

# Краткое оглавление

---

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

# Оглавление

Предисловие от издательства .....	13
Введение .....	14
Предисловие .....	16
Благодарности .....	17
Об этой книге .....	19
О авторах .....	26
Об иллюстрации на обложке .....	27
<b>ЧАСТЬ 1. Современная релевантность поиска .....</b>	<b>29</b>
<b>1 Знакомство с поиском на основе ИИ .....</b>	<b>31</b>
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
<b>2 Работа с естественным языком .....</b>	<b>58</b>
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: топливо, питающее поиск на основе ИИ .....	84
Резюме .....	85

<b>3</b>	<b><i>Ранжирование и релевантность на основе контента .....</i></b>	<b><i>87</i></b>
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
<b>4</b>	<b><i>Краудсорсинговая релевантность.....</i></b>	<b><i>120</i></b>
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
	<b>ЧАСТЬ 2. Изучение домен-специфичного намерения.....</b>	<b>147</b>
<b>5</b>	<b><i>Что такое графы знаний .....</i></b>	<b><i>149</i></b>
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 языка .....181**

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 интеллект .....249**

## **8 Модели бустинга сигналов .....251**

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 с помощью моделей кликов.....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 с помощью активного обучения.....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: реранкинг ответов.....	517
14.5.4. Шаг 4: возврат результатов путем объединения ретривера, ридера и реранкера.....	518
Резюме .....	519
<b>15 Базовые модели и новые парадигмы поиска .....</b>	<b>520</b>
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
<b>Приложение А. Запуск примеров кода .....</b>	<b>570</b>
А.1. Общая структура примеров кода.....	570
А.2. Извлечение исходного кода.....	571
А.3. Сборка и запуск кода .....	571
А.4. Работа с Jupyter .....	573
А.5. Работа с Docker .....	575
<b>Приложение В. Поддерживаемые поисковые системы и векторные базы данных .....</b>	<b>576</b>
В.1. Поддерживаемые движки .....	576
В.2. Динамическая замена движка.....	577
В.3. Абстракции движка и коллекции .....	578
В.4. Добавление поддержки для дополнительных движков .....	579
<b>Предметный указатель .....</b>	<b>581</b>



# Предисловие от издательства

---

## Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте [www.dmkpress.com](http://www.dmkpress.com), зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com); при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу [http://dmkpress.com/authors/publish\\_book/](http://dmkpress.com/authors/publish_book/) или напишите в издательство по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com).

## Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в основном тексте или программном коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com), и мы исправим это в следующих тиражах.

## Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com).

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

# Введение

---

В течение последних двух десятилетий информационный поиск был в центре почти каждого аспекта нашего технического существования как людей. Нужно найти факт? Выполните поиск. Хотите посетить новый ресторан? Выполните поиск. Нужно найти начало тропы в горах для вашего похода на выходных? Выполните поиск. Однако для многих инженеров основы того, как работает поиск или как он выходит за рамки простого сопоставления ключевых слов, чтобы действительно раскрыть то, что нужно пользователям от информационной системы, являются загадкой, о которой не рассказывают почти ни на каких курсах и учебных сборах по информатике. Учитывая этот относительный недостаток инструкций в новый золотой век ИИ, сейчас самое время дать *поиску на основе ИИ* возможность оставить свой след в мире, научив читателей основным принципам, необходимым для понимания работы ИИ в любом компьютерном приложении.

В основе всех поисковых систем лежит именно это: раскрытие информации, которое помогает пользователям принимать более обоснованные решения, необходимые для ориентирования в своем мире. Это раскрытие происходит в основном четырьмя способами.

1. Перебор данных, поиск соответствующих фрагментов информации, ранжирование и возврат наиболее важных фрагментов для их синтеза пользователем.

2. Обобщение данных в более мелкие, более удобоваримые формы для совместного использования и совместной работы с помощью визуализаций и других абстракций.

3. Соотнесение данных с другими, в идеале знакомыми фрагментами информации и концептами.

4. Подача любого из этих трех способов вместе с другим контекстом от пользователя в большой языковой модели (LLM) для дальнейшего синтеза, суммаризации и освоения при постоянном обновлении на основе отзывов пользователей во взаимодействии с ними.

За эти же два десятилетия, когда поиск в нашей жизни на уровне потребителя стал повсеместным, поисковые системы, управляющие этим миром, такие как Google, Elasticsearch, Apache Solr и др., эволюционировали, чтобы делать не только поиск и ранжирование методами, о которых сказано выше, но и решать другие задачи, и не только для текстовых данных, но и для всех других видов данных.

Поисковые системы продвинулись вперед в решении этих задач путем глубокого внедрения статистического анализа, машинного обучения, больших языковых моделей и обработки естественного языка; другими словами, интегрируя методы искусственного интеллекта

в каждый элемент своего ядра. И все же, несмотря на глубину и охват возможностей поисковых систем, эти методы обычно упускают из виду, как то, что, собственно, и производит «поиск по ключевым словам».

В книге «Поиск на основе искусственного интеллекта» Трей, Дуг и Макс изложили насыщенное и подробное руководство, предназначенное для того, чтобы провести инженеров через все детали создания интеллектуальных информационных систем, используя все доступные средства: LLM, домен-специфичные знания, базы знаний и графы и, наконец, пользовательские и краудсорсинговые сигналы. Примеры в книге иллюстрируют важнейшие понятия доступными и простыми для понимания способами.

Как человек, потративший большую часть своей карьеры на создание, обучение и продвижение поиска как средства, помогающего решать некоторые из самых важных проблем нашего времени, я был свидетелем моментов, которые дают старт и направляют инженеров (после того, как они прорываются через неопределенность, присущую работе с запутанными мультимодальными данными) на построение своей дальнейшей карьеры в одной из самых сложных и интересных областей нашего времени. Я надеюсь, что, прочитав эту книгу, вы тоже найдете бесконечное очарование в мире поиска.

Счастливого поиска!

– Грант Ингерсолл, генеральный директор  
и основатель Develomentor LLC,  
Opensearch Leadership Committee

# Предисловие

---

Спасибо за то, что вы приобрели книгу «Поиск на основе искусственного интеллекта»! Эта книга даст вам знания и навыки, необходимые для разработки высокоинтеллектуальных поисковых приложений, которые могут автоматически обучаться на основе каждого обновления контента и взаимодействия с пользователем, постоянно предоставляя все более релевантные результаты поиска.

Нет лучшего времени, чем сейчас, чтобы узнать, как реализовать поиск на основе ИИ. С развитием генеративного ИИ такие методы, как RAG (генерация, дополненная результатами поиска, англ. *Retrieval Augmented Generation*), стали фактическим способом обеспечения систем ИИ актуальными и надежными данными, на основе которых можно получать ответы. Тем не менее «R» в RAG часто является наименее понятным аспектом создания таких систем. Эта книга дает глубокое представление о том, как хорошо выполнять *поиск и извлечение* информации на основе ИИ независимо от того, используете ли вы это для поддержки системы ИИ, создания традиционного поискового приложения или создания новейшего приложения, требующего интеллектуального ранжирования и сопоставления.

За свою карьеру я имел возможность глубоко погрузиться в релевантность поиска, семантический поиск, персонализированный поиск и рекомендации, обработку поведенческих сигналов, семантические графы знаний, обучение ранжированию, LLM и другие базовые модели, поиск по плотным векторам и многие другие возможности поиска на основе ИИ; публиковать исследования в ведущих журналах и на конференциях и, что еще важнее, создавать и поставлять работающее программное обеспечение в больших масштабах. Как основатель Searchkernel, бывший главный специалист по алгоритмам Lucidworks и старший вице-президент по инжинирингу, я также помог поставить многие из этих возможностей сотням самых продвинутых инновационных компаний в мире, чтобы помочь им усилить возможности поиска, которые вы, вероятно, используете каждый день. Я также рад, что Дуг Тернбулл (Reddit, ранее Shopify) и Макс Ирвин (Max.io, ранее OpenSource Connections) также стали соавторами этой книги, опираясь на свой многолетний практический опыт помощи компаниям и клиентам в области поиска и инжиниринга релевантности. В этой книге мы собрали наш многолетний совокупный опыт в практическое руководство, которое поможет вам вывести ваши поисковые приложения на новый уровень. Вы узнаете, как заставить ваши приложения постоянно учиться лучше понимать ваш контент, пользователей и домен, чтобы предоставлять оптимально релевантный опыт при каждом взаимодействии с пользователем.

С наилучшими пожеланиями, когда вы начнете применять поиск на основе ИИ на практике!

—Трей Грейнджер

# Благодарности

---

Прежде всего я хочу поблагодарить мою жену Линдси и моих детей Мелоди, Талли и Оливию. Вы поддерживали меня все эти долгие ночи и выходные, которые я провел за написанием этой книги, и я бы не смог сделать это без вас. Я люблю вас всех!

Далее я хотел бы поблагодарить Дуга Тернбулла и Макса Ирвина за их вклад в эту книгу и в область поиска с использованием искусственного интеллекта (и в информационный поиск в целом). Дуг написал большую часть глав 10–12, а Макс – большую часть глав 13–14 и часть главы 15. Я многому научился у вас обоих и на вашей карьере, и я благодарен за возможность работать с вами над этой книгой.

Далее я хотел бы поблагодарить моего редактора по развитию издательства Manning Марину Майклс. Спасибо за Вашу поддержку и терпение, особенно с учетом того, что сроки выполнения этого масштабного проекта растянулись из-за моей работы в нескольких стартах в ходе производства этого проекта. Качество книги во многом обусловлено Вашим опытом и руководством.

Спасибо также всем остальным сотрудникам Manning, которые работали со мной над разработкой и продвижением проекта: Джону Гатри по технической разработке, Ивану Мартиновичу по ранним выпускам, Майклу Стивенсу по общему видению и направлению и всей маркетинговой команде Manning. Я также благодарю производственную команду Manning за всю их тяжелую работу по форматированию и набору этой книги.

Особая благодарность Гранту Ингерсоллу за написание предисловия. За эти годы я многому научился у Вас, и я очень благодарен за Вашу поддержку.

Далее я хотел бы поблагодарить дополнительных технических авторов книги:

- Дэниела Крауча за его тщательный обзор рукописи книги, его обширный рефакторинг кодовой базы книги и его работу над тем, чтобы сделать книгу в основном не зависящей от конкретных поисковых систем, путем интеграции поддержки plug-and-play для нескольких популярных поисковых систем и векторных баз данных;
- Алекса Отта за его многочисленные технические обзоры книги и за его многочисленные раунды вклада в улучшение кодовой базы книги;
- Мохаммеда Кораема, доктора наук, за его сотрудничество и реализацию алгоритмов обучения графа знаний на основе пользовательских сигналов (глава 6) и персонализированных методов поиска с использованием эмбедингов (глава 9);

- Чао Хан, доктора наук, за его сотрудничество в разработке алгоритмов на основе сигналов для домен-специфичного обнаружения фраз и исправления орфографии.

Я также хотел бы поблагодарить многочисленных читателей, которые предоставили отзывы о ранних версиях этой книги, пока она была в производстве. Ваши отзывы оказали значительное влияние на качество книги. Наконец, я хотел бы поблагодарить рецензентов, которые потратили свое драгоценное время на прочтение рукописи на разных этапах ее разработки и предоставили бесценные отзывы, это Абдул-Басит Хафиз, Адам Дудчак, Эл Кринкер, Ален Кунио, Альфонсо Хесус Флорес Альварадо, Остин Стори, Бхагван Коммади, Дэвид Меза, Давиде Кадамуро, Давиде Фиорентино, Дерек Хэмптон, Гаурав Мохан Тули, Джордж Сейф, Ховард Уолл, Ян Пойнтер, Ишан Курана, Джон Касевич, Кейт Ким, Ким Фальк Йоргенсен, Мария Ана, Марк Джеймс Миллер, Мартин Бир, Мэтт Уэлк, Максим Волгин, Милорад Имбра, Ник Ракочи, Пьерлуиджи Рити, Ричард Воган, Сатей Кумар Саху, Сен Сюй, Шрирам Мачарла, Стив Роджерс, Сумит Пал, Томас Хаук, Тиклу Гангули, Тони Холдройд, Венката Маррапу, Видья Винай, Юдхиеш Равиндранат и Зорозайи Мукуя. Ваши предложения помогли сделать эту книгу лучше.

– Трей Грейнджер

# Об этой книге

---

«Поиск на основе искусственного интеллекта» – пособие по тому, как создавать передовые поисковые системы, которые постоянно обучаются как у ваших пользователей, так и у вашего контента, чтобы обеспечить более предметно-ориентированный и интеллектуальный поиск. Вы изучите современные методы поиска, основанные на науке о данных, такие как:

- семантический поиск с использованием плотных векторных эмбедингов из базовых моделей;
- генерация, дополненная результатами поиска (RAG);
- ответы на вопросы и суммаризация, объединяющие поиск и большие языковые модели (LLM);
- тонкая настройка LLM на основе трансформеров;
- персонализированный поиск на основе пользовательских сигналов и векторных эмбедингов;
- сбор поведенческих сигналов пользователей и построение моделей бустинга сигналов;
- семантические графы знаний для домен-специфичного обучения;
- мультимодальный поиск (гибридные запросы по тексту, изображению, видео и другим типам);
- реализация обобщаемых моделей машинного ранжирования (обучение ранжированию);
- построение кликовых моделей для автоматизации машинного ранжирования;
- методы оптимизации векторного поиска, такие как поиск ANN, квантизация, обучение представлению и би-кодировщики в сравнении с кросс-кодировщиками;
- генеративный поиск, гибридный поиск и передний край поиска.

От современных поисковых систем ожидается, что они будут умными, понимать нюансы запросов на естественном языке, а также предпочтения и контекст каждого пользователя. Эта книга дает вам возможность создавать поисковые системы, которые используют взаимодействие с пользователем и скрытые семантические связи в вашем контенте для автоматического предоставления лучшего, более релевантного результата поиска. Вы даже узнаете, как интегрировать LLM и мультимодальные базовые модели, чтобы значительно ускорить возможности вашей поисковой технологии.



## Кому следует прочитать эту книгу

Эта книга предназначена для инженеров поисковых систем, инженеров-программистов и специалистов по данным, которые хотят узнать, как создавать передовые поисковые системы, интегрирующие новейшие методы машинного обучения, чтобы обеспечить более предметно-ориентированный и интеллектуальный поиск. В книге также представлен подробный обзор поиска на основе ИИ для продакт-менеджеров и руководителей предприятий, которые, возможно, не смогут реализовать эти методы самостоятельно, но хотят понять возможности и ограничения поиска на основе ИИ.

Технические читатели, которые хотят извлечь максимальную пользу из этой книги, могут следовать примерам кода Python. Предполагается знакомство с синтаксисом SQL (язык структурированных запросов), поскольку мы решили реализовать многие агрегации данных в этом стандартизированном представлении, когда это возможно. Базовое понимание того, как работают поисковые системы (такие как Elasticsearch, Apache Solr или OpenSearch) или векторные базы данных, также полезно, но не обязательно.

## Как организована эта книга: дорожная карта

Книга состоит из 4 разделов, которые включают 15 глав. Часть 1 знакомит с поиском на основе ИИ и современной релевантностью поиска:

- глава 1 содержит обзор поиска на основе ИИ, включая основные понятия и методы, лежащие в основе остальной части книги;
- глава 2 охватывает работу с естественным языком, предоставляя базовые сведения о структуре языка и о том, как он позволяет обучать интеллект на данных;
- глава 3 охватывает основы релевантности поиска, объясняя методы сопоставления и ранжирования, использующие векторные эмбединги и сопоставление ключевых слов;
- глава 4 знакомит с краудсорсинговой релевантностью, охватывая сбор и обработку сигналов взаимодействия с пользователем и предоставляя обзор подходов отраженного интеллекта, которые будут использоваться в последующих главах для автоматической оптимизации алгоритмов релевантности поиска.

Часть 2 охватывает домен-специфичное намерение (intent), уделяя особое внимание использованию контента и взаимодействия с пользователем для оптимизации понимания запросов:

- глава 5 знакомит с изучением графа знаний, уделяя особое внимание извлечению явных графов знаний и неявных семантических графов знаний для детального понимания и расширения запросов;
- глава 6 обучает классификации намерений запросов, устранению неоднозначности различных значений слов и фраз, а также использованию как сигналов контента, так и поведенческих сигналов пользователя для изучения домен-специфичной термино-

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

[e-Univers.ru](http://e-Univers.ru)