



Ю. Лесковец, А. Раджараман, Дж. Ульман

Анализ больших наборов данных

Юре Лесковец, Ананд Раджараман, Джеффри Д. Ульман

Анализ больших наборов данных

Mining of Massive Datasets

Jure Leskovec
Stanford Univ.

Anand Rajaraman
Milliway Labs

Jeffrey D. Ullman
Stanford Univ.



CAMBRIDGE
UNIVERSITY PRESS

Анализ больших наборов данных

Юре Лесковец
Stanford Univ.

Ананд Раджараман
Milliway Labs

Джеффри Д. Ульман
Stanford Univ.

2-е издание, электронное



Лесковец, Юре.

Л50 Анализ больших наборов данных / Ульман Дж. Д., Лесковец Ю., Раджараман А. ; пер. с англ. А. А. Слинкина. — 2-е изд., эл. — 1 файл pdf : 500 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-304-2

Эта книга написана ведущими специалистами в области технологий баз данных и веба. Благодаря популярности интернет-торговли появилось много чрезвычайно объемных баз данных, для извлечения информации из которых нужно применять методы добычи данных (data mining).

В книге описываются алгоритмы, которые реально использовались для решения важнейших задач добычи данных и могут быть с успехом применены даже к очень большим наборам данных. Изложение начинается с рассмотрения технологии MapReduce — важного средства распараллеливания алгоритмов. Излагаются алгоритмы хэширования с учетом близости и потоковой обработки данных, которые поступают слишком быстро для тщательного анализа. В последующих главах рассматривается идея показателя PageRank, нахождение частых предметных наборов и кластеризация. Во второе издание включен дополнительный материал о социальных сетях, машинном обучении и понижении размерности.

Издание будет в равной мере полезна студентам и программистам-практикам.

УДК 004.6
ББК 32.972

Электронное издание на основе печатного издания: Анализ больших наборов данных / Ульман Дж. Д., Лесковец Ю., Раджараман А. ; пер. с англ. А. А. Слинкина. — Москва : ДМК Пресс, 2016. — 498 с. — ISBN 978-5-97060-190-7. — Текст : непосредственный.

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.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

В соответствии со ст. 1299 и 1301 ГК РФ при устранении ограничений, установленных техническими средствами защиты авторских прав, правообладатель вправе требовать от нарушителя возмещения убытков или выплаты компенсации.

ISBN 978-5-89818-304-2

© 2010, 2011, 2012, 2013, 2014 Anand Rajaraman,
Jure Leskovec, Jeffrey D. Ullman
© Оформление, издание, ДМК Пресс, 2016



ОГЛАВЛЕНИЕ

Предисловие	17
О чем эта книга.....	17
Требования к читателю	18
Упражнения.....	18
Поддержка в вебе.....	18
Автоматизированные домашние задания	18
Благодарности	19
ГЛАВА 1.	
Добыча данных	20
1.1. Что такое добыча данных?	20
1.1.1. Статистическое моделирование	20
1.1.2. Машинное обучение	21
1.1.3. Вычислительные подходы к моделированию	21
1.1.4. Обобщение	22
1.1.5. Выделение признаков.....	23
1.2. Статистические пределы добычи данных	23
1.2.1. Тотальное владение информацией	24
1.2.2. Принцип Бонферрони	24
1.2.3. Пример применения принципа Бонферрони	25
1.2.4. Упражнения к разделу 1.2	26
1.3. Кое-какие полезные сведения	26
1.3.1. Важность слов в документах	27
1.3.2. Хэш-функции	28
1.3.3. Индексы	29
1.3.4. Внешняя память.....	31
1.3.5. Основание натуральных логарифмов	31
1.3.6. Степенные зависимости	32
1.3.7. Упражнения к разделу 1.3	34
1.4. План книги	35
1.5. Резюме	37
1.6. Список литературы	38

ГЛАВА 2.**MapReduce и новый программный стек 39**

2.1. Распределенные файловые системы	40
2.1.1. Физическая организация вычислительных узлов	40
2.1.2. Организация больших файловых систем.....	42
2.2. MapReduce	42
2.2.1. Задачи-распределители	44
2.2.2. Группировка по ключу	44
2.2.3. Задачи-редукторы	45
2.2.4. Комбинаторы.....	45
2.2.5. Детали выполнения MapReduce.....	46
2.2.6. Обработка отказов узлов	48
2.2.7. Упражнения к разделу 2.2	48
2.3. Алгоритмы, в которых используется MapReduce.....	48
2.3.1. Умножение матрицы на вектор с применением MapReduce	49
2.3.2. Если вектор v не помещается в оперативной памяти.....	50
2.3.3. Операции реляционной алгебры.....	51
2.3.4. Вычисление выборки с помощью MapReduce	53
2.3.5. Вычисление проекции с помощью MapReduce.....	54
2.3.6. Вычисление объединения, пересечения и разности с помощью MapReduce	54
2.3.7. Вычисление естественного соединения с помощью MapReduce.....	55
2.3.8. Вычисление группировки и агрегирования с помощью MapReduce	56
2.3.9. Умножение матриц	56
2.3.10. Умножение матриц за один шаг MapReduce.....	57
2.3.11. Упражнения к разделу 2.3	58
2.4. Обобщения MapReduce	59
2.4.1. Системы потоков работ	60
2.4.2. Рекурсивные обобщения MapReduce.....	61
2.4.3. Система Pregel	64
2.4.4. Упражнения к разделу 2.4	65
2.5. Модель коммуникационной стоимости	65
2.5.1. Коммуникационная стоимость для сетей задач.....	65
2.5.2. Физическое время	68
2.5.3. Многопутевое соединение.....	68
2.5.4. Упражнения к разделу 2.5	71
2.6. Теория сложности MapReduce	73
2.6.1. Размер редукции и коэффициент репликации	73
2.6.2. Пример: соединение по сходству.....	74
2.6.3. Графовая модель для проблем MapReduce.....	76
2.6.4. Схема сопоставления	78
2.6.5. Когда присутствуют не все входы.....	79
2.6.6. Нижняя граница коэффициента репликации	80
2.6.7. Пример: умножение матриц.....	82
2.6.8. Упражнения к разделу 2.6	86
2.7. Резюме	87

2.8. Список литературы	89
------------------------------	----

ГЛАВА 3.

Поиск похожих объектов 92

3.1. Приложения поиска близкого соседям	92
3.1.1. Сходство множеств по Жаккару	93
3.1.2. Сходство документов	93
3.1.3. Коллаборативная фильтрация как задача о сходстве множеств	94
3.1.4. Упражнения к разделу 3.1	96
3.2. Разбиение документов на шинглы	96
3.2.1. k-шинглы	97
3.2.2. Выбор размера шингла	97
3.2.3. Хэширование шинглов	98
3.2.4. Шинглы, построенные из слов	98
3.2.5. Упражнения к разделу 3.2	99
3.3. Сигнатуры множеств с сохранением сходства	100
3.3.1. Матричное представление множеств	100
3.3.2. Минхэш	101
3.3.3. Минхэш и коэффициент Жаккара	102
3.3.4. Минхэш-сигнатуры	102
3.3.5. Вычисление минхэш-сигнатур	103
3.3.6. Упражнения к разделу 3.3	105
3.4. Хэширование документов с учетом близости	107
3.4.1. LSH для минхэш-сигнатур	107
3.4.2. Анализ метода разбиения на полосы	109
3.4.3. Сочетание разных методов	110
3.4.4. Упражнения к разделу 3.4	111
3.5. Метрики	111
3.5.1. Определение метрики	112
3.5.2. Евклидовы метрики	112
3.5.3. Расстояние Жаккара	113
3.5.4. Косинусное расстояние	114
3.5.5. Редакционное расстояние	114
3.5.6. Расстояние Хэмминга	115
3.5.7. Упражнения к разделу 3.5	116
3.6. Теория функций, учитывающих близость	118
3.6.1. Функции, учитывающие близость	119
3.6.2. LSH-семейства для расстояния Жаккара	120
3.6.3. Расширение LSH-семейства	120
3.6.4. Упражнения к разделу 3.6	122
3.7. LSH-семейства для других метрик	123
3.7.1. LSH-семейства для расстояния Хэмминга	123
3.7.2. Случайные гиперплоскости и косинусное расстояние	124
3.7.3. Эскизы	125
3.7.4. LSH-семейства для евклидова расстояния	126
3.7.5. Другие примеры LSH-семейств в евклидовых пространствах	127

3.7.6. Упражнения к разделу 3.7	128
3.8. Применения хэширования с учетом близости	129
3.8.1. Отождествление объектов	129
3.8.2. Пример отождествления объектов	129
3.8.3. Проверка отождествления записей	131
3.8.4. Сравнение отпечатков пальцев	132
3.8.5. LSH-семейство для сравнения отпечатков пальцев	132
3.8.6. Похожие новости	134
3.8.7. Упражнения к разделу 3.8	135
3.9. Методы для высокой степени сходства	136
3.9.1. Поиск одинаковых объектов	137
3.9.2. Представление множеств в виде строк	137
3.9.3. Фильтрация на основе длины строки	138
3.9.4. Префиксное индексирование	138
3.9.5. Использование информации о позиции	140
3.9.6. Использование позиции и длины в индексах	141
3.9.7. Упражнения к разделу 3.9	144
3.10. Резюме	144
3.11. Список литературы	147

ГЛАВА 4.

Анализ потоков данных	149
4.1. Потокковая модель данных	149
4.1.1. Система управления потоками данных	150
4.1.2. Примеры источников потоков данных	151
4.1.3. Запросы к потокам	152
4.1.4. Проблемы обработки потоков	153
4.2. Выборка данных из потока	154
4.2.1. Пояснительный пример	154
4.2.2. Получение репрезентативной выборки	155
4.2.3. Общая постановка задачи о выборке	155
4.2.4. Динамическое изменение размера выборки	156
4.2.5. Упражнения к разделу 4.2	156
4.3. Фильтрация потоков	157
4.3.1. Пояснительный пример	157
4.3.2. Фильтр Блума	158
4.3.3. Анализ фильтра Блума	158
4.3.4. Упражнения к разделу 4.3	160
4.4. Подсчет различных элементов в потоке	160
4.4.1. Проблема Count-Distinct	160
4.4.2. Алгоритм Флажол-Мартена	161
4.4.3. Комбинирование оценок	162
4.4.4. Требования к памяти	163
4.4.5. Упражнения к разделу 4.4	163
4.5. Оценивание моментов	163
4.5.1. Определение моментов	163

4.5.2. Алгоритм Алона-Матиаса-Сегеди для вторых моментов	164
4.5.3. Почему работает алгоритм Алона-Матиаса-Сегеди	165
4.5.4. Моменты высших порядков.....	166
4.5.5. Обработка бесконечных потоков.....	166
4.5.6. Упражнения к разделу 4.5	168
4.6. Подсчет единиц в окне	169
4.6.1. Стоимость точного подсчета.....	169
4.6.2. Алгоритм Датара-Гиониса-Индыка-Мотвани	170
4.6.3. Требования к объему памяти для алгоритма DGIM	171
4.6.4. Ответы на вопросы в алгоритме DGIM.....	172
4.6.5. Поддержание условий DGIM	172
4.6.6. Уменьшение погрешности	174
4.6.7. Обобщения алгоритма подсчета единиц.....	174
4.6.8. Упражнения к разделу 4.6	175
4.7. Затухающие окна	176
4.7.1. Задача о самых частых элементах.....	176
4.7.2. Определение затухающего окна	176
4.7.3. Нахождение самых популярных элементов	177
4.8. Резюме	178
4.9. Список литературы	180

ГЛАВА 5.

Анализ ссылок 182

5.1. PageRank	182
5.1.1. Ранние поисковые системы и спам термов	183
5.1.2. Определение PageRank	184
5.1.3. Структура веба	187
5.1.4. Избегание тупиков.....	189
5.1.5. Паучьи ловушки и телепортация	192
5.1.6. Использование PageRank в поисковой системе.....	194
5.1.7. Упражнения к разделу 5.1	194
5.2. Эффективное вычисление PageRank.....	196
5.2.1. Представление матрицы переходов	196
5.2.2. Итеративное вычисление PageRank с помощью MapReduce	197
5.2.3. Использование комбинаторов для консолидации резльтирующего вектора.....	198
5.2.4. Представление блоков матрицы переходов	199
5.2.5. Другие эффективные подходы к итеративному вычислению PageRank	200
5.2.6. Упражнения к разделу 5.2	201
5.3. Тематический PageRank.....	202
5.3.1. Зачем нужен тематический PageRank	202
5.3.2. Смещенное случайное блуждание	202
5.3.3. Использование тематического PageRank.....	204
5.3.4. Вывод тем из слов	205
5.3.5. Упражнения к разделу 5.3	205

5.4. Ссылочный спам	206
5.4.1. Архитектура спам-фермы	206
5.4.2. Анализ спам-фермы	207
5.4.3. Борьба со ссылочным спамом	208
5.4.4. TrustRank	208
5.4.5. Спамная масса	209
5.4.6. Упражнения к разделу 5.4	210
5.5. Хабы и авторитетные страницы	210
5.5.1. Предположения, лежащие в основе HITS	211
5.5.2. Формализация хабов и авторитетных страниц	211
5.5.3. Упражнения к разделу 5.5	214
5.6. Резюме	214
5.7. Список литературы	218

ГЛАВА 6.

Частые предметные наборы 219

6.1. Модель корзины покупок	219
6.1.1. Определение частого предметного набора	220
6.1.2. Применения частых предметных наборов	221
6.1.3. Ассоциативные правила	223
6.1.4. Поиск ассоциативных правил с высокой достоверностью	225
6.1.5. Упражнения к разделу 6.1	225
6.2. Корзины покупок и алгоритм Apriori	226
6.2.1. Представление данных о корзинах покупок	227
6.2.2. Использование оперативной памяти для подсчета предметных наборов	228
6.2.3. Монотонность предметных наборов	230
6.2.4. Доминирование подсчета пар	230
6.2.5. Алгоритм Apriori	231
6.2.6. Применение Apriori для поиска всех частых предметных наборов	232
6.2.7. Упражнения к разделу 6.2	235
6.3. Обработка больших наборов данных в оперативной памяти	236
6.3.1. Алгоритм Парка-Чена-Ю (PCY)	236
6.3.2. Многоэтапный алгоритм	238
6.3.3. Многохэшевый алгоритм	240
6.3.4. Упражнения к разделу 6.3	242
6.4. Алгоритм с ограниченным числом проходов	244
6.4.1. Простой рандомизированный алгоритм	244
6.4.2. Предотвращение ошибок в алгоритмах формирования выборки	245
6.4.3. Алгоритм SON	246
6.4.4. Алгоритм SON и MapReduce	247
6.4.5. Алгоритм Тойвонена	248
6.4.6. Почему алгоритм Тойвонена работает	249
6.4.7. Упражнения к разделу 6.4	249
6.5. Подсчет частых предметных наборов в потоке	250
6.5.1. Методы выборки из потока	250

6.5.2. Частые предметные наборы в затухающих окнах	251
6.5.3. Гибридные методы	253
6.5.4. Упражнения к разделу 6.5	253
6.6. Резюме	254
6.7. Список литературы	256

ГЛАВА 7.

Кластеризация.....258

7.1. Введение в методы кластеризации	258
7.1.1. Точки, пространства и расстояния	258
7.1.2. Стратегии кластеризации	260
7.1.3. Проклятие размерности.....	260
7.1.4. Упражнения к разделу 7.1	262
7.2. Иерархическая кластеризация.....	262
7.2.1. Иерархическая кластеризация в евклидовом пространстве.....	263
7.2.2. Эффективность иерархической кластеризации	265
7.2.3. Альтернативные правила управления иерархической кластеризацией	266
7.2.4. Иерархическая кластеризация в неевклидовых пространствах.....	268
7.2.5. Упражнения к разделу 7.2	269
7.3. Алгоритм k средних	270
7.3.1. Основы алгоритма k средних	270
7.3.2. Инициализация кластеров в алгоритме k средних.....	271
7.3.3. Выбор правильного значения k	272
7.3.4. Алгоритм Брэдли-Файяда-Рейна	273
7.3.5. Обработка данных в алгоритме BFR.....	275
7.3.6. Упражнения к разделу 7.3	277
7.4. Алгоритм CURE	278
7.4.1. Этап инициализации в CURE.....	278
7.4.2. Завершение работы алгоритма CURE	279
7.4.3. Упражнения к разделу 7.4	280
7.5. Кластеризация в неевклидовых пространствах.....	280
7.5.1. Представление кластеров в алгоритме GRGPF	281
7.5.2. Инициализация дерева кластеров	281
7.5.3. Добавление точек в алгоритме GRGPF	282
7.5.4. Разделение и объединение кластеров	283
7.5.5. Упражнения к разделу 7.5	285
7.6. Кластеризация для потоков и параллелизм	285
7.6.1. Модель потоковых вычислений.....	285
7.6.2. Алгоритм кластеризации потока	286
7.6.3. Инициализация интервалов	286
7.6.4. Объединение кластеров	287
7.6.5. Ответы на вопросы	289
7.6.6. Кластеризация в параллельной среде	290
7.6.7. Упражнения к разделу 7.6	290
7.7. Резюме	290

7.8. Список литературы	294
------------------------------	-----

ГЛАВА 8.

Реклама в Интернете.....295

8.1. Проблемы онлайн-рекламы	295
8.1.1. Возможности рекламы.....	295
8.1.2. Прямое размещение рекламы.....	296
8.1.3. Акцидентные объявления.....	297
8.2. Онлайн-алгоритмы	298
8.2.1. Онлайн- и офлайн-алгоритмы	298
8.2.2. Жадные алгоритмы.....	299
8.2.3. Коэффициент конкурентоспособности	300
8.2.4. Упражнения к разделу 8.2	300
8.3. Задача о паросочетании	301
8.3.1. Паросочетания и совершенные паросочетания	301
8.3.2. Жадный алгоритм нахождения максимального паросочетания	302
8.3.3. Коэффициент конкурентоспособности жадного алгоритма паросочетания	303
8.3.4. Упражнения к разделу 8.3	304
8.4. Задача о ключевых словах.....	304
8.4.1. История поисковой рекламы.....	304
8.4.2. Постановка задачи о ключевых словах	305
8.4.3. Жадный подход к задаче о ключевых словах	306
8.4.4. Алгоритм Balance.....	307
8.4.5. Нижняя граница коэффициента конкурентоспособности в алгоритме Balance	308
8.4.6. Алгоритм Balance при большом числе участников аукциона.....	310
8.4.7. Обобщенный алгоритм Balance	311
8.4.8. Заключительные замечания по поводу задачи о ключевых словах.....	312
8.4.9. Упражнения к разделу 8.4	313
8.5. Реализация алгоритма Adwords	313
8.5.1. Сопоставление предложений с поисковыми запросами	314
8.5.2. Более сложные задачи сопоставления.....	314
8.5.3. Алгоритм сопоставления документов и ценовых предложений	315
8.6. Резюме	318
8.7. Список литературы	320

ГЛАВА 9.

Рекомендательные системы321

9.1. Модель рекомендательной системы	321
9.1.1. Матрица предпочтений.....	322
9.1.2. Длинный хвост	323
9.1.3. Применения рекомендательных систем.....	323
9.1.4. Заполнение матрицы предпочтений	325
9.2. Рекомендации на основе фильтрации содержимого	326

9.2.1. Профили объектов	326
9.2.2. Выявление признаков документа	327
9.2.3. Получение признаков объектов из меток	328
9.2.4. Представление профиля объекта.....	329
9.2.5. Профили пользователей	330
9.2.6. Рекомендование объектов пользователям на основе содержимого	331
9.2.7. Алгоритм классификации	332
9.2.8. Упражнения к разделу 9.2	335
9.3. Коллаборативная фильтрация	336
9.3.1. Измерение сходства	336
9.3.2. Двойственность сходства	339
9.3.3. Кластеризация пользователей и объектов	340
9.3.4. Упражнения к разделу 9.3	341
9.4. Понижение размерности	342
9.4.1. UV-декомпозиция	343
9.4.2. Среднеквадратичная ошибка.....	343
9.4.3. Инкрементное вычисление UV-декомпозиции	344
9.4.4. Оптимизация произвольного элемента.....	347
9.4.5. Построение полного алгоритма UV-декомпозиции	348
9.4.6. Упражнения к разделу 9.4	351
9.5. Задача NetFlix	351
9.6. Резюме	353
9.7. Список литературы	355

ГЛАВА 10.

Анализ графов социальных сетей 356

10.1. Социальные сети как графы	356
10.1.1. Что такое социальная сеть?	357
10.1.2. Социальные сети как графы	357
10.1.3. Разновидности социальных сетей.....	358
10.1.4. Графы с вершинами нескольких типов	360
10.1.5. Упражнения к разделу 10.1	361
10.2. Кластеризация графа социальной сети.....	361
10.2.1. Метрики для графов социальных сетей.....	361
10.2.2. Применение стандартных методов кластеризации	362
10.2.3. Промежуточность	363
10.2.4. Алгоритм Гирвана-Ньюмана.....	364
10.2.5. Использование промежуточности для нахождения сообществ.....	366
10.2.6. Упражнения к разделу 10.2	368
10.3. Прямое нахождение сообществ	368
10.3.1. Нахождение клик	368
10.3.2. Полные двудольные графы	369
10.3.3. Нахождение полных двудольных подграфов	370
10.3.4. Почему должны существовать полные двудольные графы	370
10.3.5. Упражнения к разделу 10.3	372

10.4. Разрезание графов	373
10.4.1. Какое разрезание считать хорошим?	373
10.4.2. Нормализованные разрезы	374
10.4.3. Некоторые матрицы, описывающие графы	374
10.4.4. Собственные значения матрицы Лапласа	375
10.4.5. Другие методы разрезания	378
10.4.6. Упражнения к разделу 10.4	379
10.5. Нахождение пересекающихся сообществ	379
10.5.1. Природа сообществ	379
10.5.2. Оценка максимального правдоподобия	380
10.5.3. Модель графа принадлежности	382
10.5.4. Как избежать дискретных изменений членства	384
10.5.5. Упражнения к разделу 10.5	385
10.6. Simrank	386
10.6.1. Случайные блуждания в социальном графе	386
10.6.2. Случайное блуждание с перезапуском	387
10.6.3. Упражнения к разделу 10.6	389
10.7. Подсчет треугольников	390
10.7.1. Зачем подсчитывать треугольники?	390
10.7.2. Алгоритм нахождения треугольников	390
10.7.3. Оптимальность алгоритма нахождения треугольников	392
10.7.4. Нахождение треугольников с помощью MapReduce	392
10.7.5. Использование меньшего числа редукторов	394
10.7.6. Упражнения к разделу 10.7	395
10.8. Окрестности в графах	396
10.8.1. Ориентированные графы и окрестности	396
10.8.2. Диаметр графа	397
10.8.3. Транзитивное замыкание и достижимость	399
10.8.4. Вычисление транзитивного замыкания с помощью MapReduce	399
10.8.5. Интеллектуальное транзитивное замыкание	402
10.8.6. Транзитивное замыкание посредством сокращения графа	403
10.8.7. Аппроксимация размеров окрестностей	405
10.8.8. Упражнения к разделу 10.8	407
10.9. Резюме	408
10.10. Список литературы	411

ГЛАВА 11.

Понижение размерности414

11.1. Собственные значения и собственные векторы.....	414
11.1.1. Определения	415
11.1.2. Вычисление собственных значений и собственных векторов	415
11.1.3. Нахождение собственных пары степенным методом	417
11.1.4. Матрица собственных векторов	420
11.1.5. Упражнения к разделу 11.1	421
11.2. Метод главных компонент	422
11.2.1. Иллюстративный пример	422

11.2.2. Использование собственных векторов для понижения размерности	425
11.2.3. Матрица расстояний	426
11.2.4. Упражнения к разделу 11.2	427
11.3. Сингулярное разложение	427
11.3.1. Определение сингулярного разложения	428
11.3.2. Интерпретация сингулярного разложения	429
11.3.3. Понижение размерности с помощью сингулярного разложения	431
11.3.4. Почему обнуление малых сингулярных значений работает	432
11.3.5. Запросы с использованием концептов	434
11.3.6. Вычисление сингулярного разложения матрицы	434
11.3.7. Упражнения к разделу 11.3	435
11.4. CUR-декомпозиция	436
11.4.1. Определение CUR-декомпозиции	437
11.4.2. Правильный выбор строк и столбцов	438
11.4.3. Построение средней матрицы	440
11.4.4. Полная CUR-декомпозиция	441
11.4.5. Исключение дубликатов строк и столбцов	441
11.4.6. Упражнения к разделу 11.4	442
11.5. Резюме	442
11.6. Список литературы	444

ГЛАВА 12.

Машинное обучение на больших данных 446

12.1. Модель машинного обучения	447
12.1.1. Обучающие наборы	447
12.1.2. Пояснительные примеры	447
12.1.3. Подходы к машинному обучению	449
12.1.4. Архитектура машинного обучения	451
12.1.5. Упражнения к разделу 12.1	454
12.2. Перцептроны	454
12.2.1. Обучение перцептрона с нулевым порогом	455
12.2.2. Сходимость перцептронов	457
12.2.3. Алгоритм Winnow	458
12.2.4. Переменный порог	459
12.2.5. Многоклассовые перцептроны	461
12.2.6. Преобразование обучающего набора	462
12.2.7. Проблемы, связанные с перцептронами	463
12.2.8. Параллельная реализация перцептронов	464
12.2.9. Упражнения к разделу 12.2	466
12.3. Метод опорных векторов	466
12.3.1. Механизм метода опорных векторов	466
12.3.2. Нормировка гиперплоскости	468
12.3.3. Нахождение оптимальных приближенных разделителей	470
12.3.4. Нахождение решений в методе опорных векторов с помощью градиентного спуска	472

12.3.5. Стохастический градиентный спуск	476
12.3.6. Параллельная реализация метода опорных векторов	477
12.3.7. Упражнения к разделу 12.3	477
12.4. Обучение по ближайшим соседям	478
12.4.1. Инфраструктура для вычисления ближайших соседей	478
12.4.2. Обучение по одному ближайшему соседу	479
12.4.3. Обучение одномерных функций	480
12.4.4. Ядерная регрессия	482
12.4.5. Данные в многомерном евклидовом пространстве	483
12.4.6. Неевклидовы метрики	484
12.4.7. Упражнения к разделу 12.4	485
12.5. Сравнение методов обучения	486
12.6. Резюме	487
12.7. Список литературы	489
Предметный указатель	490



ПРЕДИСЛОВИЕ

В основу этой книги положен материал односеместрового курса, который Ананд Раджараман и Джефф Ульман в течение нескольких лет читали в Стэнфордском университете. Курс CS345A под названием «Добыча данных в вебе» задумывался как спецкурс для аспирантов, но оказался доступным и полезным также старшекурсникам. Когда в Стэнфорд пришел преподавать Юре Лесковец, мы существенно изменили организацию материала. Он начал читать новый курс CS224W по анализу сетей и расширил материал курса CS345A, который получил номер CS246. Втроем авторы также подготовили курс CS341, посвященный крупномасштабному проекту в области добычи данных. В своем теперешнем виде книга содержит материал всех трех курсов.

О чем эта книга

В самых общих словах, эта книга о добыче данных. Но акцент сделан на анализе данных очень большого объема, не помещающихся в оперативную память. Поэтому многие примеры относятся к вебу или к данным, полученным из веба. Кроме того, в книге принят алгоритмический подход: добыча данных – это применение алгоритмов к данным, а не использование данных для «обучения» той или иной машины. Ниже перечислены основные рассматриваемые темы.

1. Распределенные файловые системы и технология распределения-редукции (map-reduce) как средство создания параллельных алгоритмов, успешно справляющихся с очень большими объемами данных.
2. Поиск по сходству, в том числе такие важнейшие алгоритмы, как MinHash и хэширование с учетом близости (locality sensitive hashing).
3. Обработка потоков данных и специализированные алгоритмы для работы с данными, которые поступают настолько быстро, что либо обрабатываются немедленно, либо теряются.
4. Принципы работы поисковых систем, в том числе алгоритм Google PageRank, распознавание ссылочного спама и метод авторитетных и хаб-документов.
5. Частые предметные наборы, в том числе поиск ассоциативных правил, анализ корзины, алгоритм Apriori и его усовершенствованные варианты.
6. Алгоритмы кластеризации очень больших многомерных наборов данных.

7. Две важные для веб-приложений задачи: управление рекламой и рекомендательные системы.
8. Алгоритмы анализа структуры очень больших графов, в особенности графов социальных сетей.
9. Методы получения важных свойств большого набора данных с помощью понижения размерности, в том числе сингулярное разложение и латентно-семантическое индексирование.
10. Алгоритмы машинного обучения, применимые к очень большим наборам данных, в том числе перцептроны, метод опорных векторов и градиентный спуск.

Требования к читателю

Для полного понимания изложенного в книге материала мы рекомендуем:

1. Прослушать вводный курс по системам баз данных, включая основы SQL и сопутствующих систем программирования.
2. Иметь знания о структурах данных, алгоритмах и дискретной математике в объеме второго курса университета.
3. Иметь знания о программных системах, программной инженерии и языках программирования в объеме второго курса университета.

Упражнения

В книге много упражнений, они есть почти в каждом разделе. Более трудные упражнения или их части отмечены восклицательным знаком, а самые трудные – двумя восклицательными знаками.

Поддержка в вебе

Слайды, домашние задания, проектные требования и экзаменационные задачи из курсов, примыкающих к этой книге, можно найти по адресу <http://www.mmnds.org>.

Автоматизированные домашние задания

На основе этой книги составлены автоматизированные упражнения с применением системы проверочных вопросов Gradiance, доступной по адресу www.gradiance.com/services. Студенты могут стать членами открытой группы, создав на этом сайте учетную запись и присоединившись к группе с кодом 1EDD8A1D. Преподаватели также могут воспользоваться этим сайтом, для этого нужно создать учетную запись и отправить сообщение на адрес support@gradiance.com,

указав в нем свой логин, название учебного заведения и запрос на право использования материалов к книге (MMDS).

Благодарности

Мы благодарны Фото Афрати (Foto Afrati), Аруну Маратхи (Arun Marathe) и Року Сосику (Rok Sosic) за критическое прочтение рукописи.

Об ошибках также сообщали Раджив Абрахам (Rajiv Abraham), Апурв Агарвал (Apoorv Agarwal), Арис Анагностопулос (Aris Anagnostopoulos), Атилла Сонер Балкир (Atilla Soner Balkir), Арно Бельтуаль (Arnaud Belletoile), Робин Беннетт (Robin Bennett), Сьюзан Бьянкани (Susan Biancani), Амитабх Чаудхари (Amitabh Chaudhary), Леланд Чен (Leland Chen), Анастасиос Гунарис (Anastasios Gounaris), Шрей Гупта (Shrey Gupta), Валид Хамейд (Waleed Hameid), Саман Харати-заде (Saman Haratizadeh), Лаклан Канг (Lachlan Kang), Эд Кнопп (Ed Knorr), Хэй Вун Квак (Haewoon Kwak), Эллис Лау (Ellis Lau), Грег Ли (Greg Lee), Этан Лозано (Ethan Lozano), Ю Нань Люо (Yunan Luo), Майкл Махоуни (Michael Mahoney), Джастин Мейер (Justin Meyer), Брайант Москон (Bryant Moscon), Брэд Пенофф (Brad Penoff), Филипп Коко Прасетийо (Philips Kokoh Prasetyo), Ки Ге (Qi Ge), Рич Сейтер (Rich Seiter), Хитэш Шетти (Hitesh Shetty), Ангад Сингх (Angad Singh), Сандип Срипада (Sandeep Sripada), Дэннис Сидхарта (Dennis Sidharta), Кшиштоф Стенсел (Krzysztof Stencel), Марк Стопус (Mark Storus), Рошан Сумбалай (Roshan Sumbaly), Зак Тэйлор (Zack Taylor), Тим Триш мл. (Tim Triche Jr.), Вань Бин (Wang Bin), Вэнь Цзен Бин (Weng Zhen-Bin), Роберт Уэст (Robert West), Оскар Ву (Oscar Wu), Се Ке (Xie Ke), Николас Чжао (Nicolas Zhao) и Чжу Цзинь Бо (Zhou Jingbo). Разумеется, все оставшиеся незамеченными ошибки – наша вина.

Ю. Л.

А. Р.

Дж. Д. У.

Пало-Альто, Калифорния
март 2014



ГЛАВА 1.

Добыча данных

В этой вводной главе мы опишем, в чем состоит сущность добычи данных, и обсудим, как добыча данных трактуется в различных дисциплинах, которые вносят свой вклад в эту область. Мы рассмотрим «принцип Бонферрони», предупреждающий об опасностях чрезмерного увлечения добычей данных. В этой же главе мы кратко упомянем некоторые идеи, которые, хотя сами и не относятся к добыче данных, но полезны для понимания ряда важных идей, относящихся к этой тематике. Мы имеем в виду метрику важности слов TFIDF, поведение хэш-функций и индексов, а также некоторые тождества, содержащие число e , основание натуральных логарифмов. Наконец, мы расскажем о темах, рассматриваемых в этой книге.

1.1. Что такое добыча данных?

Многие разделяют определение «добычи данных» как выявление «моделей» данных. Однако под моделью можно понимать разные вещи. Ниже описываются наиболее важные направления моделирования.

1.1.1. Статистическое моделирование

Первыми термин «добыча данных» ввели в обиход специалисты по математической статистике. Первоначально словосочетание «data mining» (добыча данных) или «data dredging» (вычерпывание данных) имело несколько пренебрежительный оттенок и обозначало попытки извлечь информацию, которая явно не присутствовала в данных. В разделе 1.2 демонстрируются различные ошибки, которые могут возникнуть, если пытаться извлечь то, чего в данных на самом деле нет. В наши дни термин «добыча данных» употребляется в положительном смысле. Теперь статистики рассматривают добычу данных как средство построения статистической модели, т. е. закона, в соответствии с которым распределены видимые данные.

Пример 1.1. Пусть данными будет множество чисел. Эти данные намного проще тех, что подвергаются добыче, но для примера вполне подойдут. Статистик может предположить, что данные имеют гауссово распределение и по известным формулам вычислить наиболее вероятные параметры этого распределения.

Среднее и стандартное отклонение полностью определяют гауссово распределение и потому могут служить моделью данных.

1.1.2. Машинное обучение

Некоторые считают, что добыча данных и машинное обучение – синонимы. Безусловно, для добычи данных иногда используются алгоритмы, применяемые в машинном обучении. Специалисты по машинному обучению используют данные как обучающий набор и на них обучают алгоритм того или иного вида, например: байесовские сети, метод опорных векторов, решающие деревья, скрытые марковские модели и т. п.

В некоторых ситуациях использование данных подобным образом имеет смысл. В частности, машинное обучение дает хороший результат, когда мы плохо представляем себе, что искать в данных. Например, совсем неясно, из-за каких особенностей одним людям фильм нравится, а другим – нет. Поэтому принявшие «вызов Netflix» – изобрести алгоритм, который предсказывал бы оценку фильма пользователями на основе выборки из их прошлых ответов, – с большим успехом применили алгоритмы машинного обучения. Мы обсудим простую форму алгоритма такого типа в разделе 9.4.

С другой стороны, машинное обучение не приносит успеха в ситуациях, когда цели добычи данных можно описать более конкретно. Интересный пример – попытка компании WhizBang! Labs¹ использовать методы машинного обучения для поиска резюме, которые люди размещают в сети. У нее не получилось добиться результатов, лучших, чем дают вручную составленные алгоритмы, которые ищут очевидные слова и фразы, встречающиеся в типичном резюме. Всякий, кто читал или писал резюме, довольно отчетливо представляет, что в нем содержится, поэтому как выглядит веб-страница, содержащая резюме, – никакая не тайна. Потому-то применение машинного обучения не дало выигрыша по сравнению с составленным в лоб алгоритмом распознавания резюме.

1.1.3. Вычислительные подходы к моделированию

Сравнительно недавно на добычу данных стали смотреть как на алгоритмическую задачу. В этом случае модель данных – это просто ответ на сложный запрос к данным. Например, если дано множество чисел, как в примере 1.1, то можно было бы вычислить их среднее и стандартное отклонение. Отметим, что эти значения обязательно являются параметрами гауссова распределения, которое лучше всего аппроксимирует данные, хотя при достаточно большом наборе данных они почти наверняка будут близки к ним.

Есть много подходов к моделированию данных. Мы уже упомянули одну возможность: построить статистический процесс, с помощью которого данные могли

¹ Эта компания пыталась использовать методы машинного обучения для анализа очень большого объема данных и наняла для этого много высококлассных специалистов. К сожалению, выжить ей не удалось.

быть сгенерированы. Большинство прочих подходов к моделированию можно отнести к одной из двух категорий.

1. Краткое и приближенное обобщение данных или
2. Извлечение из данных наиболее существенных признаков с отбрасыванием всего остального.

В следующих разделах мы исследуем оба подхода.

1.1.4. Обобщение

Одна из самых интересных форм обобщения – идея алгоритма PageRank, так успешно примененная Google; мы будем рассматривать ее в главе 5. При такой форме добычи данных вся сложная структура веба сводится к одному числу для каждой страницы. Несколько упрощая, это число, «ранг страницы» (PageRank), можно описать как вероятность того, что пользователь, случайно обходящий граф, окажется на этой странице в любой заданный момент времени. Замечательное свойство такого ранжирования заключается том, что оно очень хорошо отражает «важность» страницы – в какой мере типичный пользователь поисковой системы хотел бы видеть данную страницу в ответе на свой запрос.

Еще один важный вид обобщения – кластеризация – будет рассмотрен в главе 7. В этом случае данные рассматриваются как точки в многомерном пространстве. Те точки, которые в некотором смысле «близки», помещаются в один кластер. Сами кластеры также обобщаются, например, путем указания центроида кластера и среднего расстояния от центроида до всех точек. Совокупность обобщенных характеристик кластеров становится обобщением всего набора данных.

Пример 1.2. Знаменитый пример применения кластеризации для решения задачи имел место много лет назад в Лондоне, когда никаких компьютеров еще не было². Врач Джон Сноу, сражаясь со вспышкой холеры, нанес места проживания заболевших на карту города. На рис. 1.1 показана упрощенная иллюстрация этой процедуры.

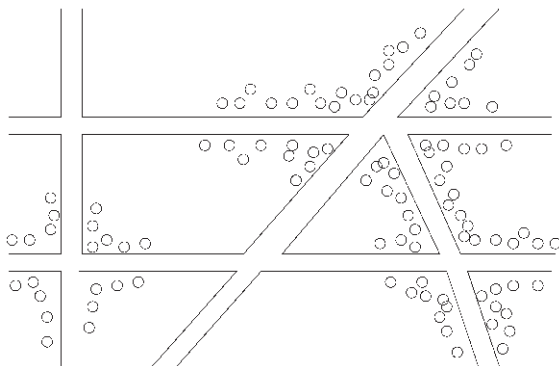


Рис. 1.1. Случаи холеры на карте Лондона

² См. http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Как видно, образовалось несколько кластеров в районе перекрестков. На этих перекрестках находились зараженные водоразборные колонки; жившие поблизости от них заболели, те же, кто жил рядом с незараженными колонками, остались здоровы. Не будь возможности кластеризовать данные, причина холеры осталась бы невыясненной.

1.1.5. Выделение признаков

В типичной модели на основе признаков ищутся экстремальные примеры некоторого явления, и данные представляются с помощью этих примеров. Если вы знакомы с *байесовскими сетями*, одной из ветвей машинного обучения, которая в этой книге не рассматривается, то знаете, что в них сложные связи между объектами представляются с помощью отыскания самых сильных статистических зависимостей и использования только их для представления всех статистических связей. Мы изучим следующие важные формы выделения признаков из больших наборов данных.

1. *Частые предметные наборы*. Эта модель имеет смысл, когда данные состоят из «корзин», содержащих небольшие наборы предметов, как, например, в задаче об анализе корзин покупок, обсуждаемой в главе 6. Мы ищем небольшие наборы предметов, которые встречаются вместе во многих корзинах, и считаем эти «частые предметные наборы» искомой характеристикой данных. Первоначально такой вид добычи данных применялся к настоящим корзинам покупок: поиску предметов, например гамбургер и кетчуп, которые люди покупают вместе в небольшой лавке или в супермаркете.
2. *Похожие предметы*. Часто данные имеют вид коллекции наборов, а цель состоит в том, чтобы найти пары наборов, в которых относительно много общих элементов. Например, покупателей в интернет-магазине типа Amazon можно рассматривать как наборы купленных ими товаров. Чтобы предложить покупателю еще что-нибудь, что могло бы ему понравиться, Amazon может искать «похожих» покупателей и порекомендовать товары, которые покупали многие из них. Этот процесс называется «коллаборативной фильтрацией». Если бы все покупатели были целеустремленными, т. е. покупали бы только одну вещь, то могла бы сработать кластеризация покупателей. Но обычно покупателей интересуют разные вещи, поэтому полезнее для каждого покупателя найти небольшое число покупателей со схожими вкусами и представить данные такими связями. Проблему сходства мы будем обсуждать в главе 3.

1.2. Статистические пределы добычи данных

Типичная задача добычи данных – обнаружение необычных событий, скрытых в массивном объеме данных. В этом разделе мы рассмотрим эту проблему и заодно

«принцип Бонферрони» – предостережение против излишне ревностных попыток добыть данные.

1.2.1. Тотальное владение информацией

В 2002 году администрация Буша выдвинула план – подвергнуть анализу все данные, до которых можно дотянуться, в том числе чеки, оплаченные кредитной картой, данные о регистрации в гостиницах, данные о поездках и многие иные виды информации, – с целью отслеживания террористической деятельности. Эта идея, естественно, вызвала недовольство у поборников защиты частной жизни, и в итоге весь проект, названный ТИА, или *Total Information Awareness (тотальное владение информацией)*, был похоронен Конгрессом, хотя не исключено, что он все же существует под другим именем. В этой книге мы не собираемся обсуждать трудную проблему поиска компромисса между безопасностью и конфиденциальностью. Однако в связи с проектом ТИА или подобной системой возникает ряд технических вопросов касательно практической осуществимости и реалистичности предположений.

Многие задавались вопросом: если исследовать так много данных, пытаясь найти следы деятельности, характерной для террористов, то не получится ли, что мы найдем много совершенно невинных действий – или даже незаконных, но не относящихся к терроризму, – и человеку придется свести знакомство с полицией, а, может, и не просто знакомство? Здесь все зависит от того, насколько узко определена интересующая нас деятельность. Статистики сталкивались с многообразными проявлениями этой проблемы и выдвинули теорию, начатки которой мы изложим в следующем разделе.

1.2.2. Принцип Бонферрони

Пусть имеются какие-то данные, и мы ищем в них события определенного вида. Можно ожидать, что такие события встретятся, даже если данные выбраны абсолютно случайно, а количество событий будет расти вместе с объемом данных. Эти события «фиктивные» в том смысле, что у них нет никакой причины, помимо случайности данных, а в случайных данных всегда встретится какое-то количество необычных признаков, которые, хотя и выглядят значимыми, на самом деле таковыми не являются. Теорема математической статистики, известная под названием *поправка Бонферрони*, дает статистически корректный способ избежать большинства таких ложноположительных ответов на поисковый запрос. Не вдаваясь в технические детали, мы предложим ее неформальный вариант, *принцип Бонферрони*, который поможет избежать трактовки случайных фактов как реальных. Вычислите ожидаемое число искомых событий в предположении, что данные случайны. Если это число существенно больше количества реальных событий, которые вы надеетесь обнаружить, то следует ожидать, что почти все найденные события фиктивные, т. е. являются статистическими артефактами, а не свидетельством в пользу того, что вы ищете. Это наблюдение и есть неформальный принцип Бонферрони.

В случае поиска террористов, когда мы ожидаем, что сколько-то террористов действуют в любой момент времени, принцип Бонферрони гласит, что обнаружить террористов можно, только выискивая события настолько редкие, что в случайных данных их появление крайне маловероятно. Развернутый пример мы приведем в следующем разделе.

1.2.3. Пример применения принципа Бонферрони

Допустим, мы полагаем, что где-то действуют «злоумышленники», и хотим их обнаружить. Допустим также, что есть основания полагать, что злоумышленники периодически встречаются в гостинице, чтобы спланировать свой злой умысел. Сделаем следующие предположения о размере задачи:

1. Есть миллиард людей, среди которых могут быть злоумышленники.
2. Любой человек останавливается в гостинице один день из 100.
3. Гостиница вмещает 100 человек. Следовательно, 100 000 гостиниц будет достаточно, чтобы разместить 1 % от миллиарда людей, которые останавливаются в гостинице в каждый конкретный день.
4. Мы изучаем данные о регистрации в гостиницах за 1000 дней.

Чтобы найти в этих данных злоумышленников, мы будем искать людей, которые в два разных дня останавливались в одной и той же гостинице. Допустим, однако, что в действительности никаких злоумышленников нет. То есть все ведут себя случайным образом, с вероятностью 0,01 решая в данный день остановиться в какой-то гостинице и при этом случайно выбирая одну из 10^5 гостиниц. Найдем ли мы пары людей, которые выглядят как злоумышленники? Можно выполнить простое вычисление. Вероятность того, что два произвольных человека решат остановиться в гостинице в данный день, составляет 0,0001. Вероятность того, что они остановятся в одной и той же гостинице в один и тот же день равна 10^{-9} . Вероятность, что они остановятся в одной и той же гостинице в два разных дня, равна квадрату этого числа, т. е. 10^{-18} . Отметим, что выбранные в эти дни гостиницы могут быть разными.

Теперь надо посчитать, сколько событий указывают на злой умысел. Под «событием» здесь понимается пара людей и пара дней такие, что оба человека в каждый из этих двух дней останавливались в одной и той же гостинице. Чтобы упростить вычисления, заметим, что для больших n $\binom{n}{2}$ приблизительно равно $n^2/2$. Таким образом, количество пар людей равно $\binom{10^9}{2} = 5 \times 10^{17}$. Количество пар дней равно $\binom{1000}{2} = 5 \times 10^5$. Ожидаемое число событий, выглядящих как злоумышление, равно произведению количества пар людей на количество пар дней и на вероятность того, что пара людей и пара дней демонстрируют искомое поведение. Это число равно

$$5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} = 250\,000.$$

То есть четверть миллиона людей будут казаться злоумышленниками, даже если не являются таковыми.

Теперь предположим, что в действительности существует 10 пар злоумышленников. Полиции придется проверить четверть миллиона других пар, чтобы найти настоящих злоумышленников. Мало того что это означает вторжение в частную жизнь полумиллиона ни в чем неповинных людей, так еще и объем работы настолько велик, что такой подход к поиску злоумышленников практически неосуществим.

1.2.4. Упражнения к разделу 1.2

Упражнение 1.2.1. Используя сведения из раздела 1.2.3, найдите количество подозрительных пар, если внести в данные следующие изменения (сохранив все прочие числа)?

- (a) Увеличить количество дней наблюдения до 2000.
- (б) Увеличить количество наблюдаемых людей до 2 миллиардов (а количество гостиниц соответственно до 200 000).
- (c) Считать двух человек подозрительными, если они останавливались в одной и той же гостинице в три разных дня.

! Упражнение 1.2.2. Предположим, что у нас есть информация о покупках 100 миллионов людей в супермаркетах. Каждый человек заходит в супермаркет 100 раз в год и покупает 10 из 1000 предлагаемых там товаров. Мы думаем, что двое террористов в какой-то день на протяжении года купят в точности один и тот же набор предметов (быть может, компонентов бомбы). Если мы будем искать пары людей, купивших одинаковые наборы предметов, то можно ли ожидать, что найденные люди действительно террористы?³

1.3. Кое-какие полезные сведения

В этом разделе содержится краткое введение в темы, с которыми вы, возможно, знакомились на других курсах. Все эти сведения будут полезны при изучении добычи данных.

1. Мера важности слов TF.IDF.
2. Хэш-функции и их применение.
3. Внешняя память (диск) и ее влияние на время работы алгоритмов.
4. Основание натуральных логарифмов e и тождества, содержащие эту константу.
5. Степенные зависимости.

1.3.1. Важность слов в документах

В нескольких приложениях добычи данных мы столкнемся с проблемой классификации документов (последовательностей слов) по тематике. Как правило, тема определяется путем поиска специальных слов, которые характеризуют относя-

³ То есть наша гипотеза состоит в том, что террористы наверняка купят набор из десяти одинаковых предметов в какой-то день на протяжении года. Мы не хотим обсуждать вопрос о том, характерно ли такое поведение для настоящих террористов.

щиеся к ней документы. Например, в статьях о бейсболе будут часто встречаться слова «мяч», «бита», «бросок», «пробежка» и т. д. После того как документы отнесены к теме бейсбола, нетрудно заметить, что подобные слова встречаются в них аномально часто. Но пока классификация не произведена, выделить эти слова как характеристические невозможно.

Таким образом, классификация часто начинается с изучения документов и отыскания в них важных слов. Первая гипотеза может состоять в том, что слова, которые чаще всего встречаются в документе, и есть самые важные. Но в данном случае интуиция подсказывает ответ, прямо противоположный истинному положению дел. Самыми частыми, конечно же, будут наиболее употребительные слова типа «the» и «and», которые помогают выразить мысль, но сами по себе не несут никакого смысла. И действительно, несколько сотен наиболее употребительных слов английского языка (они называются *stop-словами*), обычно исключаются из документов еще до попытки классификации.

На самом деле, индикаторами темы являются относительно редко встречающиеся слова. Но не все редкие слова одинаково полезны в качестве индикаторов. Некоторые слова, например «notwithstanding» (несмотря на) или «albeit» (пусть даже), хоть редко встречаются в коллекции документов, но ничего полезного не сообщают. С другой стороны, слово «chukker» (период в игре в поло), пожалуй, встречается не менее редко, но подсказывает, что данный документ посвящен игре в поло. Разница между значимыми и незначимыми редкими словами определяется концентрацией полезных слов в немногих документах. То есть присутствие в документе слова типа «albeit» не повышает вероятность его повторного появления. Но если в статье один раз встречается слово «chukker», то весьма вероятно, что оно входит в составе словосочетания «first chukker» (первый период), еще раз в «second chukker» (второй период) и т. д. То есть, если слово вообще встречается, то с большой вероятностью оно встретится несколько раз.

Формальная мера концентрации данного слова в относительно небольшом количестве документов называется TF.IDF (частота термина, помноженная на обратную частоту документа). Обычно она вычисляется следующим образом. Пусть есть коллекция из N документов. Обозначим f_{ij} частоту (количество вхождений) термина (слова) i в документ j и определим частоту термина TF_{ij} такой формулой:

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}.$$

Иначе говоря, частота термина i в документе j равна величине f_{ij} , нормированной путем деления на максимальное количество вхождений этого термина (возможно, после исключения стоп-слов) в один и тот же документ. Следовательно, у самого часто встречающегося термина в документе j TF будет равно 1, а у всех остальных меньше.

IDF термина определяется следующим образом. Пусть терм i встречается в n_i документах из коллекции, содержащей всего N документов. Тогда $IDF_i = \log_2(N/n_i)$. Оценка TF.IDF для термина i в документе j определяется как $TF_{ij} \times IDF_i$. Именно тер-

мы с наибольшей оценкой TF.IDF часто наилучшим образом характеризуют тему документа.

Пример 1.3. Предположим, что репозиторий содержит $2^{20} = 1048576$ документов. Пусть слово w встречается в $2^{10} = 1024$ документов. Тогда $IDF_w = \log_2(2^{20}/2^{10}) = \log_2(2^{10}) = 10$. Рассмотрим документ j , в котором терм w встречается 20 раз, и пусть это максимальное количество вхождений одного слова (возможно, после исключения стоп-слов). Тогда $TF_{wj} = 1$ и оценка TF.IDF терма w в документе j равна 10.

Предположим, что в документе k слово w встречается один раз, тогда как максимальное количество вхождений одного слова в этом документе равно 20. Тогда $TF_{wk} = 1/20$, а оценка TF.IDF для w в документе k равна $1/2$.

1.3.2. Хэш-функции

Вы, вероятно, слышали о хэш-таблицах и, возможно, использовали их в Java-классах или других подобных пакетах. Хэш-функции, лежащие в основе хэш-таблиц, находят важное применение и во многих алгоритмах добычи данных, в которых хэш-таблицы принимают необычную форму. В этом разделе мы рассмотрим основные понятия.

Прежде всего, хэш-функция h принимает *ключ хэширования* в качестве аргумента и возвращает *номер ячейки (bucket)*. Номер ячейки – это целое число, обычно в диапазоне от 0 до $B - 1$, где B – количество ячеек. Тип ключа хэширования может быть любым. На интуитивном уровне хэш-функция «рандомизирует» ключи хэширования. Точнее, если ключи хэширования случайным образом выбираются из разумной совокупности возможных ключей, то h поместит в каждую из B ячеек примерно одинаковое количество ключей. Это было бы невозможно, если, к примеру, размер совокупности ключей хэширования меньше B . Такая совокупность не считается «разумной». Однако есть немало более тонких причин, по которым хэш-функция может не давать приблизительно равномерного распределения по ячейкам.

Пример 1.4. Допустим, что ключи хэширования – положительные целые числа. Простая и употребительная хэш-функция $h(x) = x \bmod B$ – возвращает остаток от деления x на B . Она неплохо работает, если совокупность ключей хэширования – множество всех положительных целых чисел. Тогда доля ключей, попавших в каждую ячейку, составит $1/B$. Но предположим, что наша совокупность содержит только четные числа и пусть $B = 10$. Тогда значениями $h(x)$ могут быть только ячейки с номерами 0, 2, 4, 6, 8, так что поведение хэш-функции заведомо не случайно. С другой стороны, если взять $B = 11$, то окажется, что доля четных чисел в каждой из 11 ячеек равна $1/11$, т. е. хэш-функция работает очень хорошо.

Обобщая пример 1.4, можно сказать, что если ключами хэширования являются целые числа, то при выборе в качестве B числа, имеющего общий множитель со

всеми возможными ключами (или хотя бы с их большинством), распределение по ячейкам будет неравномерным. Поэтому обычно в качестве B берут простое число. При таком выборе снижаются шансы неслучайного поведения, хотя по-прежнему необходимо рассмотреть случай, когда все ключи делятся на B . Разумеется, есть много других типов хэш-функций, не зависящих от арифметики по модулю. Мы не станем пытаться систематизировать их здесь, но приведем несколько источников в списке литературы.

А что, если ключи хэширования не являются целыми числами? Вообще говоря, у любого типа данных имеется значение, состоящее из битов, а последовательность битов всегда можно интерпретировать как целое число. Однако существуют простые правила, позволяющие преобразовать наиболее употребительные типы в целые числа. Например, если ключ хэширования – строка, то мы можем преобразовать каждый символ в его значение в кодировке ASCII или Unicode, которое можно интерпретировать как небольшое целое число. Затем, перед делением на B , эти числа складываются. Если B меньше типичной суммы кодов символов для некоторой совокупности строк, то распределение по ячейкам будет близко к равномерному. Если же B больше, то мы можем разбить все символы строки на несколько групп. Затем конкатенируем коды символов в одной группе и рассматриваем результат как одно целое число. Складываем все получившиеся таким образом числа и делим на B , как и раньше. Например, если B порядка миллиарда, т. е. 2^{30} , то группировка символов по четыре даст 32-разрядные целые числа. Их суммы распределяются по миллиарду ячеек приблизительно равномерно.

Эта идея рекурсивно обобщается на более сложные типы данных.

- Если тип является записью, каждая компонента которой имеет свой тип, то рекурсивно преобразовать значение каждой компоненты в целое число, применяя алгоритм, соответствующий типу компоненты. Сложить целые числа, получившиеся для всех компонент, и преобразовать сумму в номер ячейки, разделив ее на B .
- Если тип является массивом, множеством или коллекцией элементов одного и того же типа, то преобразовать значения его элементов в целые числа, сложить результаты и разделить сумму на B .

1.3.3. Индексы

Индекс – это структура данных, которая позволяет эффективно находить объект по значениям одного или нескольких его элементов. Наиболее типична ситуация, когда объекты являются записями, а индекс строится по одному из полей каждой записи. Если известно значение v , то индекс позволяет найти все записи, в которых это поле имеет такое значение. Например, мы можем располагать файлом, содержащим тройки (имя, адрес, телефон), и построить индекс по полю «телефон». Зная номер телефона, мы можем с помощью индекса быстро найти одну или несколько записей с таким номером.