

УДК 004.738.52
ББК 32.971.353
T21

T21 Даг Тарибулл, Джон Берримен

Релевантный поиск с использованием Elasticsearch и Solr. / пер. с англ. Киселев А. Н. – М.: ДМК Пресс, 2018. – 408 с.: ил.

ISBN 978-5-97060-592-9

Данная книга поможет вам раскрыть суть и механику релевантного поиска на базе библиотеки Apache Lucene. На примере поисковых систем Elasticsearch и Solr вы научитесь строго контролировать ранжирование результатов поиска на основе четких критериев. Вы поймете, как программировать релевантность, как подключить вторичные источники данных, классификаторы, организовать анализ текста. Наконец вы узнаете, как можно улучшить релевантность поиска за счет применения приемов машинного обучения, персонализации и семантического поиска.

Издание предназначено разработчикам, стремящихся создавать интеллектуальные поисковые приложения на основе Elasticsearch или Solr.

Original English language edition published by Manning Publications USA, USA. Copyright © 2017 by Manning Publications. Russian language edition copyright © 2018 by DMK Press. All rights reserved.

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

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.

ISBN 978-1-61729-277-4 (англ.)
ISBN 978-5-97060-592-9 (рус.)

© 2017 by Manning Publications Co.
© Оформление, перевод на русский язык, издание,
ДМК Пресс, 2018

Оглавление

Предисловие	11
Вступление	13
Благодарности	15
Об этой книге	17
Кому адресована книга.....	17
Краткое содержание.....	17
Об исходном коде.....	18
Автор в сети.....	19
Другие онлайн-ресурсы	19
Об авторах	21
Об иллюстрации на обложке	21
Глава 1. Задача релевантного поиска	22
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
Глава 2. Поиск – взгляд изнутри	40
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
Глава 3. Отладка первой проблемы релевантного поиска	68
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
Глава 4. Угрошение лексем	109
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
Глава 5. Основы поиска по нескольким полям	147
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
Глава 6. Поиск по терминам	182
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
Глава 7. Перегрузка операторов и другие соглашения	220
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
Глава 8. Релевантная обратная связь	260
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

Глава 9. Проектирование приложений релевантного поиска	293
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
Глава 10. Предприятие, опирающееся на релевантность	324
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
Глава 11 Семантический поиск и персонализация	350
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
Приложение А. Индексирование непосредственно из TMDB	378
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
Приложение В. Справочник для пользователей Solr	384
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
Предметный указатель	400