

# Оглавление

<b>Предисловие</b> .....	<b>11</b>
<b>Вступление</b> .....	<b>13</b>
<b>Благодарности</b> .....	<b>15</b>
<b>Об этой книге</b> .....	<b>17</b>
Кому адресована книга.....	17
Краткое содержание.....	17
Об исходном коде.....	18
Автор в сети.....	19
Другие онлайн-ресурсы .....	19
<b>Об авторах</b> .....	<b>21</b>
<b>Об иллюстрации на обложке</b> .....	<b>21</b>
<b>Глава 1. Задача релевантного поиска</b> .....	<b>22</b>
1.1. Ваша цель: стать специалистом по релевантности .....	23
1.2. Сложности релевантного поиска.....	24
1.2.1. Какой результат можно назвать «релевантным»? .....	25
1.2.2. Поиск: не существует панацеи от всех бед! .....	27
1.3. Анализ релевантности.....	28
1.3.1. Информационный поиск.....	29
1.3.2. Можно ли использовать достижения информационного поиска для решения задачи релевантности? .....	30
1.4. Как решается проблема релевантности? .....	33
1.5. Не только технологии: кураторство, сотрудничество и обратная связь .....	36
1.6. В заключение .....	38
<b>Глава 2. Поиск – взгляд изнутри</b> .....	<b>40</b>
2.1. Простейший поиск.....	41
2.1.1. Что такое документ с точки зрения поиска?.....	42
2.1.2. Поиск по содержимому.....	42
2.1.3. Исследование содержимого в процессе поиска .....	44
2.2. Структуры данных механизма поиска .....	46
2.2.1. Обратный индекс.....	46
2.2.2. Другие элементы обратного индекса .....	48
2.3. Индексирование содержимого: извлечение, обогащение, анализ и индексирование.....	49
2.3.1. Извлечение содержимого в документы.....	51
2.3.2. Обогащение документов: чистка, добавление и объединение данных.....	52
2.3.3. Анализ .....	53
2.3.4. Индексирование .....	57
2.4. Поиск и извлечение документов.....	58

2.4.1. Логический поиск: И/ИЛИ/НЕ .....	58
2.4.2. Логические запросы в Lucene (ДОЛЖЕН/НЕ_ДОЛЖЕН/МОЖЕТ) .....	60
2.4.3. Сопоставление фраз и учет позиций терминов.....	61
2.4.4. Побуждение к исследованиям: фильтрация, категоризация и агрегирование .....	62
2.4.5. Сортировка, ранжирование результатов и релевантность.....	63
2.5. В заключение.....	66
<b>Глава 3. Отладка первой проблемы релевантного поиска .....</b>	<b>68</b>
3.1. Приложения для Solr и Elasticsearch: примеры в Elasticsearch .....	69
3.2. Наш главный набор данных: TMDB.....	70
3.3. Примеры на языке Python.....	71
3.4. Первое поисковое приложение.....	71
3.4.1. Первые попытки поиска в индексе TMDB .....	75
3.5. Отладка сопоставления запросов .....	77
3.5.2. Анализ запроса.....	79
3.5.3. Отладка анализа для решения проблем сопоставления .....	80
3.5.4. Сопоставление запроса с обратным индексом.....	83
3.5.5. Исправление проблем сопоставления заменой анализаторов .....	84
3.6. Отладка ранжирования .....	87
3.6.1. Объяснение формулы оценки релевантности с помощью функции explain в Lucene .....	88
3.6.2. Векторная модель, объяснение релевантности и вы.....	93
3.6.3. Практические аспекты применения векторной модели .....	96
3.6.4. Оценка совпадений для измерения релевантности.....	98
3.6.5. Вычисление весов с использованием метрики TF × IDF .....	99
3.6.6. Ложь, ужасная ложь и сходство .....	101
3.6.7. Учет важности термина.....	103
3.6.8. Исправление оценки важности термина alien в описании «Space Jam» .....	103
3.7. Проблема решена? Наша работа никогда не заканчивается!.....	106
3.8. В заключение.....	107
<b>Глава 4. Угрошение лексем .....</b>	<b>109</b>
4.1. Лексемы, как признаки документов.....	109
4.1.1. Процесс сопоставления .....	111
4.1.2. Лексемы – больше, чем слова .....	111
4.2. Управление точностью и полнотой.....	112
4.2.1. Точность и полнота на примере .....	112
4.2.2. Анализ для точности и полноты .....	115
4.2.3. Доведение полноты до крайности .....	119
4.3. Точность и полнота – совмещение несовместимого .....	121
4.3.1. Оценка силы признака в единственном поле .....	122
4.3.2. Кроме TF × IDF: поиск по нескольким терминам и полям .....	125
4.4. Стратегии анализа.....	126
4.4.1. Обработка разделителей.....	127
4.4.2. Передача смысла с применением синонимов .....	130
4.4.3. Моделирование специфичности.....	134

4.4.4. Моделирование специфичности с синонимами.....	134
4.4.5. Моделирование специфичности построением путей .....	138
4.4.6. Лексемизируем мир!.....	139
4.4.7. Лексемизация целых чисел.....	140
4.4.8. Лексемизация географических данных .....	141
4.4.9. Лексемизация мелодий.....	143
4.5. В заключение.....	146
<b>Глава 5. Основы поиска по нескольким полям .....</b>	<b>147</b>
5.1. Сигналы и моделирование сигналов.....	149
5.1.1. Что такое сигнал? .....	149
5.1.2. Модель исходных данных.....	150
5.1.3. Реализация сигнала.....	153
5.1.4. Моделирование сигнала: моделирование данных для нужд релевантности.....	154
5.2. TMDb – поиск, последний рубеж!.....	155
5.2.1. Нарушение главной заповеди.....	157
5.2.2. Упрощение вложенных документов .....	157
5.3. Моделирование сигналов при поиске по полям.....	160
5.3.1. Первая попытка с best_fields.....	164
5.3.2. Управление выбором полей в результатах поиска.....	167
5.3.3. Улучшение стратегии best_fields более точными сигналами .....	169
5.3.4. Поделимся триумфом с проигравшими: калибровка best_fields .....	172
5.3.5. Учет нескольких сигналов в стратегии most_fields.....	175
5.3.6. Форсирование оценок в стратегии most_fields.....	177
5.3.7. Когда дополнительные совпадения не имеют значения.....	178
5.3.8. Вердикт стратегии most_fields .....	180
5.4. В заключение.....	180
<b>Глава 6. Поиск по терминам .....</b>	<b>182</b>
6.1. Что такое поиск по терминам?.....	183
6.2. Что дает поиск по терминам? .....	185
6.2.1. Охота на белых слонов .....	185
6.2.2. Поиск белого слона в примере Star Trek.....	188
6.2.3. Несоответствие сигналов.....	190
6.2.4. Понимание механики несоответствия сигналов .....	191
6.3. Наш первый поиск по терминам.....	193
6.3.1. Функция ранжирования в поиске по терминам.....	194
6.3.2. Поиск по терминам с использованием парсера запросов (неудачная) .....	197
6.3.3. Синхронность полей.....	198
6.3.4. Синхронность полей и моделирование сигналов.....	199
6.3.5. Парсеры запросов и несоответствие сигналов.....	200
6.3.6. Настройка поиска по терминам .....	202
6.4. Решение проблемы несоответствия сигналов в поиске по терминам.....	204
6.4.1. Объединение полей.....	205
6.4.2. Решение проблемы несоответствия сигналов с cross_fields.....	209
6.5. Объединение стратегий поиска по полям и терминам: как рыбку съесть, и косточкой не подавиться.....	211

6.5.1. Группировка «подобных полей» .....	212
6.5.2. Ограничения группировки полей .....	213
6.5.3. Объединение жадного поиска с консервативными усилителями .....	215
6.5.4. Поиск по терминам против поиска по полям и точность против полноты .....	218
6.5.5. Фильтрация, форсирование и переупорядочение .....	218
6.6. В заключение .....	219
<b>Глава 7. Перегрузка операторов и другие соглашения .....</b>	<b>220</b>
7.1. Что означает «формирование оценки»? .....	221
7.2. Форсирование: продвижение результатов .....	223
7.2.1. Форсирование: последний рубеж .....	223
7.2.2. Форсирование – прибавлять или умножать? Логический или функциональный запрос? .....	224
7.2.3. Решение первое: аддитивное форсирование с логическими запросами .....	226
7.2.4. Решение второе: применение функциональных запросов для ранжирования .....	230
7.2.5. Практика применения функциональных запросов: простое мультипликативное форсирование .....	232
7.2.6. Основы форсирования: сигналы, сигналы повсюду .....	234
7.3. Фильтрация: исключение результатов .....	234
7.4. Стратегии формирования оценок для удовлетворения потребностей бизнеса .....	236
7.4.1. Поиск всех фильмов! .....	237
7.4.2. Моделирование форсирующих сигналов .....	239
7.4.3. Функция ранжирования: добавление уровней с высокой оценкой .....	243
7.4.4. Уровень с высокой оценкой на основе функционального запроса .....	247
7.4.5. Игнорирование метрики $TF \times IDF$ .....	249
7.4.6. Определение качественных метрик .....	250
7.4.7. Оценка свежести .....	252
7.4.8. Объединение функциональных запросов .....	255
7.4.9. Объединяем все вместе! .....	258
7.5. В заключение .....	258
<b>Глава 8. Релевантная обратная связь .....</b>	<b>260</b>
8.1. Релевантная обратная связь в строке ввода запроса .....	262
8.1.1. Синхронный поиск в процессе ввода .....	262
8.1.2. Помощь в составлении более конкретных запросов с функцией подсказки .....	264
8.1.3. Исправление опечаток и орфографических ошибок с подсказками .....	273
8.2. Релевантная обратная связь в процессе просмотра .....	276
8.2.1. Реализация возможности обзора по категориям .....	278
8.2.2. Навигационные цепочки .....	280
8.2.3. Альтернативное упорядочение результатов .....	281
8.3. Релевантная обратная связь в результатах поиска .....	282
8.3.1. Какая информация должна выводиться в списке с результатами? .....	283
8.3.2. Релевантная обратная связь через подсветку фрагментов .....	284
8.3.3. Группировка схожих документов .....	288
8.3.4. Помощь пользователю в отсутствие результатов .....	291
8.4. В заключение .....	291

<b>Глава 9. Проектирование приложений релевантного поиска .....</b>	<b>293</b>
9.1. Yow! Новый проект! .....	294
9.2. Сбор информации и требований .....	295
9.2.1. Пользователи и их информационные потребности .....	296
9.2.2. Бизнес и его потребности .....	298
9.2.3. Определение требуемой и доступной информации .....	298
9.3. Проектирование поискового приложения .....	300
9.3.1. Пользовательский интерфейс .....	301
9.3.2. Определение полей и моделирование сигналов .....	304
9.3.3. Комбинирование и балансирование сигналов .....	305
9.4. Развертывание, мониторинг и совершенствование .....	318
9.4.1. Мониторинг .....	318
9.4.2. Выявление и исправление проблем! .....	320
9.5. Важно вовремя остановиться .....	322
9.6. В заключение .....	322
<b>Глава 10. Предприятие, опирающееся на релевантность .....</b>	<b>324</b>
10.1. Обратная связь: фундамент предприятия, зависящего от релевантности .....	326
10.2. Почему пользователь важнее данных? .....	328
10.3. Полет вслепую .....	331
10.4. Создание начальной обратной связи: эксперты в предметной области и опытные пользователи .....	334
10.5. Зрелость обратной связи: курирование контента .....	336
10.5.1. Роль куратора контента .....	337
10.5.2. Риск недопонимания в отношениях с куратором контента .....	339
10.6. Рационализация релевантности: парная работа с куратором .....	340
10.7. Ускорение релевантности: настройка релевантности через тестирование .....	342
10.7.1. Понимание настройки релевантности через тестирование .....	342
10.7.2. Использование данных о поведении пользователей в тестировании релевантности .....	345
10.8. Другая сторона настройки релевантности через тестирование: обучение ранжированию .....	346
10.9. В заключение .....	348
<b>Глава 11 Семантический поиск и персонализация .....</b>	<b>350</b>
11.1. Персонализация поиска на основе профилей пользователей .....	352
11.1.1. Извлечение информации из профиля пользователя .....	353
11.1.2. Связывание информации из профиля с поисковым индексом .....	353
11.2. Персонализация поиска на основе поведения пользователя .....	355
11.2.1. Введение в совместную фильтрацию .....	355
11.2.2. Простая совместная фильтрация с использованием подсчета совместного появления .....	356
11.2.3. Связывание информации о поведении пользователя с поисковым индексом .....	362

11.3. Базовые методы концептуального поиска.....	366
11.3.1. Конструирование концептуальных сигналов.....	367
11.3.2. Дополнение содержимого синонимами.....	368
11.4. Концептуальный поиск с применением методов машинного обучения.....	369
11.4.1. Важность фраз в концептуальном поиске.....	371
11.5. Связь персонализированного и концептуального поиска .....	372
11.6. Рекомендации как обобщение поиска.....	373
11.6.1. Замена поиска рекомендациями .....	375
11.7. Пожелания успехов на стезе релевантного поиска.....	376
11.8. В заключение .....	376
<b>Приложение А. Индексирование непосредственно из TMDB .....</b>	<b>378</b>
А.1. Получение ключа TMDB API и настройка окружения.....	378
А.2. Подготовка к взаимодействиям с TMDB API .....	379
А.3. Обход TMDB API .....	380
А.4. Индексирование фильмов в Elasticsearch .....	382
<b>Приложение В. Справочник для пользователей Solr .....</b>	<b>384</b>
В.1. Глава 4: укрощение лексем в Solr .....	385
В.1.1. Краткая сводка функций анализа и отображения в Solr.....	385
В.1.2. Создание собственного анализатора в Solr .....	385
В.1.3. Отображение полей в Solr.....	387
В.2. Главы 5 и 6: поиск по нескольким полям в Solr .....	388
В.2.1. Краткая сводка возможностей управления запросами .....	388
В.2.2. Различия между запросами в Solr и Elasticsearch .....	388
В.2.3. Эргономика запросов в Solr .....	390
В.2.4. Поиск по терминам и полям с применением парсера запросов edismax .....	391
В.2.5. Методы поиска с объединением полей и cross_fields .....	392
В.3. Глава 7: формирование функции ранжирования в Solr .....	393
В.3.1. Краткая сводка средств форсирования.....	393
В.3.2. Форсирование в логических запросах Solr .....	393
В.3.3. Функциональные запросы в Solr .....	394
В.3.4. Мультипликативное форсирование в Solr .....	396
В.4. Глава 8: релевантная обратная связь.....	396
В.4.1. Краткая сводка средств поддержки релевантной обратной связи .....	396
В.4.2. Автодополнение в Solr: поиск по началу фразы.....	397
В.4.3. Обзор по категориям в Solr .....	397
В.4.4. Свертка полей.....	398
В.4.5. Подсказки и подсветка .....	398
<b>Предметный указатель .....</b>	<b>400</b>

# Предисловие

За последнее десятилетие функция поиска проникла во все сферы – поле поиска по ключевой фразе фактически стало неотъемлемой частью пользовательского интерфейса большинства веб-сайтов и приложений. В то же время поддержка по-настоящему релевантного поиска для большинства организаций оставалась слабым местом, порой не выдерживающим никакой критики.

За это время появились мощные технологии с открытым исходным кодом, обеспечивающие быстрый полнотекстовый поиск (Apache Lucene) в распределенном и легко масштабируемом окружении, практически не требующем писать дополнительный код (Apache Solr и позднее Elasticsearch). Они обеспечили необходимую инфраструктуру для создания «в целом релевантных» механизмов поиска в масштабе реального времени, соответствующих требованиям эпохи больших данных. По мере устранения проблем, характерных для инфраструктур жесткого поиска, и стандартизации решений, многие организации стали отказываться от масштабируемых механизмов быстрого поиска в пользу инструментов, гарантирующих более релевантное соответствие информационным потребностям пользователя. Иными словами, предоставление «в целом релевантных» результатов уже не является достаточным – Google и другие крупные поисковые системы уже приучили пользователей, что поисковые приложения почти читают их мысли. В этой книге рассказывается, как еще активнее двигаться в направлении понимания намерений пользователей.

Даг Тарнбулл (Doug Turnbull) и Джон Берримен (John Berryman) – два опытных эксперта в области технологий релевантного поиска. Мы знакомы уже много лет и часто встречаемся на конференциях, посвященных проблемам поиска. Я с удовольствием вспоминаю время, проведенное с ними за обсуждением идей решения самых сложных проблем, связанных с релевантностью поиска, персонализацией и автоматическим подбором рекомендаций. Никто так не волновался, ожидая знакомства с их опытом, облеченным в форму этой книги – одной из лучших технических книг, которые я когда-либо читал.

Настройка релевантности – сложная проблема. Многие неправильно ее понимают, и часто это остается неочевидным, пока не обнаружится ошибка. Обычно, чтобы заметить несоответствие требуется увидеть много неудачных примеров, и часто бывает сложно понять, как должны выглядеть удачные результаты, не имея их перед глазами. К сожалению, это редко происходит до внедрения системы в эксплуатацию, когда организации начинают замечать разрыв между желаемыми и фактическими результатами поиска.

Кроме того, навыки (знание предметной области и характерных признаков, знакомство с приемами машинного обучения, онтологии и обработки

естественного языка), необходимые для реализации релевантного поиска, сильно отличаются от тех, что требуются для создания и сопровождения масштабируемых инфраструктур (распределенные системы, структуры данных, производительность и конкурентное выполнение, знание аппаратного обеспечения, приемы организации сетевых взаимодействий). Должность специалиста по релевантности вообще отсутствует во многих организациях, из-за чего остается нереализованным потенциал релевантного поиска, который по-настоящему радует пользователей и помогает двигать компанию вперед.

Спектр персонализации между поиском по ключевым фразам, вводимым вручную, и автоматическим подбором рекомендаций также открывает широкое поле деятельности для внедрения релевантности и обеспечения более полного соответствия потребностям пользователей. Авторы проделали большую работу, объясняя некоторые из тонкостей, связанных с поиском, позволяющих в полной мере использовать этот спектр. Вооруженные приемами, описываемыми в этой книге, вы сможете взять на себя роль специалиста по релевантности и найти решение большинства проблем, присущих системам настоящего релевантного и персонализированного поиска.

*Трей Грейнджер (Trey Grainger)*

Автор книги «Solr in Action»

Первый вице-президент компании  
«Engineering at Lucidworks»



# Вступление

Джон и я познакомились, когда работали вместе как консультанты в Open-Source Connections (OSC) и решали проблемы поиска для клиентов. Для одних мы искали более производительные решения (увеличивали скорость работы!). Другим помогали в создании поисковых приложений. Все наши проекты имели простую меру успеха. Удалось увеличить производительность? Приложение было закончено?

Однако релевантность поиска – совсем другое дело. Пользователи, взращенные в эпоху Google, не хотят иметь «просто хороший» поиск. Им нужен «чертовски умный» поиск. Они желают иметь возможность определять критерии поиска, и их не устраивает, когда поисковая система подбирает результаты «наобум».

Как мотыльки, привлекаемые пламенем, мы оба чувствовали тягу к этой трудной проблеме. И так же как мотыльки, мы часто обжигали свои крылья. Усваивая эти болезненные уроки, мы неуклонно росли и набирались опыта, и теперь с успехом решаем задачи, которые раньше казались нам слишком сложными.

За это время окрепли наши голоса в блоге OSC. Мы обнаружили, что о проблемах релевантного поиска пишется непозволительно мало. Мы выдвинули и реализовали идею релевантности, управляемой тестированием. Мы подробно описали наши трудности, проблемы и победы. Вместе мы экспериментировали с приемами машинного обучения, такими как латентно-семантический анализ. Мы углубились в изучение внутренних механизмов Lucene и исследовали приемы создания компонентов поиска для решения проблем. Мы начали изучение темы поиска информации. Узнавая больше приемов решения сложных проблем, мы тут же писали о них.

Но блоги имеют свои ограничения. Мы с Джоном всегда хотели выразить наши идеи более системно, в форме книги. К счастью, произошла целая цепочка благоприятных событий, благодаря которым у нас появилась такая возможность. Я, вместе с Эндрю Монталенти (Andrew Montalenti), выступил с докладом о средствах параллельного выполнения в Python на местной встрече. Эндрю упомянул этот доклад на конференции PyCon и в результате ему позвонили из издательства Manning, чтобы обсудить возможность создания книги о параллельном программировании на Python. Эндрю ответил, что не собирается писать книгу, но его содокладчик Даг, возможно, согласится.

Так получилось, что я тоже не особенно хотел писать книгу о параллельном программировании на Python, но подкинул идею написать другую книгу. Я подошел с этой идеей к Джону и, побеседовав с ним на эту тему два-три раза, мы подготовили для издательства предложение, от которого они не смогли отказаться!

Этот важный звонок из Manning случился почти два года тому назад. С тех пор много воды утекло. За время, пока мы писали книгу, у нас обоих в семьях произошло прибавление. Я начал практиковать оказание консультационных услуг о релевантности. Джон поменял работу, став экспертом по поиску в Eventbrite. И все же мы не могли сопротивляться желанию написать об этой потрясающей теме.

Эта книга не похожа на другие технические книги. Она не является перечислением особенностей одной технологии. Скорее это карта с нашим маршрутом через многие годы боли, и решений проблем, на которые не было готовых ответов. Иными словами, мы прошли насквозь пустыню релевантного поиска, встретили множество оазисов и научились уклоняться от встреч с песчаными людьми и штурмовиками.

Мы передаем вам эту карту, чтобы вы не блуждали в пустыне, как это случилось с нами. А теперь извините, но мы собираемся отправиться на ближайший пляж и отдохнуть...

*Даг Тарнбулл*

# Благодарности

За несколько недель до начала работы над этой книгой, в наших семьях произошло прибавление. Мы выражаем самую глубокую благодарность нашим супругам, Харе Тарнбулл (Khara Turnbull) и Кумико Берримен (Kumiko Berryman). Они много недель не имели выходных, пока мы работали над книгой – за это время Хара закончила свою книгу, а Кумико занималась продажей дома и организацией переезда в другую страну. Настало время для долгого отпуска!

Эта книга не состоялась бы без основателя OpenSource Connections – Эрика Пью (Eric Pugh). Как наш «босс», он подталкивал нас писать, говорить и решать большие проблемы. Как настоящий лидер, Эрик делает свою страсть вашей страстью. Если бы Эрик не отпускал нас в свободное плавание (а иногда даже не подталкивал), мы не смогли бы поверить в свои силы и возможности, как писатели и специалисты, способные решать сложные проблемы. Эрик учил нас, что всякий, в том числе и мы, может стать лидером.

Спасибо проекту TMDb за предоставленные данные и поддержку. Мы потратили массу времени, пытаясь подобрать хорошие наборы данных. Проект TMDb (<http://themoviedb.org>) не только предоставил богатый набор поисковых данных, но также поддерживал нас и наших первых читателей, пока мы выявляли и исправляли ошибки и неточности, в основном в нашем собственном коде. Тревис Белл (Travis Bell) заслуживает особой нашей благодарности за оперативную реакцию на наши проблемы и электронные письма.

Создание книг – это командный вид спорта и мы хотели бы поблагодарить всех сотрудников издательства Manning, участвовавших в работе над этой книгой: Марину Майклз (Marina Michaels), нашего редактора-консультанта по аудитории; Аарона Колкорда (Aaron Colcord), технического редактора; Валентина Креттаза (Valentin Crettaz), технического корректора; Френка Полманна (Frank Pohlmann) и Майка Стивенса (Mike Stephens), рецензентов издательства; и Кэндис Джиллхули (Candace Gillhoolley) за маркетинг.

Мы также хотим сказать спасибо многим рецензентам, читавшим первые рукописи книги и давшим ценные советы, включая Джона Гатри (John Guthrie), Мартина Бира (Martin Beer), Артура Зубарева (Arthur Zubarev), Эльмана Кринкера (Elman Krinker), Амита Ламба (Amit Lamba), Марка-Оливера Шиила (Marc-Oliver Scheele), Яна Стерка (Ian Stirk), Джозефа Вана (Joseph Wang), Стюарта Вудворда (Stuart Woodward), Урсина Стаусса (Ursin Stauss), Расса Кама (Russ Cam), Майкла Финка (Michael Fink), Грегора Зуровски (Gregor Zurowski), Димитриоса Кузиса-Лукаса (Dimitrios Kouzis-Loukas), Джереми Гейлора (Jeremy Gailor) и Кейт Вебстера (Keith Webster).

Особое спасибо Эндрю Монталенти (Andrew Montalenti), связавшему нас с издательством Manning. Спасибо Шаю Банону (Shay Banon), создателю Elasticsearch, за его поддержку, а еще просто за то, что он отличный парень. Спасибо нашим коллегам: Трею Грейнджеру (Trey Grainger), Мэту Оверстриту (Matt Overstreet), Рене Морзе (Rena Morse), Дэвиду Смайли (David Smiley), Гранту Ингерсоллу (Grant Ingersoll), Йонику Сили (Yonik Seeley), Рене Криглеру (Rene Kriegler), Питеру Диксону-Моисею (Peter Dixon-Moses), Чарли Халлу (Charlie Hull) и Дрю Фаррису (Drew Farris) за постоянный обмен мнениями на тему поиска и релевантности на протяжении многих лет. И особое спасибо Трею за предисловие к этой книге.

Спасибо всем членам наших семей за их поддержку. Особенно нашим детям: Мегуме Берримен (Megume Berryman), Яну Тарнбуллу (Ian Turnbull) и Мюррею Тарнбуллу (Murray Turnbull). Спасибо нашим «семьям на работе» в OpenSource Connections и Eventbrite, что позволили нам вложить всю нашу энергию в эту книгу.

# Об этой книге

Книга «*Релевантный поиск*» рассказывает, как отвечать на запросы пользователей содержимым, удовлетворяющим их потребности. Она научит вас строго контролировать ранжирование результатов поиска на основе четких критериев, а не мистических прихотей поискового движка. Мы изложим наш подход к глубокой настройке релевантности в Solr или Elasticsearch, а также познакомим с методами, которые помогут вам понять, что является релевантным для вашего приложения.

## Кому адресована книга

Книга «*Релевантный поиск*» предназначена для разработчиков, использующих Solr или Elasticsearch, столкнувшихся с проблемой, когда поисковый механизм возвращает пользователям не всегда релевантные результаты поиска. Читатели, имеющие хотя бы поверхностное знакомство с их поисковым механизмом, смогут с помощью этой книги подняться на следующий уровень. Несмотря на то, что эта книга относится к разряду технических, значительная ее часть посвящена организационным стратегиям – для руководителей проектов, контент-стратегам, специалистам по маркетингу или предметной области, занимающимся проблемами поиска.

## Краткое содержание

Книга «*Релевантный поиск*» постепенно раскрывает технические детали стратегии преодоления проблем, с которыми вам придется столкнуться при определении и решении задачи релевантного поиска. Книга завершается следующими шагами: как организовать персонализированный поиск, семантический поиск и автоматический подбор рекомендаций.

*Глава 1* начинается с обсуждения проблемы релевантности. Она охватывает такие области, как веб-поиск, электронная коммерция и экспертный поиск. Глава обсуждает степень поддержки академическими кругами наших стремлений к релевантному поиску. В конце мы изложим техническую стратегию поддержки релевантности.

*Глава 2* представляет краткий обзор основных структур данных и алгоритмов в Lucene, имеющих отношение к релевантности. Вы увидите, как Lucene обеспечивает невероятную основу для организации релевантного поиска.

*Глава 3* научит приемам отладки релевантного поиска. Когда структуры данных и алгоритмы, обсуждаемые в главе 2, не дают желаемых результатов, вам придется копнуть глубже, чтобы понять, что мешает поиску.

*Глава 4* покажет, как разложить содержимое и поиск на описательные признаки, используя процесс анализа в поисковом механизме. Это фун-

даментальное умение поможет вам использовать анализ для поиска чего угодно.

*Глава 5* начинает обсуждение стратегий запросов по нескольким полям. В этой главе мы расскажем, как конструировать запросы, оценивающие конкретные факторы ранжирования, важные для ваших пользователей.

*Глава 6* продолжает обсуждение стратегий запросов. Здесь мы сосредоточимся на стратегиях поиска по ключевым терминам, которые поддерживают упрощенное понимание релевантности пользователями.

*Глава 7* демонстрирует такие приемы регулирования, как форсирование и фильтрация. Вам часто придется управлять поиском, отдавая приоритет более свежей информации, прибыльным продуктам или близлежащим местоположениям.

*Глава 8* показывает альтернативные пути, ведущие пользователей к релевантной информации. Иногда, когда ранжирование по релевантности терпит неудачу, направить пользователя в правильном направлении могут помочь некоторые особенности пользовательского интерфейса, такие как категории, автодополнение и выделение.

*Глава 9* описывает создание приложения с полноценной поддержкой релевантного поиска, что поможет вам глубже вникнуть в суть. Получив достаточно полное представление о релевантном поиске в предыдущих главах, в этой главе вы увидите весь процесс создания продукта, от начала до конца.

*Глава 10* поднимается на ступень выше в исследовании стратегий создания продукта и рассматривает культурные и организационные факторы. Какая информация считается релевантной для организации? Здесь вы увидите, как организация может реализовать быструю и точную обратную связь для направления усилий специалиста по релевантности.

*Глава 11* выведет вас за рамки поискового механизма. Здесь вы узнаете, как можно улучшить релевантность поиска за счет применения приемов машинного обучения, персонализации и семантического поиска.

*Приложение А* подробно, шаг за шагом, описывает процесс загрузки данных для примеров из книги в Elasticsearch посредством The Movie Database (TMDB) API.

*Приложение В* предназначено для знатоков Solr и описывает соответствие функций поддержки релевантности в Elasticsearch и Solr.

## **Об исходном коде**

Эта книга содержит много примеров исходного кода, и в виде листингов, и в виде фрагментов в обычном тексте. В обоих случаях исходный код оформляется моноширинным шрифтом, чтобы его можно было отличить от обычного текста. Иногда, чтобы подчеркнуть отличия от предыдущего шага и

выделить вновь добавленные особенности, код будет оформляться **жирным моноширинным шрифтом**.

Во многих случаях оригинальный исходный код был переформатирован; мы добавили переносы строк и изменили ширину отступов, чтобы уместить строки кода по ширине книжной страницы. Кроме того, мы убрали комментарии из кода, если он описывается в тексте книги. Многие листинги сопровождаются дополнительными аннотациями, подчеркивающими наиболее важные идеи.

Примеры были проверены с Elasticsearch 2.0 и Python 2.7.

Исходный код примеров для глав 3–9 можно найти на веб-сайте издательства Manning ([www.manning.com/books/relevant-search](http://www.manning.com/books/relevant-search)) и репозитории книги на сайте GitHub (<http://github.com/o19s/relevant-search-book>). Примеры создавались в среде iPython Notebook/Jupyter, чтобы упростить эксперименты с ними. В файле README вы найдете подробный перечень дополнительного программного обеспечения, которое требуется установить для опробования примеров.

## Автор в сети

Одновременно с покупкой книги «Релевантный поиск» вы получаете бесплатный доступ к частному веб-форуму, организованному издательством Manning Publications, где можно оставлять комментарии о книге, задавать технические вопросы, а также получать помощь от автора и других пользователей. Чтобы получить доступ к форуму и зарегистрироваться на нем, откройте в веб-браузере страницу [www.manning.com/relevant-search](http://www.manning.com/relevant-search). Здесь описывается, как попасть на форум после регистрации, какие виды помощи доступны и правила поведения на форуме.

Издательство Manning обязуется предоставить своим читателям место встречи, где может состояться содержательный диалог между отдельными читателями и между читателями и автором. Но со стороны автора отсутствуют какие-либо обязательства уделять форуму какое-то определенное внимание – его присутствие на форуме остается добровольным (и неоплачиваемым). Мы предлагаем задавать автору стимулирующие вопросы, чтобы его интерес не угасал!

Форум и архив с предыдущими обсуждениями остается доступным на сайте издательства, пока книга продолжает издаваться.

## Другие онлайн-ресурсы

Желающим узнать больше мы рекомендуем обращаться к следующим замечательным ресурсам:

- блог OpenSource Connection (<http://opensourceconnections.com/blog>);
- личный блог Джона Берримана (John Berryman, <http://thoughtbox.solutions>);

- блог проекта Elastic ([www.elastic.co/blog](http://www.elastic.co/blog));
- блог проекта Lucidwork (<https://lucidworks.com/blog>);
- блог «Salmon Run» Суджиты Пала, посвященный Solr (<http://sujitpal.blogspot.com/>);
- новостная рассылка Solr Start ([www.solr-start.com](http://www.solr-start.com)).

Желающим изучить тему поиска и извлечения информации в целом, мы рекомендуем следующий канонический труд:

- «Introduction to Information Retrieval» Кристофера Маннинга (Christopher Manning) с соавторами. (Cambridge University Press, 2008), <http://nlp.stanford.edu/IR-book/>.<sup>1</sup>

С вопросами, касающимися конкретно технологий Solr/Elasticsearch, мы рекомендуем обращаться на специализированные дискуссионные форумы:

- по теме Elasticsearch: <http://discuss.elastic.co>;
- по теме Solr: <http://lucene.apache.org/solr/resources.html>.

---

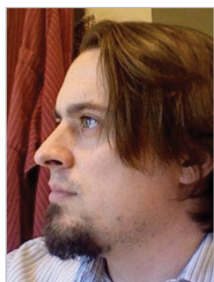
<sup>1</sup> Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск. М.: Вильямс. 2014. ISBN 978-5-8459-1623-5, 978-0-5218-6571-5. – Прим. перев.



# Об авторах



**Даг Тарнбулл (Doug Turnbull)** ведет консультационную практику в OpenSource Connections, где часто выступает и ведет свой блог. Даг конструирует системы релевантного и семантического поиска для своих клиентов в самых разных областях, используя разнообразные технологии поиска и обработки естественного языка.



**Джон Берримен (John Berryman)** начинал карьеру как инженер в аэрокосмической промышленности, но спустя несколько лет обнаружил, что ему больше нравится программировать или решать сложные математические задачи. В конце концов Джон забросил самолеты и спутники, и занялся разработкой программного обеспечения, программных архитектур и поисковых технологий. В настоящее время Джон работает в Eventbrite, помогая создавать системы обнаружения событий, поиска и автоматического подбора рекомендаций с применением Elasticsearch.

# Об иллюстрации на обложке

На обложке книги изображен рисунок, подписанный как «Homme de l'Isle de Pathmos», или человек с острова Патмос (Греция). Иллюстрация взята из каталога костюмов разных стран, составленного Жаком Грассетом де Сен-Совер (Jacques Grasset de Saint-Sauveur, 1757–1810), озаглавленного как «Costumes de Différents Pays» и опубликованного во Франции в 1797. Каждая иллюстрация в этом каталоге нарисована и раскрашена вручную. Многообразие рисунков в каталоге Грассета де Сен-Совер отчетливо демонстрирует уникальные и индивидуальные особенности городов и районов мира, существовавших 200 лет назад. Изолированные друг от друга, люди говорили на разных языках и диалектах. Встретив человека на улице, по его одежде легко можно было определить, где он живет, кем работает и какое положение в обществе занимает.

С тех пор мода изменилась и региональные различия, такие существенные в те времена, исчезли. Сейчас зачастую сложно отличить жителей разных континентов, не говоря уже о жителях из разных городов, областей или стран. Но, если посмотреть на это с оптимистической точки зрения, мы обменяли культурное и визуальное разнообразие на более разнообразную личную жизнь. Или более разнообразную и интересную интеллектуальную жизнь и техническую вооруженность.

Мы в издательстве Manning славим изобретательность, предприимчивость и радость компьютерного бизнеса обложками книг, изображающими богатство региональных различий двухвековой давности, возвращая к жизни иллюстрации из каталога Грассета де Сен-Совер.

# Глава 1

## Задача релевантного поиска

Эта глава охватывает следующие темы:

- вездесущий поиск (поиск вокруг нас!);
- сложности реализации релевантного поиска;
- примеры сложностей поиска в разных предметных областях;
- неспособность готовых решений справиться с проблемой;
- подход к организации релевантного поиска, освещаемый в этой книге.

Создание механизма поиска, обладающего желаемым поведением, может свести с ума. И только начинающие осваивать Solr или Elasticsearch, и имеющие многолетний опыт использования этих инструментов, рано или поздно сталкиваются с низким качеством результатов поиска. Стандартные настройки часто не соответствуют нашим потребностям, и нам приходится бороться даже за мало-мальски релевантные результаты поиска.

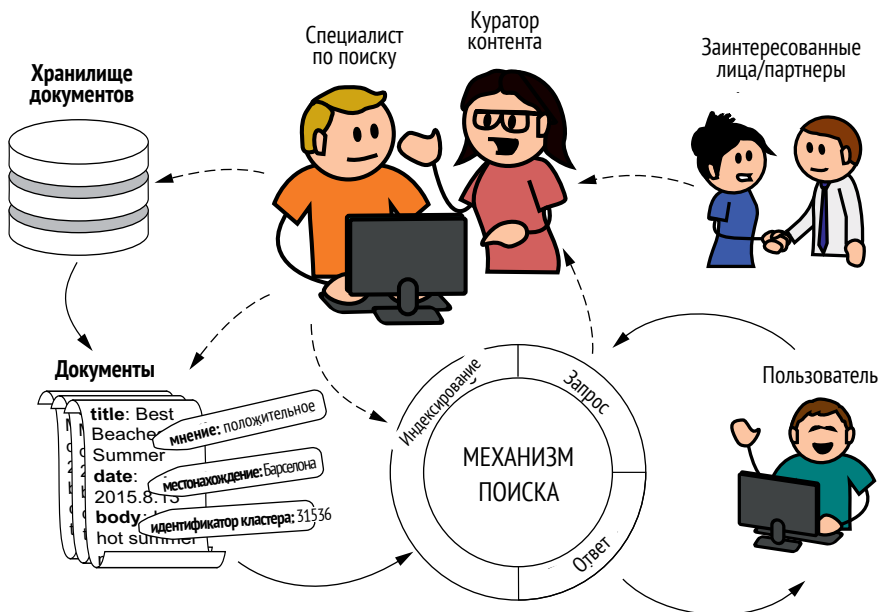
Когда дело доходит до оценки релевантности результатов, механизм поиска может показаться мистическим черным ящиком. Часто возникает желание игнорировать проблемы релевантности и обратить свое внимание на другие, менее мистические, аспекты приложения, такие как производительность или оформление пользовательского интерфейса. К сожалению, избежать оценки релевантности не удастся. Современным пользователям все чаще приходится работать с большими объемами информации, будь то каталоги продуктов, книги, сообщения в файлах журналов, электронные письма, каталоги гостиниц или медицинские статьи. Поле поиска – первое место, куда обращаются пользователи, чтобы приступить к изучению и найти ответы. Не имея возможности получать интуитивные ответы на вопросы, пользователи чувствуют себя потерявшими. Поэтому, несмотря на сумасшедший и почти мистический характер работы механизмов поиска, вы должны найти решение.

Книга «Релевантный поиск» срывает покров тайны с релевантности. Что такое релевантность на самом деле? Это основная ценность, которую может предложить механизм поиска. *Релевантность* – это искусство определения в процессе поиска ценности информации для пользователя и бизнеса. Дьявол, как известно, кроется в деталях. Какая информация оценивается? (Твиты? Продукты? Детские игрушки?) К какой категории относятся пользователи? (Врачи? Технически подкованные покупатели?) Что является предметом поиска? (Название на японском? Товары известных производителей? Юридический термин?) Что ожидают получить пользователи в ответ? (Список товаров в магазине? Каталог книг в библиотеке?) И что надеется получить от этого взаимодействия ваш работодатель? (Прибыль? Увеличение количества просмотров? Восхищенные отзывы?) Поиск все чаще становится неотъемлемой частью наших приложений, уверенно отвоевывая себе место под солнцем. Ответы на эти вопросы (в отношении релевантности) обуславливают разницу между благоприятными впечатлениями пользователей и их разочарованием.

## 1.1. Ваша цель: стать специалистом по релевантности

Как ее достичь? Книга «Релевантный поиск» обучит вас всему, что должен знать специалист по релевантности. Обладая этими знаниями, вы сможете трансформировать механизм поиска в интеллектуальную систему, которая понимает потребности пользователей и бизнеса. Для этого вы подскажете механизму поиска, какие характеристики в вашей информации считаются наиболее важными: местоположение ресторана, слова в тексте книги или цвет рубашки. Определив перечень важных признаков, вы сможете оценивать важность информации для пользователей при поиске: как далеко находится ресторан? Эта книга описывает нужную тему? Будет ли эта рубашка соответствовать только что купленным брюкам? Все эти факторы, оцениваемые в процессе поиска и сообщающие о важности или неважности информации для пользователя, называются *сигналами*. Вы увидите, что основной проблемой является выбор признаков и реализация сигналов, отражающих потребности ваших пользователей и бизнеса.

Но техническое мастерство – лишь часть работы (как показано на рис. 1.1). Понимание, что именно требуется реализовать, порой важнее, чем знание, как это сделать. Как ни странно, специалист по релевантности редко знает, что означает «релевантность» для данного приложения. Но этим знанием обладают другие – обычно нетехнические специалисты – понимающие назначение информации и цели пользователей и бизнеса. В этой книге вы научитесь добиваться релевантности, используя знания и опыт других, а также данные о поведении пользователей в процессе поиска.



**Рис. 1.1.** Специалист по релевантности использует механизм поиска и вспомогательные технологии, чтобы выразить бизнес-логику оценки результатов. Он тесно сотрудничает с другими специалистами, способными оценить релевантность, и широко использует информацию о пользователях.

Мы еще будем уточнять упомянутые идеи далее в этой главе (и книге). Но, чтобы заложить надежный фундамент, в оставшейся части этой главы мы определим основные наши задачи и проблемы. Почему релевантный поиск сопряжен с большими сложностями? Что можно предпринять, чтобы справиться с этой сложностью? Затем мы переключимся на обсуждение подхода к решению проблемы релевантности, являющегося основной темой этой книги.

## 1.2. Сложности релевантного поиска

Сложность реализации релевантного поиска отчасти объясняется тем, что мы принимаем непосредственное участие в акте поиска. Приложения, осуществляющие поиск, получают от пользователей запросы (текст, который они вводят в поле поиска) и пытаются оценить степень соответствия найденной информации.

Этот акт происходит так часто, что остается почти незамеченным. Возьмите в качестве примера самого себя. Возможно вы, проснувшись сегодня утром, включили кофеварку и начали возиться со своим смартфоном. Вы просмотрели новости, заглянули в Facebook и проверили электронную почту. Еще до того, как кофе был готов, вы наверняка успели поработать с десятком поисковых приложений, даже не задумываясь об этом. Вы от-

Конец ознакомительного фрагмента.

Приобрести книгу можно  
в интернет-магазине «Электронный универс»  
([e-Univers.ru](http://e-Univers.ru))