Оглавление

Предисловие от издательства 10
Предисловие 11

Глава 1. Введение 13

1.1. Рекомендательные системы вокруг нас 13
1.2. Эволюция технологий рекомендательных систем 15
1.3. Обзор больших моделей 17
1.4. Интеграция больших моделей и рекомендательных систем 17
Список литературы 19

Глава 2. Основы рекомендательных систем 21

2.1. Обзор базовых рекомендательных алгоритмов 21
2.2. Методы на основе коллаборативной фильтрации 23
2.2.1. Коллаборативная фильтрация 23
2.2.2. Матричная факторизация 25
2.2.3. Ограниченные машины Больцмана 27
2.2.4. Коллаборативная фильтрация с глубоким обучением 29
2.3. Методы рекомендаций на основе признаков 32
2.3.1. Методы на основе машинного обучения 32
2.3.2. Методы рекомендаций на основе глубокого обучения 38
2.4. Методы рекомендаций на основе последовательностей 43
2.4.1. GRU4Rec 44
2.4.2. SASRec 45
2.4.3. DIN 46
2.5. Проектирование и построение рекомендательных систем 48
2.5.1. Рабочий процесс рекомендательной системы 48
2.5.2. Универсальная архитектура рекомендательных систем 53
2.5.3. Холодный старт в рекомендательных системах 57
2.5.4. Конструирование признаков 57
2.6. Заключение 64
Список литературы 64

Глава 3. Основы больших моделей 67

3.1. Обработка естественного языка и языковые модели 67
3.1.1. Обзор обработки естественного языка 67
3.1.2. Языковые модели 69
3.1.3. Корпусы текстов 72
3.2. Трансформер 73
3.2.1. Механизм внимания 74
3.2.2. Архитектура трансформера 77
3.2.3. Предобученные модели на базе трансформера 81

3.3. Построение и применение больших моделей 85
3.3.1. Построение больших моделей 85
3.3.2. Промпт-инженерия для больших моделей 91
3.4. Заключение 95
Список литературы 96

Глава 4. Применение больших моделей в рекомендательных системах 98

4.1. Интеграция больших моделей с рекомендательными системами 98
4.2. Парадигмы применения больших моделей в рекомендательных системах 100
4.2.1. Большая модель как вспомогательный модуль 101
4.2.2. Большая модель непосредственно как модель рекомендаций 104
4.2.3. Большая модель как интеллектуальный агент 106
4.3. Предобучение больших рекомендательных моделей 108
4.3.1. Многозадачное предобучение больших моделей 108
4.3.2. Другие базовые большие модели 114
4.4. Тонкая настройка больших рекомендательных моделей 115
4.4.1. Полная тонкая настройка модели 116
4.4.2. Параметрически эффективная тонкая настройка 117
4.4.3. Настройка промптов 117
4.4.4. Настройка инструкций 119
4.4.5. Значение инструкций и промптов 121
4.5. Часто используемые рекомендательные датасеты 122
4.6. Заключение 123
Список литературы 123

Глава 5. Большие языковые модели как рекомендательные модели 126

5.1. Отбор кандидатов на основе больших моделей 127
5.1.1. Отбор кандидатов на основе текста 128
5.1.2. Отбор кандидатов на основе эмбеддинг-векторов 130
5.1.3. Отбор кандидатов с учетом коллаборативной информации 131
5.2. Рекомендательные системы с учетом последовательностей на основе больших моделей 133
5.2.1. Тонкая настройка и выравнивание 135
5.2.2. Оптимизация с учетом временного фактора 137
5.2.3. Оптимизация с учетом контекста 138
5.3. Ранжирование рекомендаций на основе больших моделей 140
5.3.1. Оптимизация разнообразия ранжирования 142
5.3.2. Оптимизация инвариантности ранжирования 144
5.3.3. Ранжирование элементов в нескольких предметных областях 146
5.4. Слияние коллаборативной информации и семантических знаний 148
5.4.1. Предобучение и тонкая настройка на основе коллаборативной информации 151

5.4.2. Интеграция моделей рекомендаций на основе ID с большими моделями ... 153
5.5. Рекомендательные системы на основе больших моделей, усиленные извлечением из внешних источников знаний ... 154
5.5.1. Парадигмы усиления извлечением ... 155
5.5.2. Усиление извлечением на основе коллаборативной фильтрации ... 157
5.5.3. Усиление извлечением на основе гибридных методов ... 160
5.5.4. Методы усиления извлечением на основе графов ... 162
5.6. Краткие итоги ... 165
Список литературы ... 166

Глава 6. Усиление рекомендательных систем большими моделями ... 168

6.1. Конструирование признаков с использованием больших моделей ... 168
6.1.1. Построение портретов пользователей ... 169
6.1.2. Извлечение признаков элементов ... 175
6.1.3. Комбинация признаков пользователя и элемента ... 179
6.2. Построение графов с помощью больших моделей ... 181
6.2.1. Построение графа логического вывода с помощью больших моделей ... 186
6.2.2. Извлечение отношений комплементарности с помощью больших моделей ... 189
6.2.3. Построение гиперграфа с помощью большой языковой модели ... 191
6.3. Усиление рекомендательных систем при холодном старте с помощью больших языковых моделей ... 194
6.3.1. Большая языковая модель как модель холодного старта ... 195
6.3.2. Преодоление проблемы холодного старта с помощью эмбеддингов большой языковой модели ... 196
6.3.3. Совместное обеспечение холодного и горячего старта ... 198
6.3.4. Генерация данных взаимодействий пользователь–элемент ... 200
6.4. Рекомендательные системы на основе дистилляции больших языковых моделей ... 204
6.4.1. Дистилляция больших языковых моделей ... 205
6.4.2. Дистилляция больших языковых моделей для улучшения рекомендательных систем ... 207
6.4.3. Дистилляция промптов в многозадачных рекомендательных системах ... 209
6.4.4. Интеграция знаний большой языковой модели и коллаборативной информации ... 211
6.5. Заключение ... 213
Список литературы ... 214

Глава 7. Надежные рекомендательные системы на основе больших моделей ... 217

7.1. Интерпретируемость рекомендательных систем ... 218
7.1.1. Обзор интерпретируемости ... 218

7.1.2. Тонкая настройка больших моделей для интерпретируемых рекомендательных систем 219
7.2. Справедливость рекомендательных систем 223
7.2.1. Обзор справедливости 223
7.2.2. Справедливость элементов 224
7.2.3. Справедливость пользователей 226
7.3. Приватность рекомендательных систем 229
7.3.1. Обзор приватности и федеративного обучения 229
7.3.2. Методы федеративного обучения больших моделей 230
7.3.3. Достижение баланса в федеративных рекомендательных системах 233
7.4. Актуальность рекомендательных систем 236
7.4.1. Обновление модели и катастрофическое забывание 236
7.4.2. Непрерывное обучение больших моделей 237
7.4.3. Инкрементальное обучение рекомендательных систем на основе больших моделей 239
7.5. Забывание в рекомендательных системах 240
7.5.1. Обзор забывания 240
7.5.2. Приближенное забывание 241
7.5.3. Точное забывание 243
7.6. Заключение 244
Список литературы 245

Глава 8. Рекомендательные системы на основе интеллектуальных агентов с большими моделями 249

8.1. Интеллектуальные агенты на основе больших моделей 249
8.1.1. Одноагентный интеллектуальный агент 251
8.1.2. Многоагентные интеллектуальные системы 251
8.2. Применение интеллектуальных агентов на основе больших моделей в рекомендательных системах 255
8.3. Рекомендательные системы на основе интеллектуальных агентов с большими моделями 261
8.3.1. Интеллектуальные агенты, ориентированные на пользователя 261
8.3.2. Интеллектуальные агенты, ориентированные на взаимодействие пользователь–элемент 264
8.3.3. Рекомендательные системы на основе сотрудничества многоагентных интеллектуальных систем 266
8.4. Заключение 269
Список литературы 270

Глава 9. Оценка и развертывание рекомендательных систем на основе больших моделей 272

9.1. Методы оценки больших моделей 272
9.1.1. Общие методы оценки 272
9.1.2. Методы оценки для конкретных задач 274

9.2. Методы оценки рекомендательных систем 276
9.2.1. Офлайн-метрики рекомендательных систем 276
9.2.2. Онлайн-метрики рекомендательных систем 283
Другие метрики оценки рекомендаций 286
9.3. Развертывание больших моделей 286
9.3.1. Фреймворки высокопроизводительного и пакетного развертывания 287
9.3.2. Фреймворки гибкого и кросс-платформенного развертывания 290
9.4. Сжатие больших моделей 293
9.4.1. Квантование больших моделей 293
9.4.2. Обрезка больших моделей 295
9.5. Создание рекомендательной системы на основе больших моделей с помощью фреймворка LangChain 297
9.5.1. Построение этапа отбора кандидатов 298
9.5.2. Создание интеллектуального агента для рекомендаций 301
9.6. Заключение 303
Список литературы 303

Глава 10. Итоги и перспективы 305

10.1. Является ли большая модель «серебряной пулей»? 306
10.2. Открытые вопросы 307
10.3. 60 вопросов о больших моделях и рекомендациях 307
Предметный указатель 311

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.

Предисловие

Рекомендательные системы как базовая инфраструктура эпохи интернета уже проникли во все виды цифровых платформ и прикладных сценариев. С момента появления концепции в 1992 году, пройдя более чем 30-летнюю эволюцию, рекомендательные системы превратились из теоретических исследований в зрелую технологическую область, сочетающую академическую глубину и коммерческую ценность. В последние годы под двойным воздействием взрывного роста объемов данных и скачкообразного повышения вычислительных мощностей были достигнуты прорывные успехи в технологиях машинного обучения и глубокого обучения, что привело рекомендательные системы к беспрецедентным технологическим преобразованиям. Появление нового поколения технологий искусственного интеллекта, представленного большими языковыми моделями (такими как серия GPT от OpenAI, DeepSeek от DeepSeek и др.), не только перестроило технологическую парадигму рекомендательных систем, но и обеспечило качественный скачок в точности рекомендаций, адаптивности к сценариям и пользовательском опыте взаимодействия.

В традиционном процессе разработки рекомендательных систем команды разработчиков обычно вынуждены вкладывать огромные усилия в извлечение данных и оптимизацию алгоритмов: день за днем анализировать признаки поведения пользователей, пробовать различные модели машинного и глубокого обучения, многократно обучать и подбирать гиперпараметры для повышения эффективности рекомендаций. Такой режим «конструирования признаков + итераций модели» хотя и позволяет в определенной степени поддерживать точность рекомендаций и скорость отклика системы, создавая базовую ценность для бизнеса, однако его ограничения становятся все более очевидными – перед лицом быстро меняющихся бизнес-сценариев и огромных объемов гетерогенных данных инженерам приходится непрерывно искать новые технические решения, но часто они попадают в ловушку убывающей предельной отдачи[1] и не могут добиться качественного прорыва.

Под воздействием технологий больших языковых моделей (LLM) методология и рабочие процессы традиционных рекомендательных систем переживают беспрецедентные вызовы и трансформацию. Концепция «технологического равноправия», представленная DeepSeek, позволяет малым и средним предприятиям получать эффективные возможности вывода и обобщения с относи-

[1] «Ловушка убывающей предельной отдачи» в разработке рекомендательных систем возникает, когда дополнительные инвестиции в улучшение модели – увеличение объема данных, усложнение архитектуры, расширение числа кандидатов на этапе отбора или рост вычислительных ресурсов – приводят к все меньшему приросту ключевых метрик (точность, NDCG, полнота охвата, удовлетворенность пользователей). После достижения определенного уровня качества каждая новая оптимизация дает минимальный эффект при значительных затратах, а иногда вызывает переобучение или ухудшение онлайн-результатов. – *Прим. ред.*

тельно низкими затратами за счет локального развертывания больших моделей[1], тем самым значительно повышая точность и адаптивность рекомендательных систем. Этот тренд не только ускоряет итеративное обновление технологий рекомендаций, но и продвигает общую смену парадигмы в отрасли.

В настоящее время революционные технологии искусственного интеллекта перестраивают все отрасли. Как практикующие специалисты в области рекомендательных систем, мы не только имеем возможность следовать за волной развития технологий больших моделей, но и можем активно участвовать в этой технологической революции, помогая предприятиям снижать затраты и повышать эффективность, а также создавать большую ценность для общества.

Именно поэтому мы начали работу над этой книгой. В ней мы, опираясь на собственный накопленный опыт в области больших языковых моделей и рекомендательных систем, в сочетании с передовыми технологическими прорывами отрасли и реальными примерами внедрения, систематически изложим, как применять технологии больших языковых моделей при проектировании, разработке и реализации рекомендательных систем. Наша цель – с помощью доступного и понятного изложения помочь читателям сформировать целостную когнитивную методологию применения больших языковых моделей для усиления рекомендательных систем, овладеть ключевыми путями технологического усовершенствования и развить способность к предвидению будущих тенденций развития отрасли.

Мы хотели бы выразить особую благодарность родным, друзьям и коллегам за их всестороннюю поддержку в процессе написания книги – именно благодаря вашей бескорыстной помощи мы смогли полностью сосредоточиться на творчестве. Также искренне благодарим редакторскую команду издательства за профессиональное руководство и ценные рекомендации. Наконец, выражаем глубокое уважение пионерам отрасли, продвигающим инновационное слияние больших языковых моделей и рекомендательных систем, – ваши передовые исследования предоставили богатую теоретическую основу и практические ориентиры для этой книги.

Лю Лу, Чжан Юйцзюнь

[1] Если не оговорено особо, под «большими моделями» в настоящей книге понимаются большие языковые модели. – *Прим. авт.*

Глава 1

Введение

В эпоху информационного взрыва обычным пользователям все труднее справляться с обработкой колоссального объема информации в виде описаний товаров, новостей и рекламы. Именно в таких условиях возникли рекомендательные системы и стали неотъемлемой частью интернета.

Рекомендательные системы незаметно работают как в потребительских сервисах (2C) – электронная коммерция, новостные порталы, онлайн-обучение, стриминг аудио и видео, – так и в корпоративных решениях (2B): оптимизация цепочек поставок, маркетинг, управление клиентскими отношениями, подбор персонала. Анализируя поведение и предпочтения пользователей, они позволяют бизнес-системам быстро находить в гигантских массивах данных именно тот контент, который действительно нужен пользователю, существенно повышая качество пользовательского опыта.

1.1. Рекомендательные системы вокруг нас

В условиях информационной перегрузки интернета пользователю крайне сложно самостоятельно отыскать интересный контент – здесь решающую роль играют рекомендательные системы. Их задача – предоставлять персонализированные рекомендации для онлайн-продуктов и сервисов, улучшая пользовательский опыт.

Рекомендательные системы встречаются повсюду. Вот типичные сценарии применения:

- в сфере 2C: рекомендованные плейлисты в музыкальных сервисах, товарные рекомендации в интернет-магазинах, персонализированные новостные ленты, подбор курсов на образовательных платформах;
- в сфере 2B: поиск надежных поставщиков, выявление потенциальных клиентов, оптимизация маркетинговых стратегий, подбор кандидатов.

Основной принцип – использование предпочтений пользователя (user), характеристик элемента[1] (item) и данных о взаимодействиях (покупки, клики

[1] В рекомендательных системах элементы – это объекты, предлагаемые пользователю. Основные виды: товары (e-commerce), медиаконтент (фильмы, музыка, видео, подкасты), новостные статьи, вакансии, пользователи (социальные сети, дейтинг), места и заведения (карты, туризм), образовательные курсы, реклама, игры и приложения. По свойствам элементы делятся на долгоживущие (книги, фильмы) и короткоживущие (новости, тренды), а также холодные (новые без истории) и горячие (популярные). – *Прим. ред.*

и т. д.) для фильтрации и ранжирования огромного количества элементов с целью предложить пользователю те, которые с наибольшей вероятностью ему понравятся. Таким образом, рекомендательная система выступает «мостом» между производителями контента и потребителями. Хотя область применения не ограничена интернетом, именно в информационно насыщенной веб-среде ее ценность проявляется особенно ярко.

Рекомендательная система – мощная и сложная серверная инфраструктура, включающая обработку больших данных, разработку алгоритмов и бизнес-логику. Она моделирует сложные взаимосвязи между пользователями, элементами и их взаимодействиями. Обычно состоит из модулей моделирования признаков пользователей и элементов, этапов грубого отбора (recall) и ранжирования (ranking), а также сложной инженерной практики: архитектурные итерации, обновление моделей в реальном времени и офлайн.

Основные задачи рекомендательных систем:

- снижение информационной перегрузки, помощь в быстром получении ценного или интересного контента;
- глубокое понимание пользователей для предоставления персонализированного контента и услуг;
- обнаружение новых элементов, соответствующих пользовательским предпочтениям, что помогает бизнесу расширять рынки;
- повышение кликабельности, конверсии, удержания пользователей и, как следствие, доходов компании.

Рекомендательные системы – одновременно «двигатель» эффективного доступа пользователей к контенту и «двигатель» достижения коммерческих целей интернет-компаний. Эти две роли взаимно усиливают друг друга, продвигая развитие персонализированных сервисов.

С точки зрения пользователя система анализирует исторические данные и «предугадывает» потенциально интересный контент. Потребности пользователей разнообразны: иногда нужен конкретный товар, но выбор слишком велик; иногда пользователь не может четко сформулировать запрос; иногда он сам точно не знает, что ему нужно. В отличие от поисковых систем, требующих явного запроса, рекомендательные системы работают проактивно, автоматически предлагая контент на основе накопленной истории.

С точки зрения бизнеса рекомендательные системы помогают глубже понимать пользователей, удерживать их в нужный момент, снижать затраты на взаимодействие, повышать опыт, активность и конверсию, создавая новые коммерческие возможности и обеспечивая устойчивый рост. У разных платформ – разные цели оптимизации: видеоплатформы фокусируются на длительности просмотра и количестве воспроизведений, электронная коммерция – на покупках и выручке, новостные платформы – на репостах и комментариях.

Персонализация – ключевая особенность, требующая богатых данных о поведении пользователя. Развитие больших данных и ИИ радикально ускорило персонализированные рекомендации. Пример с новостями: от традиционных порталов (Sina News) к приложениям вроде Toutiao, которые формируют полностью персонализированную ленту на главной странице. Все современные ре-

комендательные системы строятся вокруг триады «пользователь – контекст – элемент», независимо от вертикали (новости, видео, e-commerce, реклама).

В будущем развитие больших моделей откроет для рекомендательных систем новые горизонты: более глубокое понимание описаний объектов, учет долгосрочных и динамических предпочтений, точное сопоставление признаков объектов и запросов пользователей, что сделает рекомендации интеллектуальнее и создаст дополнительную бизнес-ценность.

1.2. Эволюция технологий рекомендательных систем

По количеству обслуживаемых пользователей различают персонализированные и групповые рекомендательные системы. Персонализированные предсказывают следующий объект для конкретного пользователя – именно они наиболее глубоко исследованы и являются основной темой книги. Основные подходы: коллаборативная фильтрация, контентные и гибридные методы, среди которых коллаборативная фильтрация остается самой распространенной.

- 1992 – Белкин и Крофт показали, что рекомендательные системы основаны на информационной фильтрации [1]. Вскоре Голдберг с соавт. при разработке Tapestry ввели термин «коллаборативная фильтрация» (Collaborative Filtering, CF) [2]. В последующие годы CF стала доминирующей технологией в новостях, музыке, видео, e-commerce и кино.
- 2006–2009 – конкурс Netflix Prize привлек внимание к матричным разложениям. Работа Корен с соавт. (2009) систематизировала подход и изменила стандарты оценки [3].
- 2007 – логистическая регрессия[1] (LR) для предсказания CTR[2] в рекламе [4].
- 2010 – факторизационные машины (FM) [5], 2016 – FFM с учетом полей признаков[3] (Field-aware Factorization Machines) [6].
- 2014 – Facebook предложил гибрид GBDT + LR [8].

---

[1] Логистическая регрессия – это статистический метод машинного обучения для задач бинарной классификации. Модель предсказывает вероятность принадлежности объекта к положительному классу (от 0 до 1) с помощью сигмоидной функции. Обучается путем максимизации логарифмического правдоподобия (или минимизации кросс-энтропии) с помощью градиентного спуска. Несмотря на название «регрессия», это классификатор. Простая, интерпретируемая, хорошо работает на линейно разделимых данных и часто используется как базовая модель. – *Прим. ред.*

[2] CTR (Click-Through Rate, или коэффициент кликабельности) – это показатель, который измеряет процент пользователей, кликнувших на определенный элемент (например, рекламу или ссылку), относительно общего числа пользователей, которые его увидели. В контексте искусственного интеллекта CTR часто используется для оценки эффективности рекламных кампаний, рекомендационных систем и других сервисов, зависящих от пользовательского взаимодействия. – *Прим. ред.*

[3] FFM (Field-aware Factorization Machines) – это расширение классической модели FM, которое специально предназначено для задач с категориальными признаками, разбитыми на разные «поля» (fields), например пользователь, товар, категория, бренд, время суток и т. д. В отличие от обычных FM, где все признаки взаимодействуют через одну общую матрицу скрытых факторов, в FFM каждый признак из одного поля имеет отдельный набор скрытых векторов для каждого другого поля. Это позволяет модели более точно захватывать специфические взаимодействия между разными типами признаков. – *Прим. ред.*

- 2017 – Alibaba представила смешанную логистическую регрессию (MLR, Mixed Logistic Regression) [7].
- После триумфа AlphaGo начался бум глубоких нейронных сетей (Deep Neural Networks, DNN) в рекомендательных системах. YouTube существенно повысил точность видеорекомендаций с помощью DNN [9]; технологию быстро подхватили Baidu, Alibaba, Tencent и др.

В области рекомендательных систем одним из важнейших применений глубокого обучения стали модели векторных представлений, или эмбеддинги (embedding). Они способны преобразовывать дискретную информацию в непрерывные числовые представления. Технология векторных представлений изначально возникла в обработке естественного языка; фундамент для нее заложил алгоритм word2vec, предложенный Миколовым с соавт. [10]. Благодаря своей простоте и высокой эффективности векторные представления очень быстро были перенесены в рекомендательные системы.

Применение глубокого обучения (Deep Learning, DL) в рекомендательных системах в основном развивалось вокруг двух ключевых направлений: автоматизации конструирования признаков и повышения способности моделировать пересечения признаков. Эта эволюция прошла через ряд знаковых моделей – от DSSM [11], Wide&Deep [12] и DeepFM [13] до более поздних решений. Начиная с 2018 года в рекомендательные системы активно внедряется механизм внимания (attention mechanism), что позволило значительно точнее моделировать предпочтения пользователей. SASRec [14] – одна из первых последовательностных рекомендательных моделей, полностью построенная на механизме самовнимания (self-attention); она способна одновременно улавливать долгосрочные и краткосрочные предпочтения пользователя из истории его поведения. Alibaba последовательно предложила целую линейку моделей на основе внимания: DIN [15] (Deep Interest Network), DIEN [16] (Deep Interest Evolution Network), MIMN [17] (Multi-channel Interest Memory Network), DSIN [18] (Deep Session Interest Network).

Все они используют механизмы внимания для отслеживания эволюции интересов пользователя во времени. Ранние рекомендательные модели в основном опирались на коллаборативную фильтрацию[1] и матричную факторизацию[2] (Matrix Factorization, MF) – они были сравнительно простыми и содержали ограниченное число параметров. С появлением глубоких нейронных сетей выразительная способность моделей резко возросла, а производительность существенно улучшилась. В частности, последовательные модели на основе внимания способны выделять ключевую информацию из длинных цепочек взаимодействий пользователя с элементами, что заметно повышает точность рекомендательных систем. В последние годы, с раз-

[1] Коллаборативная фильтрация – класс методов рекомендаций, которые предсказывают интерес пользователя к элементу, опираясь только на историю взаимодействий множества пользователей и элементов (матрица «пользователь–элемент»), без использования содержательных признаков контента. – *Прим. ред.*

[2] Матричная факторизация – семейство методов совместной фильтрации, которые раскладывают разреженную матрицу «пользователь × элемент» (размером m×n) на произведение двух матриц низкого ранга. – *Прим. ред.*

витием больших мультимодальных моделей (Large-Scale Models, LSM), появилась возможность решать множество рекомендательных задач внутри одной единой модели и гибко адаптировать ее под различные даунстрим-сценарии. Это кардинально меняет традиционные бизнес-процессы построения рекомендательных систем.

1.3. Обзор больших моделей

Большие языковые модели (LLM, Large Language Models), или просто большие модели, – языковые модели на базе глубоких нейронных сетей с десятками и сотнями миллиардов параметров, обученные методом самообучения на огромных объемах неразмеченного текста [19].

После появления трансформеров[1] (Transformer) [20] развитие ускорилось. К 2024 году OpenAI, Google, Meta, Baidu, iFlyTek, Alibaba, Huawei и многие другие выпустили десятки мощных моделей, показывающих выдающиеся результаты в задачах обработки естественного языка. Релиз ChatGPT в ноябре 2022 года стал поворотным моментом, продемонстрировав, что одна модель (GPT-3.5/4) способна решать задачи понимания и генерации текста, включая программирование, перевод и диалог.

В отличие от традиционных моделей глубокого обучения с умеренным числом параметров, большие модели содержат от десятков миллионов до триллионов параметров. Чем больше масштаб, тем выше обобщающая способность. При достижении критического размера возникают «эмерджентные способности» (emergent abilities) [21] – новые сложные свойства, которые невозможно предсказать по поведению малых моделей.

Большие модели не только демонстрируют владение знаниями о мире и глубокое понимание языка, но и обладают большей выразительной силой по сравнению с традиционными моделями машинного обучения и глубокого обучения. Они могут объединять знания о мире для выявления потенциальных закономерностей и ключевой информации в данных, а также адаптироваться к различным задачам и бизнес-потребностям на основе бизнес-сценариев и характеристик данных.

1.4. Интеграция больших моделей и рекомендательных систем

Возникновение больших моделей (LLM) принесло революционные изменения в обработку естественного языка, и индустрия активно использует их возможности для значительного повышения эффективности рекомендательных

[1] Трансформер – это архитектура нейронной сети, разработанная для обработки последовательностей данных, таких как текст. Она основана на механизме внимания, который позволяет модели фокусироваться на разных частях входной последовательности при генерации выходных данных. Трансформеры способны параллельно обрабатывать данные, что значительно ускоряет обучение и улучшает качество работы с длинными последовательностями. Эта архитектура стала основой для многих современных моделей в области обработки естественного языка, таких как BERT и GPT. – *Прим. ред.*

систем [22]. По сравнению с традиционными рекомендательными системами ключевое преимущество интеграции больших моделей заключается в следующем: они способны извлекать высококачественные векторные представления текстовых признаков и обладают огромным объемом знаний о мире, накопленным в процессе предобучения [23].

Большие модели мастерски улавливают контекстную информацию, что позволяет им глубже понимать текстовые данные пользователей (запросы, отзывы на заказы, описания товаров и т. д.) [24]. В результате существенно повышается точность рекомендательных систем и удовлетворенность пользователей.

Кроме того, способности больших моделей к нулевому обучению[1] и обучение на нескольких примерах[2] по промптам (zero-shot learning и few-shot learning) [25] открывают эффективный путь решения проблемы холодного старта[3] в условиях разреженных исторических взаимодействий, создавая новые практические возможности для рекомендательных систем.

Выдающаяся обобщающая способность больших моделей позволяет рекомендовать ранее не встречавшиеся элементы-кандидаты. Это достигается благодаря тому, что в процессе предобучения модель накопила огромное количество фактических знаний, профессиональных сведений в предметной области (доменных сведений) и здравого смысла. Даже столкнувшись с совершенно новым пользователем или новым товаром, она способна выдавать обоснованные и качественные рекомендации.

В настоящее время интеграция больших моделей и рекомендательных систем стала одним из важнейших исследовательских направлений как в академической, так и в среде разработчиков. Большие модели находят широкое применение на всех этапах рекомендательной воронки: от извлечения (retrieval) и ранжирования (ranking) до конструирования признаков[4] и рекомендаций при холодном старте. При этом в процессе практического внедрения больших моделей в рекомендательные системы по-прежнему сохраняется ряд крити-

---

[1] Нулевое обучение – модель решает задачу без единого примера из этой задачи. Ей дается только описание задачи на естественном языке (промпт). Пример: «Переведи на французский: …» – и модель переводит, хотя ее никогда специально не настраивали на перевод. – *Прим. ред.*

[2] Обучение на нескольких примерах – модели дают очень мало примеров (обычно 1–64) прямо в промпте, и она сразу умеет решать задачу. Пример: в промпте 5 примеров «кошка → котенок, собака → щенок», дальше «медведь → ?», модель отвечает «медвежонок». – *Прим. ред.*

[3] Проблема холодного старта (cold start problem) в рекомендательных системах – это ситуация, когда модель не может дать точные рекомендации новому пользователю (нет истории взаимодействий) или новому элементу (нет оценок/просмотров от пользователей). Кратко: отсутствие данных о «холодном» объекте приводит к низкому качеству рекомендаций на старте; решается через контентные признаки, популярные объекты или гибридные подходы. – *Прим. ред.*

[4] Конструирование признаков – это процесс выбора, преобразования и создания новых признаков (или «фич»), которые используются для обучения моделей машинного обучения. Основная цель конструирования признаков заключается в улучшении производительности модели, обеспечивая ей более информативные и релевантные данные. – *Прим. ред.*

чески важных вызовов: интерпретируемость результатов, обеспечение справедливости[1] (fairness), защита конфиденциальности пользователей и другие этические и технические проблемы.

Список литературы

[1] Belkin N.J., Croft W.B. Information filtering and information retrieval: two sides of the same coin? [J]. Communications of the ACM, 1992, 35 (12): 29–38.
[2] Goldberg D., et al. Using collaborative filtering to weave an information TAPESTRY [J]. Communications of the ACM, 1992, 35 (12): 61–70.
[3] Koren Y., Bell R., Volinsky C. Matrix Factorization Techniques for Recommender Systems [J]. Computer, 2009, 42 (8): 30–37.
[4] Richardson M., et al. Predicting Clicks: Estimating the Click-Through Rate for New Ads [C] // WWW2007, 2007.
[5] Rendle S. Factorization Machines [J]. IEEE, 2010.
[6] Juan Y., et al. Field-aware Factorization Machines for CTR Prediction [C] // RecSys, 2016.
[7] Gai K., et al. Learning Piece-wise Linear Models from Large Scale Data for Ad Click Prediction [J]. 2017.
[8] He X., et al. Practical Lessons from Predicting Clicks on Ads at Facebook [M]. 2014.
[9] Covington P., et al. Deep Neural Networks for YouTube Recommendations [C] // RecSys, 2016.
[10] Mikolov T., et al. Efficient Estimation of Word Representations in Vector Space [J]. 2013.
[11] Huang P.S., et al. Learning deep structured semantic models for web search using clickthrough data [C] // CIKM, 2013.
[12] Cheng H.T., et al. Wide & Deep Learning for Recommender Systems [J]. ACM, 2016.
[13] Guo H., et al. DeepFM: An End-to-End Wide & Deep Learning Framework for CTR Prediction.
[14] Kang W.C., Mcauley J. Self-Attentive Sequential Recommendation [C] // ICDM, 2018.
[15] Zhou G., et al. Deep Interest Network for Click-Through Rate Prediction [J]. 2017.
[16] Zhou G., et al. Deep Interest Evolution Network for Click-Through Rate Prediction [C] // AAAI, 2019.
[17] Pi Q., et al. Practice on Long Sequential User Behavior Modeling for Click-Through Rate Prediction [J]. ACM, 2019.

[1] Обеспечение справедливости (fairness) в контексте ИИ относится к процессам и методам, направленным на предотвращение предвзятости и дискриминации в алгоритмах и моделях машинного обучения. Это включает в себя разработку и тестирование систем, которые обеспечивают равное обращение с различными группами пользователей независимо от их расы, пола, возраста или других характеристик. Главное внимание уделяется тому, чтобы результаты, выдаваемые ИИ, были объективными, прозрачными и не приводили к угнетению или стигматизации определенных категорий людей. – *Прим. ред.*

[18] Feng Y., et al. Deep Session Interest Network for Click-Through Rate Prediction [J]. 2019.
[19] Zhao W.X., et al. A Survey of Large Language Models [J]. ArXiv, 2023, abs/2303.18223.
[20] Vaswani A., et al. Attention is All you Need [C], 2017.
[21] Wei J., et al. Emergent Abilities of Large Language Models [J]. ArXiv, 2022, abs/2206.07682.
[22] Chen J. A Survey on Large Language Models for Personalized and Explainable Recommendations [J]. ArXiv, 2023, abs/2311.12338.
[23] Liu P., et al. Pre-train, Prompt, and Recommendation: A Comprehensive Survey [J]. TACL, 2023, 11: 1553–1571.
[24] Geng S., et al. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5) [C] // RecSys, 2022.
[25] Sileo D., et al. Zero-Shot Recommendation as Language Modeling [C], 2021.