

УДК 004.594

ББК 32.972

Ф19

Ким Фальк

Ф19 Рекомендательные системы на практике / пер. с англ. Д. М. Павлова. – М.: ДМК Пресс, 2020. – 448 с.: ил.

ISBN 978-5-97060-774-9

Книга посвящена рекомендательным системам, которые собирают данные о пользователе и выводят для него персональные рекомендации, основываясь на его предпочтениях. Ким Фальк, специалист по обработке и анализу данных, предоставляет читателю самые важные сведения о рекомендательных системах – начиная с общего обзора и описания ключевых алгоритмов до рассмотрения тонких нюансов работы, благодаря которым система с максимальной точностью учитывает интересы пользователя. Помимо прочего, обсуждаются методы оценки рекомендательной системы вне интернета и возможности совмещения различных рекомендательных систем.

Книга снабжена многочисленными примерами программного кода.

Издание предназначено для широкого круга разработчиков и специалистов по анализу данных.

УДК 004.594

ББК 32.972

Original English language edition published by Manning Publications. Copyright © 2019 by Manning Publications. Russian language edition copyright © 2020 by DMK Press. All rights reserved.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-1-61729-2705 (англ.)

ISBN 978-5-97060-774-9 (рус.)

Copyright © 2019 by Manning Publications Co.

© Оформление, перевод, издание, ДМК Пресс, 2020

Оглавление

Предисловие от издательства	13
Предисловие	14
Благодарности	16
О книге	17
Об авторе	20
Об обложке	21
ЧАСТЬ I. Подготовка к рекомендательным системам	23
Глава 1. Что такое рекомендательная система?	25
1.1. Рекомендации в реальной жизни	25
1.1.1. Рекомендательные системы дома в интернете	27
1.1.2. Длинный хвост	28
1.1.3. Рекомендательная система Netflix	28
1.1.4. Определение рекомендательной системы.....	35
1.2. Таксономия рекомендательных систем	38
1.2.1. Специализация	39
1.2.2. Задача	39
1.2.3. Контекст.....	40
1.2.4. Степень персонализации	40
1.2.5. Чье мнение	42
1.2.6. Конфиденциальность и надежность.....	42
1.2.7. Интерфейс.....	43
1.2.8. Алгоритмы.....	46
1.3. Машинное обучение и Netflix Prize	48
1.4. Интернет-сайт MovieGEEKs.....	49
1.4.1. Оформление и характеристики	51
1.4.2. Архитектура.....	51
1.5. Создание рекомендательной системы	53
Резюме	54
Глава 2. Поведение пользователя, и как собирать о нем данные	55
2.1. Как (по моему мнению) Netflix собирает факты, пока вы пользуетесь сервисом	56
2.1.1. Какие факты собирает Netflix.....	58

2.2. Поиск полезных данных о поведении пользователя	60
2.2.1. Как узнать мнение посетителя	61
2.2.2. Что можно узнать по поведению обозревателя в магазине	62
2.2.3. Совершение покупки	67
2.2.4. Пользование товаром	68
2.2.5. Оценки посетителей	69
2.2.6. Знакомство с клиентами по (старому) методу Netflix.....	73
2.3. Идентификация пользователей.....	73
2.4. Получение данных о посетителях из других источников	74
2.5. Сборщик данных.....	74
2.5.1. Создание файлов проекта	76
2.5.2. Модель данных.....	76
2.5.3. Сборщик данных на стороне клиента	77
2.5.4. Интеграция сборщика в MovieGEEKs	78
Регистрация наведения курсора.....	80
Регистрация просмотра подробностей.....	80
Регистрация «сохранения на потом»	80
2.6. Какие пользователи есть в системе, и как их моделировать	81
Резюме.....	84
Глава 3. Мониторинг состояния системы.....	85
3.1. Почему панель аналитики – это круто	86
3.1.1. Ответ на вопрос «Как там дела у сайта?»	86
3.2. Реализация аналитики	88
3.2.1. Веб-аналитика.....	88
3.2.2. Базовые статистические данные	88
3.2.3. Конверсии.....	89
3.2.4. О пути к конверсиям.....	92
3.2.5. Путь конверсии	94
3.3. Архетипы	97
3.4. Панель сайта MovieGEEKs	100
3.4.1. Автоматическая генерация данных в журнале.....	100
3.4.2. Характеристики и дизайн панели аналитики.....	101
3.4.3. Основа панели аналитики.....	101
3.4.4. Архитектура.....	102
Резюме.....	104
Глава 4. Оценки, и как их рассчитывать	105
4.1. Предпочтения элементов пользователями.....	106
4.1.1. Определение оценок.....	106
4.1.2. Матрица пользователь–элемент.....	107
4.2. Явные или неявные оценки	109
4.2.1. Как мы используем доверенные источники для составления рекомендаций	110

4.3. Переоценка.....	111
4.4. Что такое неявные оценки?.....	111
4.4.1. Предложения людей	113
4.4.2. Что учитывать при расчете оценок	113
4.5. Расчет неявных оценок.....	116
4.5.1. Просмотр поведенческих данных.....	117
4.6. Как реализовать неявные оценки.....	122
4.6.1. Добавление учета времени	126
4.7. Более редкие элементы имеют большую ценность.....	128
Резюме.....	131
Глава 5. Неперсонализированные рекомендации	132
5.1. Что такое неперсонализированные рекомендации?	133
5.1.1. Что такое реклама?	133
5.1.2. Что делает рекомендация?.....	134
5.2. Как сделать рекомендации, когда у вас нет данных.....	135
5.2.1. Топ-10: диаграмма элементов	136
5.3. Реализация диаграмм и основы для рекомендатора	138
5.3.1. Компонент рекомендательной системы	138
5.3.2. Код MovieGEEKs на сайте GitHub	139
5.3.3. Рекомендательная система	140
5.3.4. Добавление диаграмм на MovieGEEKs	140
5.3.5. Заставим контент выглядеть более привлекательно	142
5.4. Выборочные рекомендации.....	144
5.4.1. Часто покупаемые элементы, похожие на тот, который вы просматриваете	144
5.4.2. Ассоциативные правила	145
5.4.3. Реализация ассоциативных правил.....	150
5.4.4. Сохранение ассоциативных правил в базе данных.....	154
5.4.5. Запуск калькулятора ассоциаций	155
5.4.6. Использование различных событий для создания ассоциативных правил	157
Резюме.....	158
Глава 6. «Холодные» пользователи и контент.....	159
6.1. Что такое холодный старт?.....	159
6.1.1. Холодные товары	161
6.1.2. Холодный посетитель	161
6.1.3. Серые овцы.....	163
6.1.4. Посмотрим на примеры из реальной жизни	163
6.1.5. Что вы можете сделать с холодным стартом?.....	164
6.2. Отслеживание посетителей.....	165
6.2.1. Анонимные пользователи	165

6.3. Решение проблемы холодного старта алгоритмами	165
6.3.1. Использование ассоциативных правил для создания рекомендаций для холодных пользователей	166
6.3.2. Использование знаний предметной области и бизнес-правил	168
6.3.3. Использование сегментов	168
6.3.4. Использование категорий с целью обойти проблему серых овец и холодных продуктов	170
6.4. Кто не спрашивает, тот не будет знать	172
6.4.1. Когда посетитель уже не новый	173
6.5. Использование ассоциативных правил с целью ускорить показ рекомендаций	173
6.5.1. Поиск собранных элементов	174
6.5.2. Получение ассоциативных правил и сортировка в соответствии со значениями уверенности	174
6.5.3. Отображение рекомендаций	176
6.5.4. Оценка реализации	179
Резюме	179
Часть II. Рекомендательные алгоритмы	181
Глава 7. Выявление общих черт у пользователей и контента	183
7.1. Что за сходство?	184
7.1.1. Что такое функция подобия?	185
7.2. Основные функции подобия	185
7.2.1. Расстояние Жаккара	187
7.2.2. Измерение расстояния с помощью Lp-норм	189
7.2.3. Коэффициент Отиаи	192
7.2.4. Вычисление сходства с помощью коэффициента корреляции Пирсона	194
7.2.5. Испытание сходства коэффициентом Пирсона	195
7.2.6. Коэффициент корреляции Пирсона на коэффициент Отиаи	198
7.3. Кластеризация k-средних	198
7.3.1. Алгоритм кластеризации k-средних	199
7.3.2. Реализация кластеризации k-средних на Python	201
7.4. Реализация вычисления сходства	206
7.4.1. Реализация вычисления сходства на сайте MovieGEEKs	208
7.4.2. Реализация кластеризации на сайте MovieGEEKs	210
Резюме	214
Глава 8. Совместная фильтрация в окрестностях	215
8.1. Совместная фильтрация: историческая справка	217
8.1.1. Когда начали использовать совместную фильтрацию	217
8.1.2. Взаимопомощь	217
8.1.3. Матрица оценок	219

8.1.4. Процедура совместной фильтрации.....	220
8.1.5. Нужно использовать совместную фильтрацию пользователь– пользователь или элемент–элемент?	221
8.1.6. Требования к данным	222
8.2. Расчет рекомендации.....	222
8.3. Расчет сходства	223
8.4. Алгоритм вычисления сходства элементов с Amazon.....	223
Если проблема повторяется – берегись!.....	227
8.5. Способы выбора окрестности.....	228
8.6. Поиск правильной окрестности	230
8.7. Методы прогнозирования оценок.....	230
8.8. Прогнозирование с фильтрацией по элементам.....	232
8.8.1. Вычисление прогнозов.....	233
8.9. Проблема холодного старта	233
8.10. Пара слов о терминах машинного обучения	234
8.11. Совместная фильтрация на сайте MovieGEEKs.....	235
8.11.1. Фильтрация элементов	236
8.12. В чем разница между правилами ассоциации и совместной фильтрацией?.....	242
8.13. Эксперименты с совместной фильтрацией	242
8.14. Преимущества и недостатки совместной фильтрации	244
Резюме	245
Глава 9. Оценка и тестирование рекомендательной системы.....	246
9.1. Бизнесу нужен подъем, перекрестные продажи, рост продаж и конверсии.....	247
9.2. Зачем оценивать?.....	248
Гипотеза.....	249
9.3. Как интерпретировать поведение пользователей	249
9.4. Что измерять.....	249
9.4.1. Понимание вкусов пользователя: сведение к минимуму ошибки предсказания.....	250
9.4.2. Разнообразие	251
9.4.3. Охват	252
9.4.4. Приятные неожиданности	254
9.5. Перед реализацией рекомендатора.....	255
9.5.2. Регрессионное тестирование	256
9.6. Виды оценки.....	257
9.7. Офлайн-оценка.....	257
9.7.1. Что делать, если алгоритм не дает рекомендаций	258
9.8. Офлайн-эксперименты	259

9.8.1. Подготовка данных для эксперимента	264
9.9. Реализация эксперимента в MovieGEEKs.....	270
9.9.1. Список дел.....	271
9.10. Оценка тестового набора.....	274
9.10.1. Начнем с базовых прогнозов	275
9.10.2. Поиск правильных параметров	277
9.11. Онлайн-оценка.....	278
9.11.1. Контролируемые эксперименты.....	279
9.11.2. А/В-тестирование.....	279
9.12. Непрерывное тестирование с использованием/исследованием	280
9.12.1. Петли обратной связи.....	281
Глава 10. Фильтрация по контенту	283
10.1. Описательный пример	283
10.2. Фильтрация на основе контента	286
10.3. Анализатор контента	288
10.3.1. Выделение признаков для профиля элемента.....	288
10.3.2. Редко встречающиеся данные	290
10.3.3. Преобразование года в сопоставимую функцию.....	290
10.4. Извлечение метаданных из описаний	291
10.4.1. Составление описаний	291
10.5. Поиск важных слов методом TF-IDF.....	295
10.6. Моделирование темы с использованием LDA.....	297
10.6.1. Какими крутилками настраивать LDA?	303
10.7. Поиск подобного контента	306
10.8. Создание профиля пользователя.....	307
10.8.1. Создание профиля пользователя с помощью модели LDA	307
10.8.2. Создание профиля пользователя с помощью модели TF-IDF	308
10.9. Рекомендации на основе контента на сайте MovieGEEKs	310
10.9.1. Загрузка данных.....	310
10.9.2. Обучение модели	313
10.9.3. Создание профилей элементов.....	314
10.9.4. Создание пользовательских профилей	314
10.9.5. Отображение рекомендаций.....	316
10.10. Оценка рекомендатора на основе контента	317
10.11. Плюсы и минусы фильтрации на основе контента	319
Резюме.....	320
Глава 11. Определение скрытых жанров с помощью матричной факторизации	321
11.1. Иногда чем меньше данных, тем лучше.....	322
11.2. Пример задачи	324

11.3. Немножко линейной алгебры	327
11.3.1. Матрица	327
11.3.2. Что за факторизация?	329
11.4. Выполнение факторизации с использованием SVD.....	331
11.4.1. Добавление новых пользователей путем складывания	336
11.4.2. Как формировать рекомендации с помощью SVD	338
11.4.3. Базисные предикторы	339
11.4.4. Временная динамика.....	342
11.5. Построение факторизации с помощью Funk SVD	342
11.5.1. Корень средней квадратичной ошибки	343
11.5.2. Градиентный спуск	344
11.5.3. Стохастический градиентный спуск.....	347
11.5.4. Перейдем, наконец, к факторизации	347
11.5.5. Добавление отклонений	349
11.5.6. Как начать и когда остановиться	350
11.6. Генерация рекомендаций с помощью Funk SVD	354
11.7. Реализация Funk SVD на MovieGEEKs	356
11.7.1. Что делать с неподходящими рекомендациями	361
11.7.2. Поддержание актуальности модели	362
11.7.3. Более быстрая реализация.....	363
11.8. Явные данные против неявных данных.....	363
11.9. Оценка	363
11.10. Эксперименты с моделью Funk SVD	365
Резюме.....	367

Глава 12. С каждого по способностям – реализуем гибридный алгоритм рекомендательной системы..... 368

12.1. Сложности мира гибридов	369
12.2. Монолитные рекомендаторы.....	370
12.2.1. Смешивание функций контента с поведенческими данными для улучшения алгоритмов на основе совместной фильтрации	371
12.3. Смешанный гибридный рекомендатор	372
12.4. Ансамбль.....	372
12.4.1. Переключаемый ансамбль рекомендаторов.....	374
12.4.2. Взвешенная ансамбль рекомендаторов	375
12.4.3. Линейная регрессия.....	376
12.5. Признако-взвешенное линейное сочетание (FWLS)	377
12.5.1. Представляем веса в виде функций.....	378
12.5.2. Алгоритм.....	380
12.6. Реализация	387
Резюме.....	396

Глава 13. Ранжирование и обучение ранжированию	397
13.1. Обучение ранжированию на примере Foursquare	398
13.2. Переранжирование	402
13.3. Еще раз – что такое обучение ранжированию?	403
13.3.1. Три типа алгоритмов LTR	403
13.4. Байесовское персонализированное ранжирование	405
13.4.1. Ранжирование с BPR	407
13.4.2. Магия математики (продвинутое колдовство)	409
13.4.3. Алгоритм BPR	412
13.4.4. BPR с матричной факторизацией	413
13.5. Реализация BPR	413
13.5.1. Генерация рекомендаций	419
13.6. Оценка	421
13.7. Эксперименты с BPR	423
Резюме	424
Глава 14. Будущее рекомендательных систем	425
14.1. Вся книга в паре предложений	426
14.2. Темы для дальнейшего изучения	429
14.2.1. Дальнейшее чтение	429
14.2.2. Алгоритмы	430
14.2.3. Контекст	430
14.2.4. Взаимодействие «Человек–Машина»	431
14.2.5. Выбор подходящей архитектуры	431
14.3. Что ждет рекомендательные системы в будущем?	432
14.4. Послесловие	436
Предметный указатель	438