Краткое оглавление

Оглавление

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*, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу *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/* или напишите в издательство по адресу *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*, 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*.

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 стартапах 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 эмбеддингов (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 Зородзайи 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 эмбеддингов 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 эмбеддингов;
- 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 эмбеддинги 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-