отличить одну компьютерную книгу от другой, издательство Manning привносит в компьютерную индустрию элемент новизны и свежести, создавая обложки, воспроизводящие богатое региональное разнообразие, которое было присуще миру два века назад и запечатлено на рисунках Жака Грассе де Сен-Совера.

ЧАСТЬ I

Подготовка к рекомендательным системам

Окружающая среда – это все, что не является мной.
Альберт Эйнштейн

Применение рекомендательных систем и, по сути, большинства методов машинного обучения в условиях промышленной эксплуатации подразумевает не только применения наиболее подходящего алгоритма, но и требует понимания пользователей и сферы деятельности.

Главы 1–6, т. е. часть I «Рекомендательных систем на практике» познакомит вас с экосистемой и инфраструктурой рекомендательных систем. Вы научитесь собирать данные и применять их, добавив рекомендательную систему в приложение. Вы узнаете, чем отличается рекомендация от рекламы и персональная рекомендация от неперсональной. Вы также узнаете, как собирать данные для создания собственной рекомендательной системы.

Глава 1

Что такое рекомендательная система?

Чтобы понять, что такое рекомендательная система, необходимо разобраться в огромном объеме информации, поэтому мы начнем эту книгу со следующих вопросов: какие проблемы решает рекомендательная система и как она применяется. **Вот о чем мы будем говорить:**

- выясним, какую задачу пытается выполнить рекомендательная система;
- разберемся, что такое персональные и неперсональные рекомендации;
- разработаем терминологию, позволяющую описывать рекомендательные системы;
- обсудим сайт-образец MovieGEEKs.

Налейте себе чашечку кофе, закутайтесь в одеяло и устройтесь поудобнее – сейчас вы познакомитесь с миром рекомендательных систем. Мы постараемся упростить задачу и сначала рассмотрим примеры из жизни, а только потом, в следующих главах, перейдем к математическим тонкостям рекомендательных систем. Возможно, вам захочется пропустить первую главу и читать дальше, но не следует этого делать. Чтобы получить представление о том, как должны выглядеть результаты вашей работы над рекомендательной системой, необходимо начать с самых азов.

1.1. Рекомендации в реальной жизни

Я много лет жил в Италии, в Риме. Рим – это красивый город, в котором множество продуктовых рынков – не тех, которые в центре и о которых рассказывается в путеводителях, где среди прочего торгуют поддельными сумками Gucci (да, Gucci на продуктовом рынке), а тех, которые остались за рамками маршрутов туристических автобусов, где закупаются местные жители и торгуют фермеры.

Каждое воскресенье мы покупали овощи и фрукты у продавца по имени Марино. Мы были хорошими покупателями, настоящими гурманами, поэтому он знал, что, если посоветует нам какие-то качественные продукты, мы их купим, хотя изначально собирались закупаться строго по списку. Восхитительные сезонные арбузы, множество разновидностей помидоров, даривших нам настоящий фейерверк вкусов, и потрясающая свежая моцарелла, которую я никогда не смогу забыть. Иногда Марино не советовал покупать нам продукты, качество которых было не на высоте, и мы доверяли его мнению. Это пример рекомендаций. Марино постоянно советовал нам одно и то же, что нормально, когда дело касается еды, но не вариант, когда речь идет о других вещах – например, книгах, фильмах или музыке.

Когда я был моложе, еще до того, как стриминговые сервисы типа Spotify захватили музыкальный рынок, мне нравилось покупать компакт-диски. Я шел в музыкальный магазин, который был ориентирован в первую очередь на диджеев, набирал огромную кипу компакт-дисков, находил у прилавка свободные наушники и начинал слушать. Я подолгу разговаривал с человеком за прилавком по поводу этих компакт-дисков. Он смотрел, какие диски мне понравились (и не понравились) и, исходя из этого, советовал другие. Я очень ценил то, что он запоминал мои предпочтения и не советовал мне одно и то же по нескольку раз. Это тоже пример рекомендаций.

Возвращаясь домой с работы (теперь, когда я старше), я всегда заглядываю в почтовый ящик, проверяя, нет ли писем. Как правило, ящик забит рекламой из супермаркетов, где перечислены товары по акции. Обычно в этих брошюрах на одной странице изображены свежие фрукты, а на другой – средство для посудомоечных машин, т. е. то, что супермаркеты рекомендуют вам купить, подчеркивая, что это выгодно. Это не рекомендации, а *реклама*.

Раз в неделю в почтовом ящике появляется местная газета. В газете публикуют список десяти самых популярных фильмов, которые показывают в кино на этой неделе. Это *неперсональные рекомендации*. На ТВ много внимания уделяется тому, чтобы вставить рекламный блок в подходящий материал. Это *таргетированная реклама*, поскольку считается, что ее смотрят люди определенного склада.

В феврале 2015 года сотрудники копенгагенского аэропорта заявили, что на территории аэропорта было установлено 600 мониторов, на которых, наряду с информацией о рейсе и воротах, транслируется реклама, подборка которой зависит от предполагаемого возраста и пола человека. Данные о возрасте и поле строились на основе полученного с камер изображения и специального алгоритма. В пресс-релизе об этом нововведении говорилось следующее: «Женщине, летящей в Брюссель, будет интересна реклама хороших часов или, например, финансового журнала. Семья, отправившаяся в отпуск, возможно, заинтересуется рекламой солнцезащитного крема или сервиса аренды машин»[1]. Это *релевантная реклама*, т. е. *максимально таргетированная*.

[1] Подробнее об этом читайте по ссылке **mng.bz/ka6j**.

Реклама по телевизору или в аэропорту обычно раздражает людей, но в сети границы того, что мы считаем назойливостью, несколько иные. Тому есть ряд причин, это само по себе является отдельной темой для обсуждения.

Интернет – это по-прежнему Дикий Запад, и, хотя я считаю, что реклама в аэропорту Копенгагена достаточно навязчива, меня не меньше раздражает реклама в интернете, если она предназначена для целевой группы, к которой я не отношусь. Чтобы показывать рекламу определенной целевой группе, веб-сайты должны немного представлять, кто вы такой.

В этой и следующих главах вы узнаете, что такое рекомендации, как собирать информацию о тех, кто будет эти рекомендации получать, как хранить данные и как их применять. Составлять рекомендации можно разными способами, и вы познакомитесь с наиболее распространенными приемами.

Рекомендательная система – это не только хитросплетенный алгоритм. В ее основе также лежит понимание данных и пользователей. Специалисты по анализу данных давно спорят о том, что важнее: наличие суперумного алгоритма или большого объема данных. Везде есть свои сложности. Суперумный алгоритм требует много супертехнологичного оборудования. Большой объем данных влечет за собой другие трудности, например как обеспечить быстрый доступ к этим данным. Читая эту книгу, вы узнаете, где можно пойти на компромисс, и научитесь принимать удачные решения.

Вышеописанные примеры призваны показать, что пользователь может не видеть разницы между рекламой и рекомендациями. Но на самом деле разница в предназначении: *рекомендация* выдается на основе: вкусов пользователя (если он достаточно активен), вкусов других пользователей и того, что чаще всего ищет этот человек. Реклама работает только на благо рекламодателя и обычно навязывается получателю. Различие может быть очень размытым. В этой книге я называю рекомендацией все то, что строится на основании полученных данных.

1.1.1. Рекомендательные системы дома в интернете

Рекомендательные системы чаще всего предполагают домашнее использование через интернет, поскольку так можно не только охватить отдельных пользователей, но и получить данные об их поведении. **Давайте рассмотрим несколько примеров.**

Сайт со списком 10 самых популярных хлебопечек предоставляет *неперсональные* рекомендации. Если сайт с товарами для дома или с билетами на концерты предлагает вам рекомендации, ориентируясь на демографические данные или на ваше текущее местоположение, это *полуперсональные* рекомендации. Персональные рекомендации можно увидеть, например, на сайте Amazon, где зарегистрированные пользователи видят рекомендации в специальном разделе. Потребность в персональных рекомендациях связана с тем, что людям интересны не только популярные товары, но и те, которые не входят в число наиболее продаваемых, а также те, которые находятся в длинном хвосте.

1.1.2. Длинный хвост

Понятие *длинного хвоста* ввел Крис Андерсон в своей статье для журнала Wired в 2004 году, а в 2006 году на основе этой статьи была написана книга[1]. В статье Андерсон описал новую бизнес-модель, которая часто встречается в интернете. Основная идея Андерсона заключалась в том, что, если у вас магазин не в интернете, у вас ограниченное пространство для хранения и, что еще важнее, ограненное пространство для демонстрации товаров покупателям. Также у вас ограничен круг покупателей, поскольку людям приходится идти в магазин. Если бы не было этих ограничений, вам не пришлось бы продавать только популярные товары, как часто бывает в традиционной торговой бизнес-модели. В офлайн-магазинах держать на складе непопулярные товары считается проигрышной стратегией, поскольку необходимо место для хранения большого количества товаров, которые, возможно, никто не купит. Но, если у вас интернет-магазин, в вашем распоряжении место для бесконечного количества товаров, поскольку плата за него невысока или, если вы продаете цифровой контент, складское пространство вам вообще не требуется, т. е. плата минимальна или равна нулю. Суть экономики «длинного хвоста» сводится к тому, что можно получать доход, продавая много разных товаров, каждый из них – лишь по нескольку экземпляров множеству разных людей.

Я всеми руками за разнообразие, поэтому мне кажется, что огромный каталог товаров – это прекрасно, но вопрос, на который трудно дать ответ, заключается в том, как именно пользователи находят то, что им нужно? Именно тут на сцену выходят рекомендательные системы. Именно такие системы помогают людям находить различные вещи, о существовании которых те даже не догадывались.

В сети титанами в отношении контента и рекомендаций считаются сервисы Amazon и Netflix, поэтому именно они фигурируют в различных примерах в данной книге. В следующем разделе мы рассмотрим сервис Netflix как наглядный пример рекомендательной системы.

1.1.3. Рекомендательная система Netflix

Как вы, вероятно, знаете, Netflix – это стриминговый сервис. Его основной контент – это фильмы и сериалы, подборка которых постоянно обновляется. Задача, которую выполняют рекомендации Netflix, состоит в том, чтобы поддерживать ваш интерес к представленному контенту как можно дольше, стимулируя вас оплачивать подписку из месяца в месяц.

Сервис работает на различных платформах, поэтому рекомендации могут выглядеть по-разному. На рис. 1.1 показан снимок экрана моего ноутбука с открытым сервисом Netflix. Также я могу запустить Netflix на телевизоре, планшете и даже телефоне. На разных устройствах хочется смотреть разное – я никогда не смотрю эпичные фильмы жанра фэнтези на телефоне, но мне нравится смотреть их на телевизоре.

---

[1] Более подробную информацию см. на странице **www.wired.com/2004/10/tail/**. Информацию о книге см. на странице **en.wikipedia.org/wiki/The_Long_Tail_(book)**.

Рис. 1.1. Главная страница сервиса Netflix (до того, как сменился макет)

Начнем нашу ознакомительную экскурсию с главной страницы. Стартовая страница представляет собой панель с названиями категорий, например **Top Picks** (Самое интересное), **Drama** (Драмы) и **Popular on Netflix** (Популярное на Netflix). Верхняя категория – это мои просмотры. Netflix уделяет большое внимание этой категории, поскольку она показывает не только то, что я посмотрел и что смотрю сейчас, но и то, что меня (хотя бы в какой-то степени) заинтересовало.

Netflix стремится обратить ваше внимание на категорию, расположенную строчкой ниже, – **Netflix Originals** – поскольку в нее входят сериалы производства Netflix. Они важны для Netflix по двум причинам, обе из которых связаны с финансами:

- Netflix тратит огромные деньги на производство собственного контента, который преимущественно можно посмотреть только через сервис Netflix;

- Netflix платит деньги владельцам контента, когда пользователи смотрят этот контент. Если владелец сам Netflix, сервис не просто экономит, а зарабатывает.

Последний пункт – это отдельная пища для размышлений: даже если на странице все персонализировано, категория Netflix Originals все равно идет второй строкой, и сам этот факт, как правило, не свидетельствует о том, что я смотрю этот контент, а демонстрирует стоящие перед компанией бизнес-задачи.

Подборки и тренды

Ниже расположена категория **Trending Now** (Сейчас в тренде). Тренд – это достаточно размытое понятие, под которым многое может иметься в виду, но в данном случае речь идет о контенте, который пользовался популярностью в самое недавнее время. Далее следует категория **Popular on Netflix** (Популярное на Netflix), которая также включает в себя популярный контент, но период времени, в течение которого этот контент пользуется популярностью, дольше – например, около недели. Мы подробно поговорим о трендах и подборках в главе 5.

Рекомендации

Четвертая категория – это персональная подборка **Top Picks** (Лучшее), которая отражает мои предпочтения. Сюда входит все то, что большинство людей назвали бы рекомендациями. Здесь показан контент, который, по прогнозам рекомендательной системы Netflix, я захочу посмотреть в ближайшее время. Часто эти прогнозы верны. Я не фанат кровавых фильмов со сценами жестокости и предпочту не видеть физического насилия ни в каком варианте. Не все предложенные варианты мне по вкусу, но я думаю, что Netflix составляет рекомендации на основе не только моих предпочтений. Другие члены семьи тоже иногда смотрят фильмы и сериалы через мой профиль. *Профили* позволяют сервису Netflix идентифицировать человека, который пользуется данной учетной записью в настоящий момент.

До появления профилей сервис Netflix составлял рекомендации для всей семьи, а не для отдельных пользователей[1]. Он пытался всегда подобрать что-то для мамы, папы и детей. Однако теперь Netflix больше так не делает, поэтому в моем списке отсутствуют детские передачи. Но, даже несмотря на то что сейчас сервис Netflix позволяет создать персональный профиль для каждого пользователя, я считаю, что система обязательно должна принимать во внимание всех зрителей – не только того человека, которому принадлежит профиль, но и остальных. До меня дошли слухи, что другие компании пытаются разработать решение, с помощью которого можно будет проинформировать систему о наличии других зрителей, помимо вас. Так сервис получит возможность предлагать рекомендации, ориентированные на всех присутствующих. На практике пока что я такого ни разу не встречал.

Технология Kinect компании Microsoft была способна идентифицировать сидящих перед телевизором людей с помощью функции распознавания лиц и движений. Разработчики Microsoft пошли еще дальше, и система могла уз-

[1] Netflix Recommendations: Beyond the 5 stars (Part 1), **mng.bz/bG2x**.

навать не только членов семьи, но и других людей из полного каталога пользователей, т. е. она идентифицировала пользователей, находящихся в гостях у других пользователей. Несмотря на этот шаг в сторону распознавания аудитории, сенсор Kinect для Xbox One был снят с производства в октябре 2017 года, что ознаменовало закрытие линейки Kinect.

Категории и разделы

Вернемся к категории **Top Picks** (Лучшее) на Netflix. Если навести курсор на один из предложенных фильмов или сериалов, появится более подробная информация о контенте. Наряду с всплывающим описанием (рис. 1.2) отобразится предполагаемая оценка, которую я, по прогнозу рекомендательной системы, скорее всего, дам этому контенту. Логично предположить, что все видео из категории **Top Picks** будут с высокими оценками, как на рис. 1.1. Однако, просматривая рекомендации, можно встретить варианты с прогнозом на низкие оценки, как на рис. 1.3.

Netflix формирует рекомендации множеством разных способов, и можно найти множество возможных объяснений тому, почему Netflix рекомендует контент, который, по собственным же прогнозам сервиса, получит от пользователя невысокую оценку. Одна из причин, возможно, заключается в том, что Netflix ставит разнообразие превыше точности попадания. Другая вероятная причина заключается в том, что, пусть я и не поставлю фильму максимальное количество звезд, вдруг именно этот фильм сейчас подходит под мое настроение. Кроме того, это первый признак, что Netflix не придает этим оценкам большого значения.

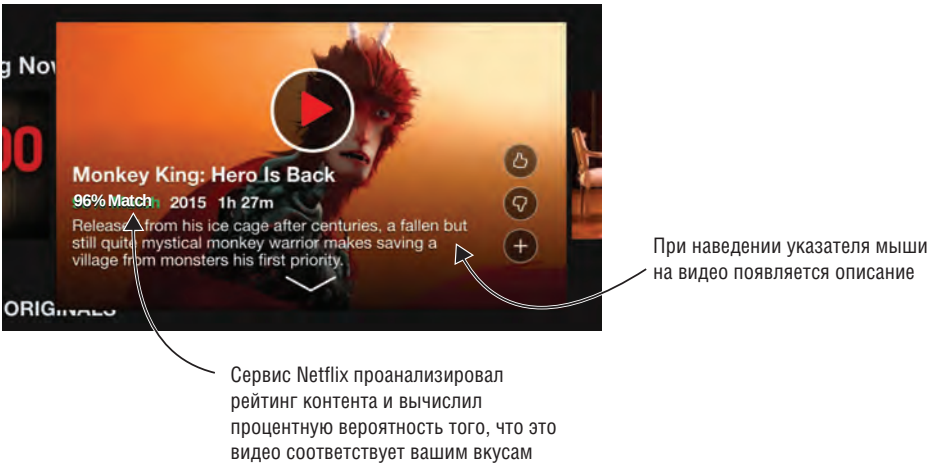

Рис. 1.2. Мультфильм из категории Top Picks на Netflix и прогнозируемое соответствие вкусам пользователя

У всех разделов разные названия. Некоторые формируются по принципу «Потому что вы посмотрели “Форс-мажоры”». Туда включены фильмы и сериалы, похожие на «Форс-мажоры». Другие разделы озаглавлены в зависимости от представленных жанров: например, раздел **Comedies** (Комедии), как ни удивительно,

включает в себя комедии. В принципе, названия разделов – это тоже в какой-то степени рекомендации, т. е. эти *категории* также можно считать *рекомендациями*.

Тут можно было бы и остановиться, но тогда мы бы пропустили самый важный аспект персонализации на Netflix.

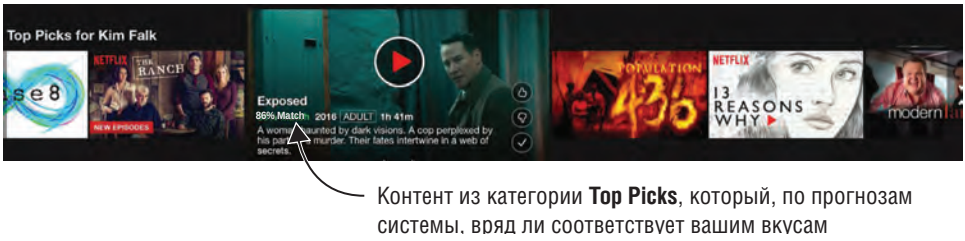

Рис. 1.3. Фильм из категории Top Picks на Netflix с прогнозируемо низкой оценкой.

Оценки

Заголовок каждой категории описывает данную подборку контента. Видео в подборке расположено в соответствии с алгоритмом рекомендательной системы – в порядке релевантности или по оценке, слева направо, как показано на рис. 1.4.

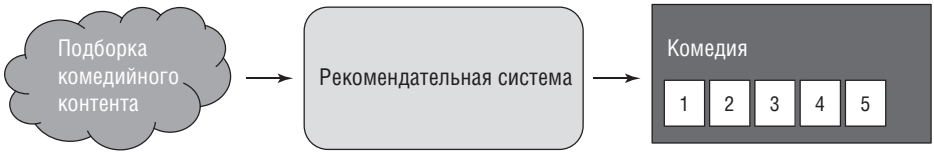

Рис. 1.4. На Netflix контент в каждой строке расположен по релевантности.

Даже в категории **My List** (Мой список), куда входит тот контент, который я выбрал самостоятельно, видео располагается в порядке, предусмотренном рекомендательной системой, – в соответствии с предполагаемой релевантностью моим предпочтениям. На рис. 1.1 представлен вчерашний снимок экрана. Сегодня видео в моем списке уже расположено по-другому, как показано на рис. 1.5.

Рис. 1.5. Netflix выстраивает видео в моем списке в порядке релевантности.

Рекомендательная система Netflix также пытается предлагать контент, который будет актуален в определенное время или при определенных условиях. На-

пример, воскресное утро больше подходит для просмотра мультфильмов и комедий, а вечер – для просмотра «серьезных» сериалов, таких как «Форс-мажоры».

Еще одна категория, содержимое которой может вас удивить, это – **Popular on Netflix** (Популярное на Netflix), куда входит контент, пользующийся популярностью в данный момент. Однако крайнее слева видео в этой категории вовсе не обязательно самое популярное. Netflix находит несколько самых популярных вещей, а затем располагает их в таком порядке, который, с точки зрения рекомендательной системы, в наибольшей степени отвечает вашим предпочтениям.

Продвижение

Любопытный вопрос – почему Netflix поставил на одно из первых мест в категории **My List** сериал «Последний кандидат», учитывая, что я уже и так его смотрю. Netflix добавил пометку, что вышел новый сезон «Последнего кандидата». Возможно, это и есть причина появления этого сериала в категории **My List**.

Продвижение – это один из способов склонить чашу весов в ту или иную сторону при составлении рекомендаций. Например, Netflix хочет, чтобы я обратил внимание на сериал «Форс-мажоры», потому что это новый контент, а значит, его ценность выше. Netflix продвигает новый контент. Под *новизной* можно понимать, что этот контент только появился или мелькнул в новостях. В главе 6 мы подробно поговорим о продвижении, поскольку именно продвижение начинает интересовать многих владельцев сайтов, как только система отлажена и запущена.

> **ПРИМЕЧАНИЕ.** Существует еще понятие бустинга (продвижения), относящееся к алгоритмам машинного обучения. Но я имею в виду не его[1].

Синхронизация с социальными сетями

На протяжении короткого периода времени сервис Netflix также пытался применять данные из социальных сетей[2]. Тогда на главной странице Netflix можно было увидеть что-то вроде того, что показано на рис. 1.6.

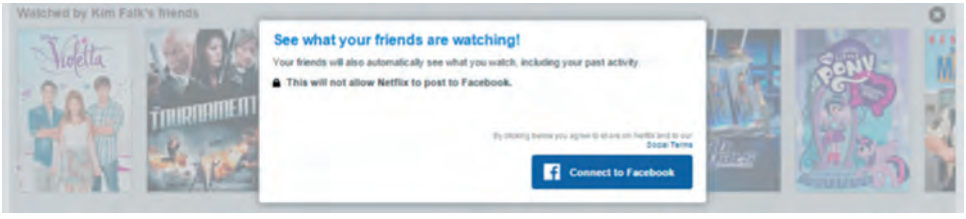

Рис. 1.6. Netflix хочет знать, что смотрят мои друзья

Netflix предлагал привязать к учетной записи свой аккаунт на Facebook, предоставив себе доступ к списку ваших друзей, а также к другой информации. Одним из преимуществ этой привязки для Netflix было то, что так систе-

[1] См. en.wikipedia.org/wiki/Boosting_(machine_learning)

[2] См. **mng.bz/6yHM.**

ма могла изучать ваших друзей и составлять общие рекомендации, опираясь на их предпочтения. Привязка к Facebook также могла превратить просмотр фильмов в совместное мероприятие – очень популярное направление, которое развивают многие медиакомпании.

В наши дни люди не сидят и не смотрят фильмы просто так. Они многозадачны: смотрят кино и одновременно зависают в другом устройстве (например, планшете или смартфоне). От того, чем вы занимаетесь на втором устройстве, может во многом зависеть, что вы будете смотреть дальше. Представьте себе, что, посмотрев что-то на Netflix, вы видите на телефоне уведомление о том, что кому-то из ваших друзей понравился какой-нибудь фильм, и вуаля – Netflix рекомендует вам посмотреть этот фильм следующим.

И все-таки этот функционал был отключен году в 2015 или 2016 по той причине, что люди не хотели делиться информацией о своих фильмах с контактами на Facebook. Как сказал Нил Хант, директор по контенту Netflix: «Это очень печально, поскольку я считаю, что совместив рекомендации по алгоритму и персональные рекомендации, можно получить ценный ресурс»[1].

Профиль предпочтений

Если страница практически полностью строится на основе ваших предпочтений, неплохо бы предоставить системе как можно больше информации о своих вкусах. Если у Netflix не будет ясного понимания, что вам нравится, вероятно, вам будет трудно найти что-нибудь интересное для себя.

В 2016 году сервис Netflix предоставлял пользователям возможность настроить профиль. Меню **Taste Profile** (Профиль предпочтений), показанное на рис. 1.7, позволяло оценивать сериалы и фильмы, выбирать жанры, указывая, как часто вам хочется посмотреть что-нибудь остросюжетное (Netflix называет этот тип фильмов **Adrenaline Rush** (Адреналиновая лихорадка), см. рис. 1.8) или проверить оценки, которые вы выставили ранее, и убедиться, что ваше мнение не изменилось.

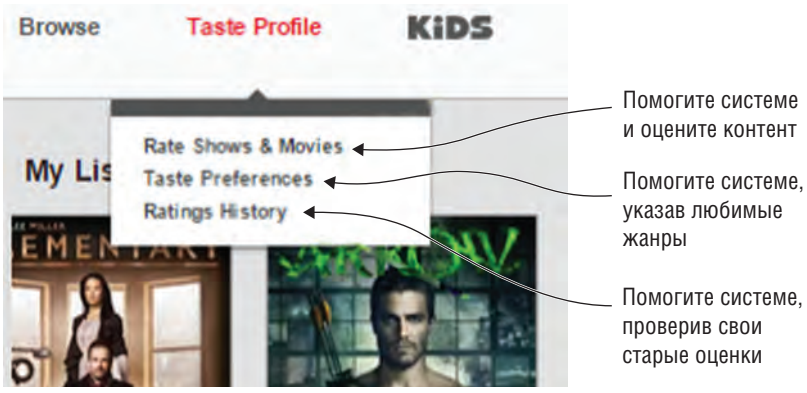

Рис. 1.7. Пример того, как выглядели профили предпочтений Netflix в 2015 году

[1] См. **mng.bz/jc7M**.

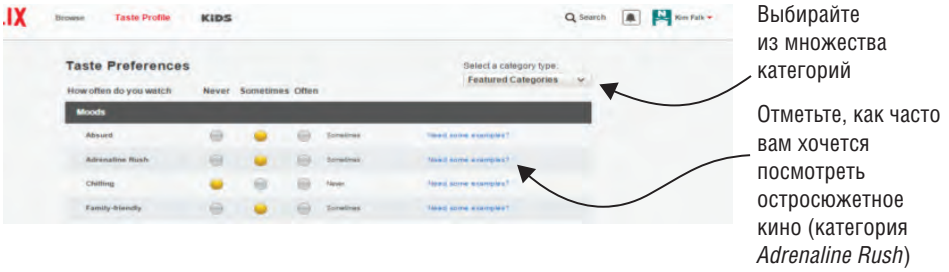

Рис. 1.8. Меню профиля предпочтений Netflix

Благодаря этим настройкам, рекомендации Netflix стали точнее. Обратиться к пользователю за помощью при составлении профиля предпочтений – это достаточно распространенный прием, к которому прибегают системы при формировании рекомендаций для новых пользователей. Но, как часто бывает, между тем, что пользователям нравится, по их словам, и тем, что им нравится на самом деле, существует большая разница.

Профиль предпочтений – это первая ступенька на пути к знакомству с пользователем. Чем дольше пользователь взаимодействует с системой, тем больше данных о нем собирает Netflix, и, как правило, на эти данные можно полагаться целиком и полностью. К настоящему моменту Netflix отказался от заполняемого пользователями профиля предпочтений.

1.1.4. Определение рекомендательной системы

Чтобы убедиться, что мы правильно понимаем друг друга, посмотрите табл. 1.1, где указаны определения основных понятий.

Таблица 1.1. Рекомендательные системы: термины и понятия

Термин	Пример из Netflix	Определение
Прогноз	Netflix угадывает, какую оценку вы поставите контенту	Прогноз – это предположение относительно того, насколько пользователю понравится контент
Релевантность	Расположение категорий контента на странице (например, **Top Picks** (Лучшее) и **Popular on Netflix** (Популярное на Netflix)) по степени интересности	Расположение контента в соответствии с тем, что больше всего подходит пользователю в данный момент. Релевантность сочетает в себе контекст, демографические данные и (ожидаемые) оценки
Рекомендация	**Top Picks** (Лучшее) для меня	Лидеры по релевантности
Персонализация	Заголовки категорий на странице – это пример персонализации	Сочетание релевантности и наглядности
Профиль предпочтений	См. рис. 1.8.	Список характеристик и их значений

Разобравшись с этими терминами, мы, наконец, можем дать определение рекомендательной системе.

Определение: рекомендательная система

Рекомендательная система подбирает и предлагает пользователю релевантный контент, основываясь на своих знаниях о пользователе, контенте и взаимодействии пользователя и контента.

Прочитав это определение, вы, возможно, решили, что вам все понятно. Но давайте рассмотрим пример того, как подбираются рекомендации и как работает система. На рис. 1.9 показано, каким образом сервис Netflix мог сформировать для меня подборку **Top Picks** (Лучшее). Вот как мог выглядеть процесс подбора контента в данном случае (по шагам):

1. Получен запрос на подборку **Top Picks** (Лучшее).
2. Сервер обращается к рекомендательной системе, в этом участвует целый ряд методов. Этот шаг называется *получить варианты объектов*. С его помощью сервер из базы данных каталога получает объекты, которые в наибольшей степени релевантны предпочтениям пользователя в данный момент.
3. Пять лидирующих объектов (в обычной ситуации этих объектов может быть 100 или больше) отправляются на следующий этап – анализ и прогнозирование.
4. Прогноз составляется на основе предпочтений пользователя, информация о которых берется из базы данных о пользователе. При этом, скорее всего, один или несколько объектов будут исключены из списка вследствие прогнозируемо низкой оценки. На рис. 1.9 исключены объекты C и E.
5. В результате анализа и прогноза остаются наиболее значимые объекты, которые теперь сопровождаются прогнозируемыми оценками. Они переходят на следующий этап – выстраивание в порядке релевантности.
6. Релевантные объекты располагаются в соответствии с предпочтениями пользователя, контекстом и демографическими данными. Этот процесс может даже предполагать максимальное расширение разнообразия подборки.
7. Теперь объекты расставлены по релевантности. Объект F был исключен, поскольку анализ релевантности показал, что данный контент не релевантен интересам конечного пользователя.
8. Система выдает список.
9. Сервер выдает результат.

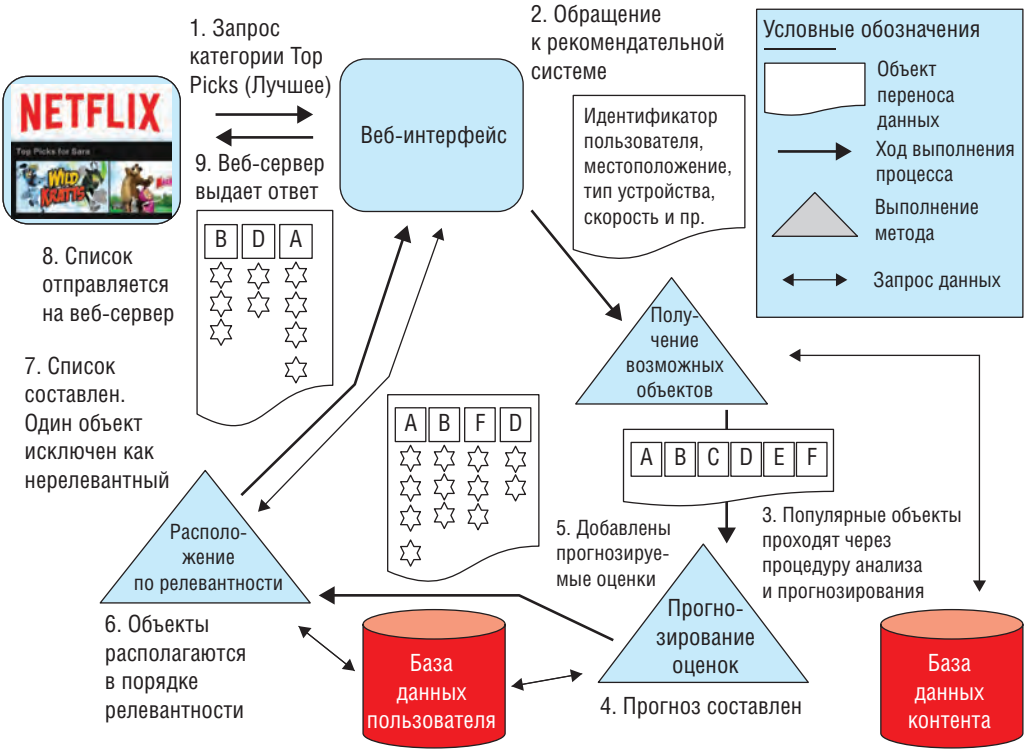

Рис. 1.9. Как может выглядеть процесс отбора контента в категорию Top Picks (Лучшее) на Netflix

Изучив рис. 1.9, легко понять, что при работе с рекомендательными системами необходимо учитывать множество аспектов. В этом описании отсутствуют этапы сбора данных и построения моделей. Большинство рекомендательных систем пытаются тем или иным образом применять данные, показанные на рис. 1.10.

Кроме того, рис. 1.9 иллюстрирует еще один факт, который необходимо учитывать: прогнозирование оценок/рейтингов – это всего лишь одна из множества задач рекомендательной системы. Другие факторы также могут оказывать существенное влияние на то, что именно система будет показывать пользователю. В данной книге очень много внимания уделяется прогнозированию оценок, и это важный аспект, даже если вам показалось, что я не придаю ему большого значения.

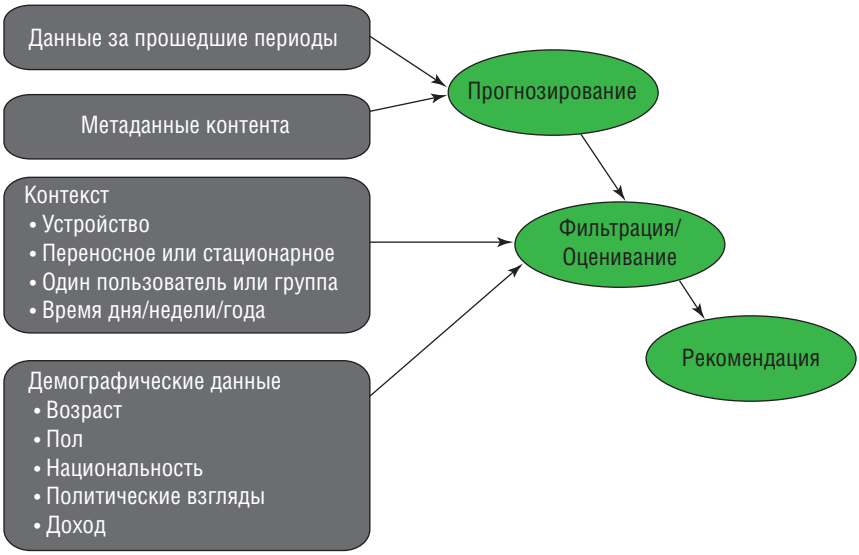

Рис. 1.10. Данные, на которые может опираться рекомендательная система

1.2. Таксономия рекомендательных систем

Прежде чем начать разработку рекомендательной системы, хорошо бы немного поразмыслить, какую именно рекомендательную систему вы хотите получить на выходе. Отличной отправной точкой будет изучение аналогичных систем. В этом разделе вы познакомитесь с основными понятиями, необходимыми для понимания рекомендательных систем.

В предыдущем разделе мы совершили обзорную экскурсию по сервису Netflix и получили общее представление о возможностях рекомендательной системы. В данном разделе мы поговорим об основных характеристиках и классификации рекомендательных систем. Эту информацию я изначально почерпнул из курса профессора Джозефа А. Констана (Joseph A. Konstan) и Майкла Д. Экстранда (Michael D. Ekstrand) под названием «Introduction to Recommender Systems» («Введение в рекомендательные системы»)[1], и она мне много раз пригождалась. *Таксономия* позволяет описать систему по следующим параметрам: специализация, задача, контекст, степень персонализации, чьи мнения, конфиденциальность и надежность, интерфейс и алгоритмы[2]. Давайте рассмотрим каждый из этих параметров.

---

[1] Подробнее об этом вы найдете по адресу **www.coursera.org/learn/recommender-systems-introduction/**.

[2] Этот вариант таксономии впервые был обозначен в книге Джона Риэля (John Riel) и Джозефа А. Констана (Joseph A. Konstan) «Word of Mouse: The Marketing Power of Collaborative Filtering» (Business Plus, 2002).

1.2.1. Специализация

Специализация – это тип рекомендуемого контента. В случае с Netflix специализация – это фильмы и телесериалы, но специализация может быть абсолютно любой: подборки контента в виде плейлистов, советы по применению цифровых обучающих курсов для достижения своих целей, работа, книги, машины, продукты, отдых, путешествия или даже знакомства.

Специализация играет важную роль, поскольку служит ориентиром при организации работы с рекомендациями. Специализация также важна, поскольку позволяет оценить, насколько плохими будут последствия ошибок. Если у вас музыкальная рекомендательная система, ничего страшного не произойдет, если вы порекомендуете неудачную музыку. Если вы рекомендуете опекунскую семью для ребенка в трудной жизненной ситуации, цена ошибки будет очень высока. Кроме того, от специализации зависит, можно ли рекомендовать одно и то же по нескольку раз.

1.2.2. Задача

Какова задача сайта Netflix – как с точки зрения конечного пользователя, так и с точки зрения провайдера? Конечным пользователям рекомендации Netflix нужны, чтобы найти подходящий контент, который будет интересно посмотреть в определенное время. Представьте себе, что весь контент предлагается без какой-либо фильтрации или упорядочивания. Как в таком случае можно было бы найти хоть что-нибудь в каталоге Netflix, если в нем более 10 000 наименований? А задача провайдера (в данном случае Netflix) – подтолкнуть пользователей месяц за месяцем платить за подписку, предлагая контент, который человек захочет посмотреть, избавляя пользователей от лишних телодвижений.

Для Netflix главным индикатором успешности системы является объем просмотренного контента. Когда мы оцениваем степень достижения цели, которая не является нашей непосредственной целью, мы говорим о вспомогательной цели. Ставя перед собой вспомогательную цель, необходимо соблюдать осторожность, поскольку все может закончиться тем, что вы начнете следить вовсе не за теми параметрами, которые вам были важны изначально. Если человек много времени проводит на Netflix, возможно, он просто растерян, поскольку ищет, ищет и все никак не может найти то, что ему нужно. А может, он нашел, что искал, но сайт «затормозил»[1].

Возможно также, что подспудно Netflix стремится сделать все так, чтобы за просмотренный вами контент сервису пришлось платить как можно меньше. Вероятно, Netflix меньше платит за первые сезоны сериала «Друзья», чем за более свежие серии. А еще лучше, если вы будете смотреть сериалы производства Netflix, – в этом случае сервис не платит никаких отчислений вообще.

---

[1] Если вам интересно, что может пойти не так, когда вы ориентируетесь на вспомогательные цели, прочитайте книгу Кэти О'Нил (Cathy O'Neil) «Weapons of Math Destruction» (Broadway Books, 2016).

Задача также может заключаться в том, чтобы предоставить информацию, оказать помощь или просветить человека в каком-либо вопросе. Однако чаще всего задача – совершить как можно больше продаж.

На каких пользователей вы ориентированы в первую очередь: тех, которые обратятся к вам лишь однажды и будут ждать от вас хороших рекомендаций, или тех, которые зарегистрируются и будут заходить регулярно? Будет ли сайт загружать контент автоматически (как, например, радио на сервисе Spotify, где нон-стоп проигрывается музыка, выбор которой зависит от первой песни или исполнителя)?

1.2.3. Контекст

Контекст – это условия, в которых пользователь получает рекомендацию. В нашем примере это может быть устройство, с которого человек заходит на Netflix, или текущее местонахождение, время дня (или ночи), а также то, чем человек занят. Есть ли у пользователя время изучить предложенные рекомендации, или ему нужно быстро принять решение? Контекст может также включать в себя погоду и даже настроение пользователя!

Возьмем, например, поиск кафе через Google Карты. Сидит ли человек в офисе за компьютером и ищет хорошую кофейню или стоит на улице, когда начинается дождь? В первом случае лучшей выдачей будет список хороших кофеен на достаточно обширной территории, во втором – рекомендация в идеале будет содержать только ближайшее место, где можно выпить кофе, пережидая дождь. Примером приложения, с помощью которого можно найти кафе, может быть Foursquare. О приложении Foursquare мы подробнее поговорим в главе 12.

1.2.4. Степень персонализации

Степень персонализации рекомендаций бывает самой разной, от применения наиболее обобщенных данных до изучения информации о конкретном пользователе. Эти уровни отображены на рис. 1.11.

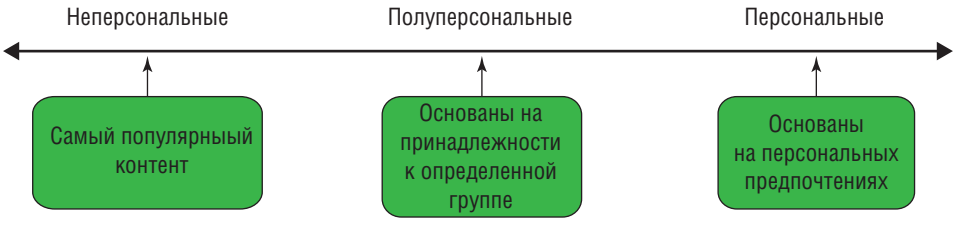

Рис. 1.11. Степени персонализации

Неперсональные

Список наиболее популярных объектов считается *неперсональной рекомендацией*: расчет здесь на то, что данному пользователю понравятся те же объекты, что и большинству других. К неперсональным рекомендациям также

относится выстраивание объектов в списке по дате их появления, например когда самыми первыми отображаются наиболее новые объекты. Любой человек, который взаимодействует с рекомендательной системой, видит те же рекомендации, что и остальные люди. Сюда же можно отнести спецпредложения в кафе, когда посетителям предлагаются алкогольные напитки в пятницу вечером, капучино по утрам и поздний завтрак утром в выходные.

Полуперсональные и частично персональные

Рекомендации следующего уровня предполагают деление пользователей по группам – это *полуперсональные* и *частично персональные рекомендации*. Пользователей можно разбить на группы по множеству разных признаков: по возрасту, по национальности или по характерному признаку – например, бизнесмены и студенты, водители автомобилей и велосипедисты.

В качестве примера: система, предназначенная для продажи билетов на мероприятия, рекомендует шоу-программы, основываясь на стране или городе пребывания пользователя. Другой пример: если пользователь слушает музыку на смартфоне, система может попытаться определить, перемещается это устройство или нет. Если да, то, возможно, человек занимается спортом или едет на машине или велосипеде. Если устройство неподвижно, пользователь, вероятно, сидит на диване дома, и ему подойдет несколько иная музыка.

Подобная рекомендательная система не знает ничего лично о вас, для нее вы являетесь частью определенной группы или сегмента. Другие люди, входящие в эту группу, получат те же рекомендации, что и вы.

Персональные

Персональные рекомендации базируются на данных о конкретном пользователе, а именно на информации о том, каким образом пользователь взаимодействовал с системой ранее. Так формируются рекомендации специально для данного пользователя.

Большинство рекомендательных систем при составлении персональных рекомендаций также учитывают принадлежность пользователя к группе и популярность контента. Пример персональных рекомендаций можно увидеть на сайте Amazon, где в личном кабинете пользователя присутствует раздел **Recommended for You** (Рекомендуется вам). Главная страница сервиса Netflix – это ярчайший пример персональных рекомендаций.

Обычно сайты комбинируют различные типы рекомендаций. Лишь несколько сайтов, включая Netflix, предлагают только персональные рекомендации. На Amazon также мы увидим раздел **Most Sold Items** (Самые продаваемые товары), где отсутствует персонализация, а также раздел **Customers Who Bought This Also Bought This** (Вместе с этим покупают эти товары), т. е. *выборочные рекомендации*. Эти рекомендации предлагаются выборочно – например, тем людям, которые смотрят данный товар.

1.2.5. Чье мнение

Экспертные рекомендательные системы формируют блок рекомендаций на основе ассоциативных правил, составленных вручную специалистами. Эти системы рекомендуют хорошие вина, книги и подобные вещи и применяются в областях, где, как правило, необходимо быть экспертом, чтобы давать советы.

Однако дни экспертных систем практически на исходе, поэтому сейчас параметр «*чьи мнения*» утратил свою актуальность. Почти все сайты опираются на мнение большинства. Говорят, что у каждого правила есть исключения: несколько экспертных сайтов все же продолжают работать. Один из них – специализирующийся на вине сайт **www.vivino.com** (рис. 1.12), на котором представлены рекомендации сомелье. Vivino также применяет рекомендательную систему по винам. В 2017 году в приложение Vivino была добавлена рекомендательная система, которая помогает пользователю выбрать новое, соответствующее его вкусам вино, опираясь на предыдущие оценки этого пользователя[1].

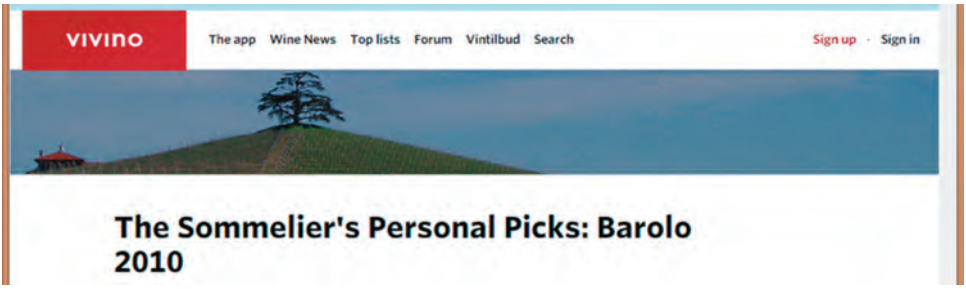

Рис. 1.12. Сайт **Vivino.com** предоставляет пользователям экспертные рекомендации по винам (сами рекомендации вырезаны для экономии пространства)

1.2.6. Конфиденциальность и надежность

Насколько хорошо система защищает данные о пользователе? Каким образом применяется полученная информация? Например, в Европе принято перечислять деньги на пенсионный счет, который находится в ведении банка. Часто эти банки предлагают различные накопительные пенсионные программы. Система, специализирующаяся на таких рекомендациях, должна очень строго следить за конфиденциальностью. Представьте себе, что вы заполняете заявление на участие в накопительной пенсионной программе, где указываете, что у вас проблемы со спиной, а через минуту после этого вам звонит мануальный терапевт с отличным предложением специально для решения вашей проблемы. Или еще хуже – вы покупаете специальную кровать для людей с больной спиной, а через час получаете электронное письмо, в котором говорится, что стоимость вашей медицинской страховки выросла.

[1] Подробнее читайте по ссылке **mng.bz/1jFR**.

Многие люди считают рекомендации своеобразной разновидностью манипуляций, поскольку с их помощью людям предлагают такие варианты товаров и услуг, на которые те с большей вероятностью согласятся, чем на предложения из случайной подборки. А большинство магазинов стремятся увеличить объем продаж. Тот факт, что благодаря рекомендациям магазины продают больше товаров и услуг, вызывает у людей ощущение, что ими манипулируют. Но если при этом просмотр фильма принесет удовольствие, а не навеет скуку, то я не возражаю. Манипуляция – это не столько сам факт того, что вам предлагают определенный объекта, сколько *мотив*, для чего это *делается*. Если вам порекомендовали неподходящее и не самое оптимальное лекарство лишь на том основании, что его продавец предоставляет владельцу сайта лучшие условия сотрудничества, то это манипуляция, и это заслуживает порицания.

Когда рекомендательная система запущена и бизнес активно развивается, у многих может возникнуть соблазн продвинуть интересы поставщиков, быстрее сбыть залежалый товар или по каким-либо иным причинам подтолкнуть пользователей к покупке определенной торговой марки таблеток. Учтите: если пользователи почувствуют, что ими манипулируют, они перестанут доверять вашим рекомендациям и в итоге найдут то, что им нужно, где-нибудь в другом месте.

> *В тот момент, когда рекомендации обладают властью влиять на решение, они становятся мишенью для спамеров, мошенников и других лиц, желающих повлиять на наши решения из далеко не лучших побуждений.*
>
> Дэниел Тункеланг[1]

Надежность – это показатель того, насколько пользователь доверяет рекомендациям, в противовес тому, чтобы расценивать их как рекламу или попытки манипулирования. Говоря о Netflix, я рассказывал, что прогнозы могут отталкивать пользователей, если прогнозируемые оценки пользователя оказываются очень далекими от реальных. Все это – вопрос надежности. Если пользователь прислушивается к рекомендациям, система надежна, ей доверяют.

1.2.7. Интерфейс

Интерфейс рекомендательной системы отображает тип ввода и вывода данной системы. Давайте рассмотрим каждый из них.

Ввод

Пользователи сервиса Netflix с некоторых пор могут указывать, нравится им контент или не нравится, выставляя оценки и обозначая предпочитаемые жанры и темы. Эти данные могут служить в качестве ввода для рекомендательной системы.

---

[1] Чтобы больше узнать о том, какую роль играют доверие и вкус в рекомендациях, см. **www.linkedin.com/pulse/taste-trust-daniel-tunkelang**.

В примере с Netflix мы говорили о *явном вводе*, при котором вы, пользователь, вручную указываете информацию о том, что вам нравится. Другой вид ввода – *неявный*, когда система пытается установить ваши вкусы, опираясь на то, как вы с ней взаимодействуете. В главе 4 мы подробнее рассмотрим обратную связь.

Вывод

К разновидностям *вывода* относятся прогнозы, рекомендации или фильтрация. Например, Netflix выдает рекомендации разными путями. Сервис прогнозирует оценки, составляет персональный набор предложений и отображает популярный контент, как правило, в виде списка 10 лидирующих объектов (но даже этот список Netflix формирует персонально для данного пользователя).

Если рекомендации естественным образом вписаны в страницу, это *органичная подача*. Расположенные рядами категории контента в сервисе Netflix – это пример органичных рекомендаций: Netflix не указывает, что это рекомендации. Это обычная часть сайта.

На рис. 1.13 показаны неорганичные результаты. То, что мы видим на сайте Hot Network Questions, – это разновидность неперсонализированных рекомендаций, поскольку все приведенные утверждения воспринимаются пользователем так, будто они сделаны от имени сайта. Amazon предлагает неорганичные персонализированные рекомендации в разделе **Recommended for You** (Рекомендуется вам), а New York Times применяет неорганичные рекомендации, когда показывает наиболее часто отправляемые по электронной почте статьи.

Рис. 1.13. Примеры неорганичных, неперсонализированных рекомендаций: раздел **Hot Network Questions** (Горячие вопросы в сети) с сайта Cross Validated, раздел **Most Emailed** (Наиболее часто пересылаемые) с сайта New York Times и персонализированный раздел **Recommended for You** (Рекомендуется вам) на сайте Amazon

Некоторые системы объясняют предоставленные рекомендации. Это системы, применяющие принцип «*белого ящика*». Те системы, которые свои

рекомендации не объясняют, построены на принципе «*черного ящика*». На рис. 1.14 показаны примеры каждого из типов. Это важное различие, которое необходимо учитывать при выборе алгоритма, поскольку не каждый алгоритм предусматривает возможность отследить процесс принятия решения до самых его истоков.

При выборе типа рекомендательной системы (построенной по принципу «белого ящика» или «черного ящика») нужно принимать во внимание особенности каждого из алгоритмов. Чем больше объяснений потребуется от системы, тем проще алгоритм. Часто решение можно проанализировать, как показано на рис. 1.15. Чем выше качество рекомендации, тем сложнее объяснение. Эта проблема известна как *компромисс между точностью модели и сложностью интерпретации модели*.

Как-то я работал над проектом, в котором огромное значение уделялось объясняемости и качеству. Чтобы справиться с задачей, пришлось надстроить еще один алгоритм поверх нашей рекомендательной системы. Это позволило нам получить рекомендации высокого качества, и при этом система была способна сопоставить использованные данные с результатом.

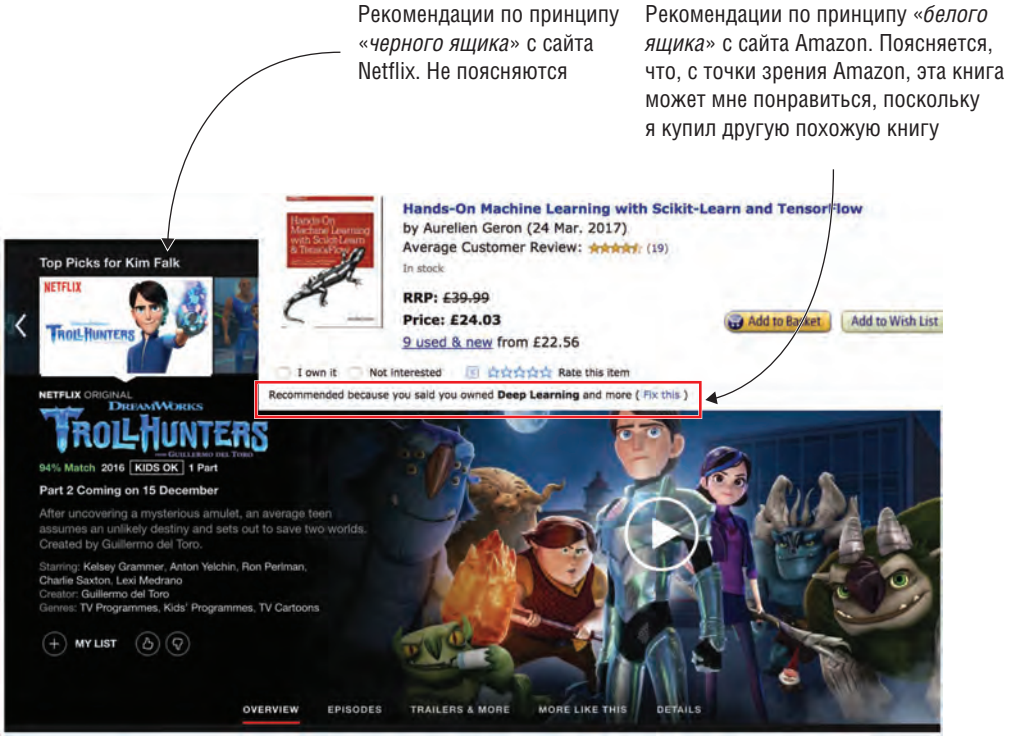

Рис. 1.14. Рекомендации по принципу «черного ящика» (с сайта Netflix) и «белого ящика» (с сайта Amazon)

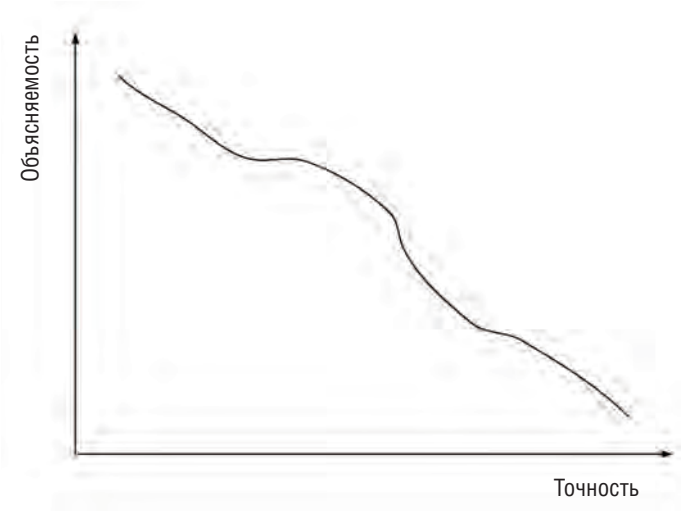

Рис. 1.15. Объясняемость и качество рекомендаций

В последнее время рекомендательные системы получили очень широкое распространение, поэтому в нашем распоряжении множество примеров. Чаще всего рекомендательные системы заточены под фильмы, музыку, книги, новости, исследовательские статьи и вообще под большинство товаров. Но рекомендательные системы применяются также и во многих других сферах, включая финансовые услуги, страхование жизни, цифровые данные, поиск работы, т. е., по сути, везде, где требуется сделать выбор. В этой книге в качестве примеров приводятся в основном сайты, но с другими платформами тоже вполне можно работать.

1.2.8. Алгоритмы

В этой книге приводится несколько алгоритмов. Алгоритмы можно разделить на две группы по типу данных, на основе которых они формируют рекомендации. Алгоритмы, которые опираются на данные о предыдущих сеансах работы пользователя с системой, называются *совместной фильтрацией*. Алгоритмы, которые при составлении рекомендаций анализируют метаданные контента и пользовательские профили, называются *контентной фильтрацией*. Сочетание этих двух типов образует гибридные рекомендательные системы.

Совместная фильтрация

На рис. 1.16 представлен один из способов реализации совместной фильтрации. Внешний контур – это весь каталог. Контур посередине – это группа людей, которые приобрели/воспользовались одними и теми же объектами. Рекомендательная система рекомендует объекты из малого овала (на переднем плане), исходя из того, что если пользователям нравится то же самое, что и текущему пользователю, то текущему пользователю понравятся и осталь-

ные объекты, заинтересовавшие данную группу. Группа определяется путем поиска соответствий между тем, что понравилось отдельным пользователям, и тем, что понравилось текущему пользователю. В этом случае в рекомендации попадет тот пласт контента, который упускает текущий пользователь (та часть средней окружности, которая не попадает в овал, включающий заинтересовавшие текущего пользователя объекты).

Существует множество способов формирования рекомендаций на основе совместной фильтрации. Простой способ описан в главе 8, а способ посложнее – в главе 11, где мы поговорим об алгоритмах факторизации матриц.

Рис. 1.16. Схема совместной фильтрации

Контентная фильтрация

Контентная фильтрация опирается на метаданные объектов из вашего каталога. Например, Netflix пользуется описаниями фильмов. В зависимости от алгоритма, система может формировать рекомендации путем подбора объектов, аналогичных тем, которые понравились пользователю ранее, путем сопоставления объектов с данными из пользовательского профиля, либо, если пользователь не задействован, путем поиска схожих объектов из всего контента. При наличии пользовательского профиля система анализирует каждый профиль, в котором указаны категории контента. Если бы сервис Netflix применял контекстную фильтрацию, он мог бы составлять пользовательские профили с разбивкой по жанрам – например, триллеры, комедии, драмы и новинки – и указывать рейтинг каждого из жанров. В этом варианте фильм попадает в рекомендации, только если его рейтинг соответствует рейтингу, указанному пользователем.

Приведем пример. Пользователю Томасу понравились фильмы «Стражи галактики», «Интерстеллар» и сериал «Игра престолов». Каждому он дал оценку по пятибалльной шкале. В табл. 1.2 показан один из способов интерпретации этих оценок.

Таблица 1.2. Система оценок на примере двух фильмов и телесериала

Фильмы и сериалы	Научная фантастика	Приключения
«Интерсталлар»	3	3
«Игра престолов»	1	5
«Стражи галактики»	5	4

С опорой на эту информацию составляется профиль Томаса, в котором научной фантастике дана оценка 3, а приключениям 4. Для подбора и рекомендации других фильмов просматривается каталог и выделяются те фильмы, которые соответствуют профилю Томаса.

Гибридная рекомендательная система

И у совместной, и у контентной фильтрации есть сильные и слабые стороны. Для нормального функционирования совместной фильтрации от пользователей должна стабильно поступать обратная связь, а контентная фильтрация подразумевает наличие подробного описания у каждого объекта. Часто рекомендации формируются на основе результатов работы этих двух алгоритмов и анализа данных другого типа, например удаленность от какой-либо локации или время суток.

1.3. Машинное обучение и Netflix Prize

Рекомендательная система предназначена, для того чтобы предугадывать, какой контент нужен пользователю в данный момент. Предугадать это можно множеством различных способов. Создание рекомендательной системы превратилось в междисциплинарное мероприятие, в котором огромную роль играют различные информационные технологии, включая машинное обучение, интеллектуальный анализ данных, поиск информации и даже человеко-компьютерное взаимодействие. Машинное обучение и интеллектуальный анализ данных позволяют вычислительной машине строить прогнозы на основе изучения примеров того, что прогнозируется. Следовательно, эти же возможности прогнозирования могут участвовать в формировании рекомендаций.

Многие рекомендательные системы построены вокруг алгоритмов машинного обучения, направленных на прогнозирование оценки, которую пользователь поставит объекту, или на то, чтобы составить наиболее подходящий для данного пользователя порядок расположения объектов. Одна из причин активного развития направления машинного обучения заключается в том, что с его помощью специалисты пытаются решить проблему рекомендательных систем. Они стремятся получить действующий алгоритм, с помощью которого компьютер мог бы угадывать наши тайные желания еще до того, как мы сами поняли, чего хотим.

Многие утверждают, что апогеем интереса к внедрению технологий машинного обучения в работу рекомендательных систем стало знаменитое соревнование Netflix Prize. Соревнование проводил сервис Netflix, пообещав-

ший выплатить 1 млн долларов каждому, кто сумеет разработать алгоритм, который улучшит рекомендации сервиса на 10 %. Соревнование началось в 2006 году, а победитель появился только через три года. В итоге выиграл гибридный алгоритм. Помните, мы говорили, что гибридный алгоритм запускает несколько алгоритмов, а затем объединяет полученные результаты и выдает общий итог? О гибридных алгоритмах мы поговорим в главе 11.

Netflix так и не применил ставший победителем алгоритм. Вероятно, причина в том, что он был настолько сложным, что неоправданно сильно возрастала нагрузка на систему. Поэтому, к сожалению, Netflix ничем не поможет в нашем рассказе об этом аспекте рекомендательных систем. Вместо этого я разработал небольшой демосайт под названием MovieGEEKs, на примере которого буду объяснять описываемые в данной книге вещи. Над сайтом необходимо как следует потрудиться, чтобы привести его в рабочее состояние. Его главная задача – это демонстрация основных принципов работы рекомендательных систем.

1.4. Интернет-сайт MovieGEEKs

Эта книга рассказывает о том, как построить рабочую рекомендательную систему. Она вооружит вас необходимыми для этого средствами независимо от того, на какой платформе будет работать ваша рекомендательная система. Но, для того чтобы придумать какую-нибудь интересную рекомендательную систему, необходимо собрать данные и понять, как все устроено, а для этого недостаточно просто посмотреть на цифры.

В центре внимания данной книги прежде всего сайты, однако это не означает, что все написанное здесь неприменимо к любому другому типу системы. Это краткое вступление подводит нас к тому фреймворку, от которого мы будем плясать.

Интернет-сайт MovieGEEKs (**mng.bz/04k5**) с помощью фреймворка Django. Я советую вам скачать MovieGEEKs и обращаться к нему по ходу чтения книги, поскольку так вам проще будет понять, о чем идет речь. Тот факт, что при создании сайта применялся фреймворк Django, не имеет большого значения. Приводя какие-либо примеры, я буду подсказывать вам, на что надо обратить внимание.

> **Интернет-сайт и фреймворк Django**
>
> Если словосочетание фреймворк Django вам ни о чем не говорит, почитайте документацию по Django на странице **www.djangoproject.com/start/overview/**.

Зайти нужно только на один этот интернет-сайт. Он содержит в себе весь функционал, описанный в данной книге. Он отражает именно тот вымышленный сценарий, которого мы будем придерживаться.

Представьте себе, что у вас есть клиент, который хочет продавать DVD-диски через интернет. Мне сразу приходит в голову старый магазинчик с прокатом DVD в английском городке Бате, владелец которого хочет попробовать продавать фильмы через интернет. К сожалению, этого магазина больше нет (рис. 1.17).

Магазин ни с какой точки зрения нельзя было назвать электронным. Учет дисков велся с помощью бумажных карточек и – хоть вам, возможно, это покажется невероятным – система работала! В реальной жизни вряд ли владелец когда-либо смог перевести свой бизнес в интернет-среду, но одной из уникальных особенностей этого места было то, что посетители всегда получали там отличные рекомендации. Владелец составлял ежемесячные обзоры – рекомендации на основе экспертного мнения, – и люди, которые там работали, всегда знали о фильмах все.

Мне хочется верить, что рекомендательная система – это попытка предложить индивидуальный подход людям в интернете. Ниже в краткой форме изложены пожелания нашего выдуманного владельца.

Рис. 1.17. Витрина вымышленного магазина On the Video Front

1.4.1. Оформление и характеристики

Для начала необходимо обозначить несколько базовых идей оформления. На главной странице сайта должно присутствовать следующее:

- область с обложками фильмов;
- обзор каждого фильма, для прочтения которого не требуется переходить на следующую страницу;
- рекомендации, как можно более персонализированные;
- меню со списком жанров.

У каждого фильма должна быть собственная страница, на которой должны быть:

- киноафиша;
- описание;
- рейтинг.

Для каждой категории должна быть выделена собственная страница, отображающая:

- ту же структуру, что и главная страница;
- рекомендации по данной категории.

1.4.2. Архитектура

Для разработки данного сайта будет применяться язык программирования Python и фреймворк Django. Django позволяет разбить проект на несколько отдельных приложений. На рис. 1.18 показана высокоуровневая архитектура и показан пример того, какие приложения участвуют в создании сайта.

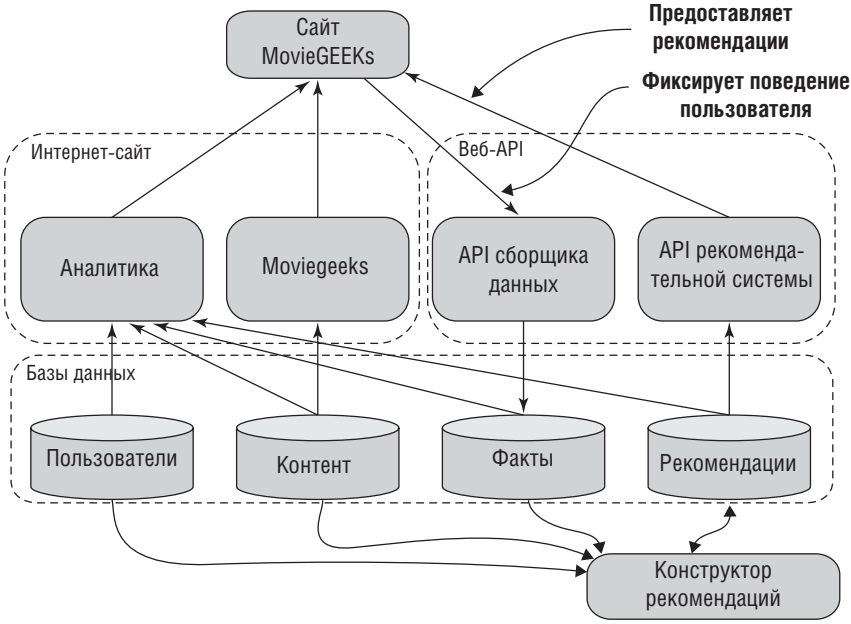

Рис. 1.18. Архитектура сайта MovieGEEKs

Давайте по-быстрому пробежимся по основным моментам:

- *MovieGEEKs* – это основная часть сайта. Здесь клиентская логика (HTML, CSS, JavaScript), наряду с программным кодом на языке Python, ориентирована на получение данных о фильмах.
- *Аналитика* – это капитанский мостик, откуда можно контролировать все процессы. Эта часть опирается на данные из всех баз. Об аналитике рассказывается в главе 4.
- *Сборщик данных* – отвечает за отслеживание шаблонов поведения пользователя и сохраняет их в базу фактов. Журнал фактов описан в главе 2.
- *Рекомендации* – сердце всей этой системы, без которого существование сайта было бы бессмысленно. Отсюда рекомендации поступают на сайт MovieGEEKs. Об этой части рассказывается в главе 5 и далее в книге.
- *Конструктор рекомендаций* – занимается предварительным составлением рекомендаций, на базе которых формируются тщательно проработанные рекомендации для пользователя. Разбираться с конструктором рекомендаций мы начнем в главе 7.

Каждый из этих компонентов или приложений содержит интересные модели данных и функции. Это станет наживкой для будущих посетителей.

MovieGEEKs – это сайт с фильмами, в основном потому что в нашем распоряжении имеется большой набор данных и контента, касающихся фильмов, пользователей и рейтингов. Более того, этот контент включает в себя URL-адреса, содержащие киноафиши, а с таким контентом работать очень приятно.

На рис. 1.19 показана главная страница сайта MovieGEEKs, т. е. целевая страница. Когда пользователь щелкает по значку фильма, появляется всплывающее окно с дополнительной информацией и ссылкой на более подробное описание.

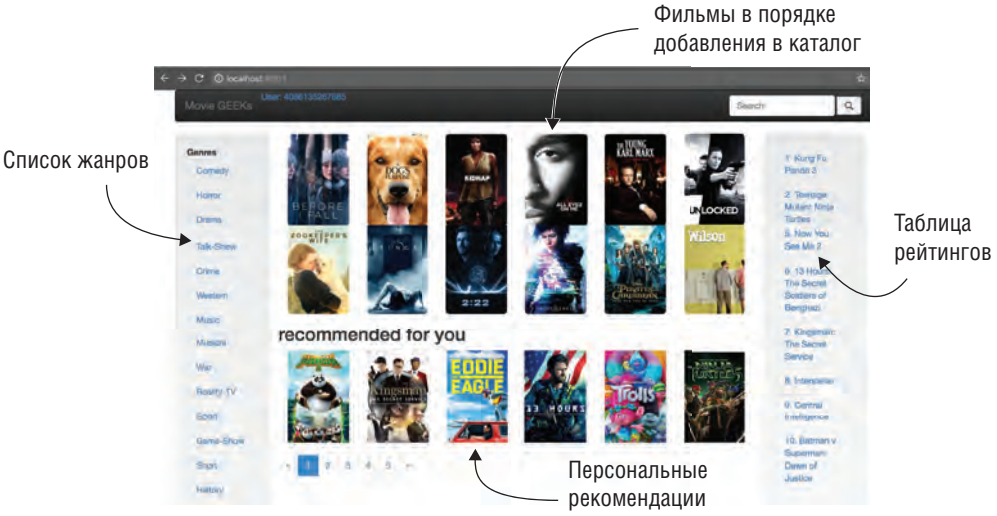

Рис. 1.19. Главная страница сайта MovieGEEKs

Вот и все! Просто, но эффективно. Скачайте материалы прямо сейчас. Инструкции по установке вы найдете в readme-файле на сервисе GitHub по ссылке **mng.bz/04k5**. Сайт MovieGEEKs опирается на набор данных под названием MovieTweetings. В этот набор данных входят оценки и рейтинги фильмов, полученные из качественных записей в Twitter[1].

1.5. Создание рекомендательной системы

Прежде чем двигаться дальше, давайте посмотрим, как создается рекомендательная система. Предположим, что у вас уже есть готовая платформа – например, сайт или приложение – и вам нужно просто добавить туда рекомендательную систему. В этом случае весь процесс будет выглядеть примерно так, как показано на рис. 1.20.

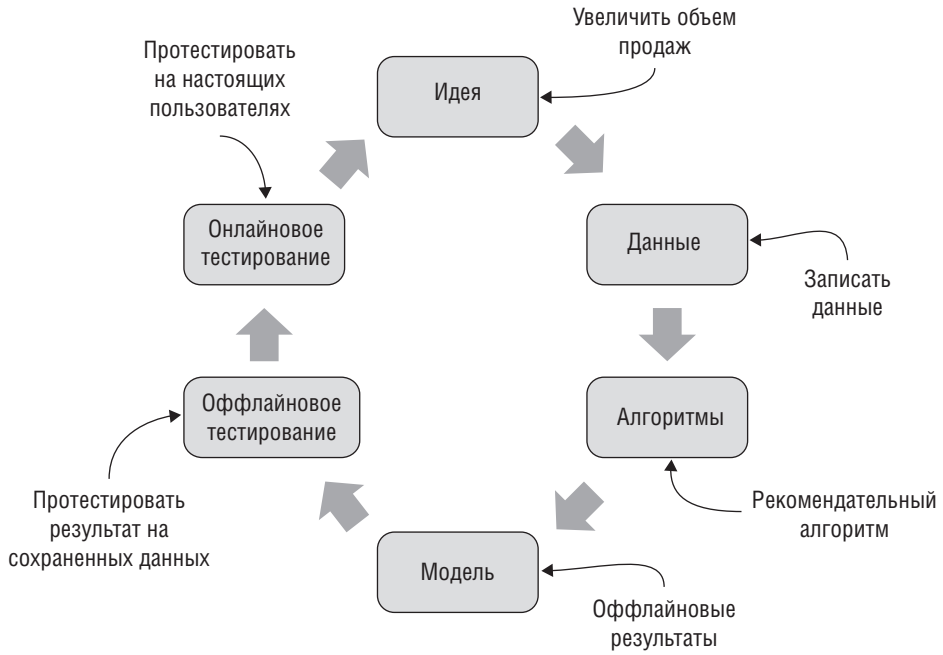

Рис. 1.20. Ориентированный на данные подход к созданию рекомендательной системы

Сначала у вас появляется желание увеличить объем продаж и возникает идея, что достичь этого можно с помощью рекомендательной системы. Вы начинаете собирать данные о поведении пользователей и на основе этих данных создаете алгоритм, при выполнении которого запускается модель. Эту модель можно также считать *функцией*, которая, получив идентификационные данные пользователя, просчитает рекомендации.

Вы опробуете эту модель на ретроспективных данных, чтобы проверить, можно ли с их помощью спрогнозировать поведение пользователя в буду-

[1] Подробнее читайте по ссылке **github.com/sidooms/MovieTweetings**.

щем. Например, если у вас есть данные о покупках пользователей в прошлом месяце, вы можете создать модель на основе данных за первые три недели и посмотреть, насколько хорошо она спрогнозирует поведение пользователя за последующие три недели того месяца, за который у вас собраны данные. Возможно, такая модель предугадает, какие товары купили пользователи, даже точнее, чем это сделала бы базовая рекомендательная система, которая может просто выполнять метод, выдающий наиболее популярные товары. В случае успеха можно испытать этот метод на группе пользователей и проверить, есть ли изменения к лучшему. Если есть, тогда метод работает и может применяться, в ином случае необходимо вернуться назад и доработать его.

Теперь вы уже должны понимать, что такое рекомендательная система, в каких данных она нуждается и что она может дать на выходе. Зная основные принципы работы рекомендательных систем, можно переходить к главе 2, в которой говорится о том, как собирать данные о пользователях.

Резюме

- Сервис Netflix с помощью рекомендаций обеспечивает персональный подход к каждому пользователю и помогает пользователям находить контент, который им понравится.
- В широкое понятие «рекомендательная система» входит множество различных компонентов и методов.
- Прогноз и рекомендация – это не одно и то же. Прогнозирование – это попытка угадать, какую оценку пользователь даст тому или иному контенту, а рекомендация – это список объектов, соответствующих вкусам пользователя.
- Контекст рекомендации – это все, что происходит вокруг пользователя (окружающая пользователя обстановка) в момент формирования рекомендации. Объекты, которые, возможно, по прогнозу, получат не самые высокие оценки, могут попасть в рекомендации, если вписываются в контекст.
- Описанная в данной главе таксономия может пригодиться, когда вы изучаете другие рекомендательные системы или пытаетесь разработать собственную. К данной таксономии не мешает обратиться, перед тем как начинать работу над своей рекомендательной системой.