

Поиск на основе искусственного интеллекта

Трей Грейнджер
Дуг Тернбулл
Макс Ирвин



MANNING



Краткое оглавление

ЧАСТЬ 1. Современная релевантность поиска	29
1  Знакомство с поиском на основе ИИ	31
2  Работа с естественным языком	58
3  Ранжирование и релевантность на основе контента	87
4  Краудсорсинговая релевантность	120
ЧАСТЬ 2. Изучение домен-специфичного намерения	147
5  Что такое графы знаний	149
6  Использование контекста для изучения домен-специфичного языка	214
7  Интерпретация намерения запроса через семантический поиск	215
ЧАСТЬ 3. Отраженный интеллект	249
8  Модели бустинга сигналов	251
9  Персонализированный поиск	278
10  Обучение ранжированию для обобщаемой релевантности поиска	356
11  Автоматизация обучения ранжированию с помощью моделей кликов	357
12  Преодоление предвзятости ранжирования с помощью активного обучения.....	390
ЧАСТЬ 4. Передний край поиска.....	421
13  Семантический поиск с плотными векторами	423
14  Ответы на вопросы с помощью тонко настроенной большой языковой модели.....	486
15  Базовые модели и новые парадигмы поиска	520

Оглавление

Предисловие от издательства	13
Введение	14
Предисловие	16
Благодарности	17
Об этой книге	19
О авторах	26
Об иллюстрации на обложке	27
ЧАСТЬ 1. Современная релевантность поиска	29
1 Знакомство с поиском на основе ИИ	31
1.1. Что такое поиск на основе ИИ?	34
1.2. Что такое намерения пользователя	38
1.2.1. Что такое поисковая система?	39
1.2.2. Что предлагают рекомендательные системы	40
1.2.3. Спектр персонализации между поиском и рекомендациями	41
1.2.4. Семантический поиск и графы знаний	43
1.2.5. Что такое измерения намерения пользователя	44
1.3. Как работает поиск на основе ИИ?	45
1.3.1. Основа поиска	47
1.3.2. Отраженный интеллект в петлях обратной связи	47
1.3.3. Бустинг сигналов, совместная фильтрация и обучение ранжированию	48
1.3.4. Интеллект контента и предметной области	50
1.3.5. Генеративный ИИ и RAG	52
1.3.6. Контролируемый искусственный интеллект в сравнении с искусственным интеллектом «черного ящика»	54
1.3.7. Архитектура поисковой системы на основе ИИ	54
Резюме	56
2 Работа с естественным языком	58
2.1. Миф о неструктурированных данных	59
2.1.1. Типы неструктурированных данных	60
2.1.2. Типы данных в традиционных структурированных базах данных	61
2.1.3. Объединения, нечеткие объединения и разрешение сущностей в неструктурированных данных	63
2.2. Структура естественного языка	68
2.3. Распределительная семантика и эмбединги	70
2.4. Моделирование домен-специфичных знаний	76
2.5. Проблемы понимания естественного языка для поиска	79
2.5.1. Проблема неоднозначности (полисемия)	80
2.5.2. Проблема понимания контекста	80
2.5.3. Проблема персонализации	81
2.5.4. Проблемы интерпретации запросов по сравнению с интерпретацией документов	82
2.5.5. Проблемы интерпретации намерения запроса	82
2.6. Контент + сигналы: питающее поиск на основе ИИ	84
Резюме	85

3	Ранжирование и релевантность на основе контента	87
3.1.	Оценка векторов запросов и документов с помощью косинусного сходства	88
3.1.1.	Преобразование текста в векторы	89
3.1.2.	Вычисление сходства между плотными векторными представлениями	90
3.1.3.	Расчет сходства между разреженными векторными представлениями	91
3.1.4.	Частота термина: измерение того, насколько хорошо документы соответствуют термину	94
3.1.5.	Обратная частота документа: измерение важности термина в запросе	99
3.1.6.	TF-IDF: сбалансированная метрика взвешивания для текстовой релевантности	101
3.2.	Управление расчетом релевантности	102
3.2.1.	BM25: стандартный алгоритм сходства текста по умолчанию	102
3.2.2.	Функции, функции, везде!	107
3.2.3.	Выбор мультипликативного или аддитивного бустинга для функций релевантности	111
3.2.4.	Различение сопоставления (фильтрации) и ранжирования (оценки) документов	112
3.2.5.	Логическое соответствие: взвешивание взаимосвязей между терминами в запросе	114
3.2.6.	Разделение задач: фильтрация и оценка	116
3.3.	Реализация пользовательского и домен-специфичного ранжирования релевантности	118
	Резюме	119
4	Краудсорсинговая релевантность	120
4.1.	Работа с пользовательскими сигналами	121
4.1.1.	Контент, сигналы и модели	121
4.1.2.	Настройка наших наборов данных о продуктах и сигналах (RetroTech)	123
4.1.3.	Изучение данных сигналов	127
4.1.4.	Моделирование пользователей, сеансов и запросов	129
4.2.	Знакомство с отраженным интеллектом	130
4.2.1.	Что такое отраженный интеллект?	131
4.2.2.	Популяризированная релевантность посредством бустинга сигналов	132
4.2.3.	Персонализированная релевантность через совместную фильтрацию	138
4.2.4.	Обобщенная релевантность через обучение ранжированию	140
4.2.5.	Другие модели отраженного интеллекта	142
4.2.6.	Краудсорсинг из контента	143
	Резюме	145
	ЧАСТЬ 2. Изучение домен-специфичного намерения	147
5	Что такое графы знаний	149
5.1.	Работа с графами знаний	150
5.2.	Использование нашей поисковой системы в качестве графа знаний	152

5.3. Автоматическое извлечение графов знаний из контента.....	152
5.3.1. Извлечение произвольных отношений из текста.....	153
5.3.2. Извлечение гипонимов и гиперонимов из текста.....	156
5.4. Изучение намерений путем обхода семантических графов знаний	159
5.4.1. Что такое семантический граф знаний?	159
5.4.2. Индексирование наборов данных.....	161
5.4.3. Структура SKG	161
5.4.4. Расчет весов ребер для измерения связанности узлов	164
5.4.5. Использование SKG для расширения запроса	168
5.4.6. Использование SKG для рекомендаций на основе контента.....	173
5.4.7. Использование SKG для моделирования произвольных отношений.....	176
5.5. Использование графов знаний для семантического поиска	179
Резюме	179

6 Использование контекста для изучения домен-специфичного языка181

6.1. Классификация намерения запроса	182
6.2. Устранение неоднозначности смысла запроса	185
6.3. Изучение связанных фраз из сигналов запроса.....	191
6.3.1. Просмотр журналов запросов для поиска связанных запросов	192
6.3.2. Поиск связанных запросов через взаимодействие продуктов	198
6.4. Обнаружение фраз из пользовательских сигналов	203
6.4.1. Обработка запросов как сущностей.....	204
6.4.2. Извлечение сущностей из более сложных запросов.....	205
6.5. Ошибки и альтернативные представления	205
6.5.1. Изучение исправлений орфографии из документов	206
6.5.2. Изучение исправлений орфографии по сигналам пользователя	207
6.6. Собираем все вместе.....	214
Резюме	214

7 Интерпретация намерения запроса через семантический поиск ... 215

7.1. Механика интерпретации запроса.....	216
7.2. Индексирование и поиск в наборе данных локальных отзывов.....	219
7.3. Пример сквозного семантического поиска	222
7.4. Конвейеры интерпретации запроса.....	224
7.4.1. Парсинг запроса для семантического поиска	224
7.4.2. Обогащение запроса для семантического поиска	234
7.4.3. Разреженные лексические модели и модели расширения.....	240
7.4.4. Преобразование запроса для семантического поиска.....	243
7.4.5. Поиск с помощью семантически улучшенного запроса	245
Резюме	246

ЧАСТЬ 3. Отраженный интеллект249

8 Модели бустинга сигналов251

8.1. Базовый бустинг сигналов.....	252
8.2. Нормирование сигналов.....	253
8.3. Борьба со спамом сигналов	256
8.3.1. Использование спама сигналов для манипулирования результатами поиска	256
8.3.2. Борьба со спамом сигналов с помощью фильтрации на основе пользователей.....	259

8.4. Объединение нескольких типов сигналов.....	261
8.5. Временные спады и короткоживущие сигналы	264
8.5.1. Обработка нечувствительных ко времени сигналов.....	265
8.5.2. Обработка сигналов, чувствительных ко времени.....	266
8.6. Бустинг во время индексирования или во время запроса: балансировка масштаба и гибкости	269
8.6.1. Компромиссы при использовании бустинга во время запроса	269
8.6.2. Реализация бустинга сигналов во время индексирования.....	271
8.6.3. Компромиссы при реализации бустинга во время индексирования	274
Резюме	277

9 Персонализированный поиск.....278

9.1. Персонализированный поиск или рекомендации	279
9.1.1. Персонализированные запросы.....	281
9.1.2. Рекомендации, управляемые пользователем	282
9.2. Приближения алгоритмов рекомендаций.....	283
9.2.1. Рекомендации на основе контента.....	283
9.2.2. Рекомендации на основе поведения.....	284
9.2.3. Мультимодальные рекомендательные системы.....	286
9.3. Реализация совместной фильтрации.....	287
9.3.1. Изучение скрытых признаков пользователя и предмета с помощью матричной факторизации.....	287
9.3.2. Реализация совместной фильтрации с помощью метода чередующихся наименьших квадратов	292
9.3.3. Персонализация результатов поиска с бустингом рекомендаций....	299
9.4. Персонализация поиска с использованием эмбедингов на основе контента	303
9.4.1. Генерация латентных признаков на основе контента	304
9.4.2. Реализация категориальных ограничений для персонализации.....	307
9.4.3. Интеграция персонализации на основе эмбедингов в результаты поиска	313
9.5. Проблемы с персонализацией результатов поиска.....	319
Резюме	321

10 Обучение ранжированию для обобщаемой релевантности поиска.....322

10.1. Что такое LTR?	323
10.1.1. Выход за рамки ручной настройки релевантности.....	323
10.1.2. Реализация LTR в реальном мире	324
10.2. Шаг 1: список суждений, начиная с обучающих данных.....	327
10.3. Шаг 2: логирование признаков и инжиниринг.....	329
10.3.1. Хранение признаков в современном поисковом движке.....	330
10.3.2. Логирование признаков из корпуса нашего поискового движка	331
10.4. Шаг 3: преобразование LTR в традиционную задачу машинного обучения.....	333
10.4.1. SVMrank: преобразование ранжирования в бинарную классификацию	335
10.4.2. Преобразование нашей задачи обучения LTR в бинарную классификацию	337

10.5. Шаг 4: обучение (и тестирование!) модели	345
10.5.1. Превращение вектора разделяющей гиперплоскости в оценочную функцию	346
10.5.2. Тест-драйв модели	347
10.5.3. Валидация модели	348
10.6. Шаги 5 и 6: загрузка модели и поиск	350
10.6.1. Запуск и использование модели LTR	350
10.6.2. Примечание о производительности LTR	353
10.7. Почистить и повторить	355
Резюме	356

11

Автоматизация обучения ранжированию

<i>с помощью моделей кликов</i>	357
11.1. (Повторное) создание списков суждений из сигналов	359
11.1.1. Генерация неявных вероятностных суждений из сигналов	360
11.1.2. Обучение модели LTR с использованием вероятностных суждений	362
11.1.3. Показатель кликабельности: ваша первая модель кликов	363
11.1.4. Распространенные предвзятости в суждениях	367
11.2. Преодоление предвзятости позиции	368
11.2.1. Определение предвзятости позиции	369
11.2.2. Предвзятость позиции в данных RetroTech	369
11.2.3. Упрощенная динамическая байесовская сеть: модель кликов, которая преодолевает предвзятость позиции	371
11.3. Управление предвзятостью уверенности: не пересматривать свою модель из-за нескольких удачных кликов	377
11.3.1. Проблема низкой достоверности в данных о кликах	377
11.3.2. Использование бета-приорного распределения для моделирования достоверности вероятностным образом	380
11.4. Изучение ваших обучающих данных в системе LTR	387
Оценки SDBN с использованием бета-распределения	388
Резюме	389

12

Преодоление предвзятости ранжирования

<i>с помощью активного обучения</i>	390
12.1. Наш автоматизированный движок LTR в нескольких строках кода	392
12.1.1. Превращение кликов в обучающие данные (глава 11 в одной строке кода)	393
12.1.2. Обучение и оценка модели в нескольких вызовах функций	395
12.2. А/В-тестирование новой модели	397
12.2.1. Выбираем лучшую модель для тестирования	397
12.2.2. Определение А/В-теста в контексте автоматизированного LTR	398
12.2.3. Перевод лучшей модели в А/В-тест	399
12.2.4. Когда «хорошие» модели работают плохо: чему мы можем научиться из неудачного А/В-теста	401
12.3. Преодоление предвзятости представления: знание того, когда исследовать, а когда эксплуатировать	403
12.3.1. Предвзятость представления в обучающих данных RetroTech	405
12.3.2. За пределами ad hoc: вдумчивое исследование с помощью гауссова процесса	407
12.3.3. Изучение результата наших исследований	414

12.4. Разработка, исследование, сбор, сортировка, повторение: надежный автоматизированный цикл LTR	417
Резюме	419

ЧАСТЬ 4. Передний край поиска.....421

13 Семантический поиск с плотными векторами.....423

13.1. Представление смысла посредством эмбедингов.....	424
13.2. Поиск с использованием плотных векторов	426
13.2.1. Краткая информация о разреженных векторах	426
13.2.2. Система поиска по плотным векторам.....	427
13.3. Получение текстовых эмбедингов с помощью трансформер-кодировщика.....	432
13.3.1. Что такое трансформер?.....	432
13.3.2. Открытые предобученные модели трансформеров	434
13.4. Применение трансформеров для поиска.....	435
13.4.1. Использование набора данных Stack Exchange outdoors	436
13.4.2. Тонкая настройка и семантический анализ сходства текста....	439
13.4.3. Знакомство с библиотекой трансформеров SBERT.....	440
13.5. Автозаполнение естественного языка.....	443
13.5.1. Получение фраз существительных и глаголов для нашего словаря ближайшего соседа	444
13.5.2. Получение эмбедингов.....	446
13.5.3. Приближенный поиск ближайших соседей	451
13.5.4. Реализация индекса ANN.....	453
13.6. Семантический поиск с эмбедингами LLM	456
13.6.1. Получение заголовков и их эмбедингов.....	457
13.6.2. Создание и поиск индекса ближайшего соседа.....	458
13.7. Квантизация и обучение представлениям для более эффективного векторного поиска	461
13.7.1. Скалярная квантизация	463
13.7.2. Бинарная квантизация	470
13.7.3. Продуктовая квантизация	472
13.7.4. Репрезентативное обучение Matryoshka	475
13.7.5. Объединение нескольких методов оптимизации векторного поиска	478
13.8. Кросс-кодировщики и би-кодировщики – сравнение	481
Резюме	485

14 Ответы на вопросы с помощью тонко настроенной большой языковой модели.....486

14.1. Обзор модели вопрос–ответ	487
14.1.1. Как работает модель вопрос–ответ.....	487
14.1.2. Шаблон ретривер–ридер.....	493
14.2. Создание обучающего набора данных для модели вопрос–ответ.....	496
14.2.1. Сбор и очистка набора данных вопрос–ответ.....	497
14.2.2. Создание набора silver: автоматическая маркировка данных из предварительно обученной модели.....	499
14.2.3. Обучение с участием человека: ручная коррекция набора silver для получения набора golden.....	502
14.2.4. Форматирование набора golden для обучения, тестирования и проверки.....	503
14.3. Тонкая настройка модели вопрос–ответ	505

14.3.1. Токенизация и формирование наших маркированных данных	507
14.3.2. Настройка модели-тренера	510
14.3.3. Делаем обучение и оцениваем потери	511
14.3.4. Валидация и подтверждение с удерживанием.....	512
14.4. Создание ридера с новой тонко настроенной моделью	513
14.5. Инкорпорация ретривера: использование модели вопрос-ответ с поисковым движком	515
14.5.1. Шаг 1: запрос ретривера.....	515
14.5.2. Шаг 2: вывод ответов из модели ридера	516
14.5.3. Шаг 3: ранкинг ответов.....	517
14.5.4. Шаг 4: возврат результатов путем объединения ретривера, ридера и ранкера.....	518
Резюме	519
15 Базовые модели и новые парадигмы поиска	520
15.1. Что такое базовые модели.....	521
15.1.1. Что можно считать базовой моделью?	522
15.1.2. Обучение, тонкая настройка в сравнении с подсказкой	523
15.2. Генеративный поиск	526
15.2.1. Генерация с дополненной выборкой	529
15.2.2. Суммаризация результатов с использованием базовых моделей.....	532
15.2.3. Генерация данных с использованием базовых моделей	535
15.2.4. Оценка генеративного вывода.....	539
15.2.5. Построение вашей собственной метрики	541
15.2.6. Оптимизация алгоритмических подсказок	543
15.3. Мультимодальный поиск	545
15.3.1. Общие режимы мультимодального поиска.....	545
15.3.2. Реализация мультимодального поиска	548
15.4. Другие появляющиеся парадигмы поиска на основе ИИ.....	554
15.4.1. Разговорный и контекстный поиск	554
15.4.2. Поиск на основе агентов.....	556
15.5. Гибридный поиск	557
15.5.1. Алгоритм RRF	557
15.5.2. Другие алгоритмы гибридного поиска.....	564
15.6. Конвергенция контекстных технологий	566
15.7. Все вышеперечисленное, пожалуйста!.....	568
Резюме	568
Приложение А. Запуск примеров кода	570
A.1. Общая структура примеров кода.....	570
A.2. Извлечение исходного кода.....	571
A.3. Сборка и запуск кода	571
A.4. Работа с Jupyter	573
A.5. Работа с Docker	575
Приложение В. Поддерживаемые поисковые системы и векторные базы данных	576
V.1. Поддерживаемые движки	576
V.2. Динамическая замена движка.....	577
V.3. Абстракции движка и коллекции	578
V.4. Добавление поддержки для дополнительных движков	579
Предметный указатель	581