

Рекомендательные сервисы в продакшене

Николай Анохин

30 сентября 2021 г.

Обзор модуля

Прежде чем начать

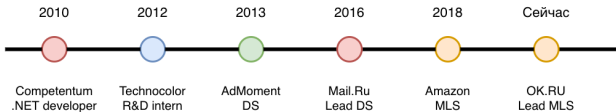
1. Пройдите опрос <https://forms.gle/k1To7wi3QuXXgP2D7>
2. Клонировать репозиторий курса
<https://gitlab.com/fpmi-atp/atp-mobod2021>

Обо мне

Академический опыт



Индустриальный опыт



Навыки



Telegram

- Вопросы вне занятия можно задать в личных сообщениях и чате группы (лучше)
- Тегайте меня, чтобы я не пропустил ваш комментарий в общем потоке сообщений
- Если ответа не последовало в течении 24 часов, то я, вероятно, не увидел ваше сообщение. Не стесняйтесь его продублировать

Как задать вопрос

- Голосом
- В специально выделенное для этого время
- Перед тем как спросить будет хорошим тоном поставить несколько знаков вопроса

20:23 Саша: ????

20:23 Преподаватель: Ждём вопроса от Саши

20:24 Саша: Какая метрика хорошо работает в задаче рекомендаций?

Если что-то пошло не так

- Пропал голос
- Исчезло изображение
- Плохо слышно
- Любые проблемы другого характера

Сразу пишем в чат много минусов и не ждем других участников. Если вы увидели, что в чате кто-то написал много минусов, а у вас всё хорошо, то поставьте несколько плюсов:

20:24 Петя: - - - - -

20:25 Саша: + + + +

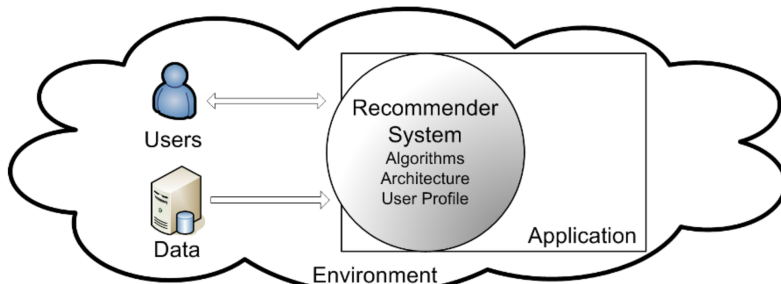
20:25 Ольга: + + + + + + + +

Программа модуля

Дата	Тема	Семинар	Домашка
2021-09-30	Рекомендательные сервисы в продакшене	✓	
2021-10-07	Метрики и базовые подходы	✓	
2021-09-14	Классические алгоритмы	✓	✓
2021-09-21	Нейросетевые рекомендеры	✓	
2021-09-28	Нерешенные проблемы и новые направления	✓	

Зачем нужны рекомендательные сервисы

Recommender Systems (RS) are software tools and techniques providing suggestions for **items** to be of use to a **user** [RRSK10].



Зачем RS бизнесу

- Увеличить продажи

Зачем RS бизнесу

- Увеличить продажи
- Продвигать более разнообразные товары

Зачем RS бизнесу

- Увеличить продажи
- Продвигать более разнообразные товары
- Улучшить пользовательский опыт

Зачем RS пользователям

- Найти лучший товар

Зачем RS пользователям

- Найти лучший товар
- Найти **все** подходящие товары

Зачем RS пользователям

- Найти лучший товар
- Найти **все** подходящие товары
- Найти последовательность или набор товаров

Зачем RS пользователям

- Найти лучший товар
- Найти **все** подходящие товары
- Найти последовательность или набор товаров
- Залипнуть

Зачем RS пользователям

- Найти лучший товар
- Найти **все** подходящие товары
- Найти последовательность или набор товаров
- Залипнуть
- Найти рекоммендер, которому можно доверять

Зачем RS пользователям

- Найти лучший товар
- Найти **все** подходящие товары
- Найти последовательность или набор товаров
- Залипнуть
- Найти рекоммендер, которому можно доверять
- Реализовать творческие потребности

Зачем RS пользователям

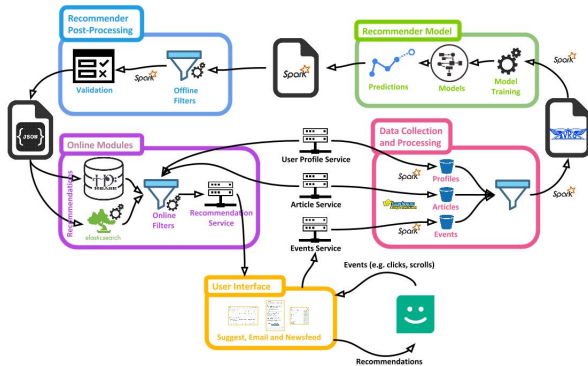
- Найти лучший товар
- Найти **все** подходящие товары
- Найти последовательность или набор товаров
- Залипнуть
- Найти рекоммендер, которому можно доверять
- Реализовать творческие потребности
- Помочь другим сделать выбор

Зачем RS инженерам

- Делать высоконагруженный отказоустойчивый сервис
- Анализировать большие данные
- Окунуться в волшебный мир магии машинного обучения
- Объективно измерять результат своей работы
- Все это за зарплату

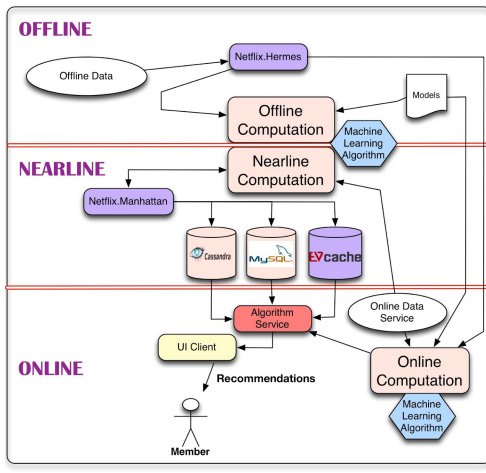
Архитектуры рекомендательных сервисов

Обзор типичных компонентов RS / Mendeley (2016) [JH16]



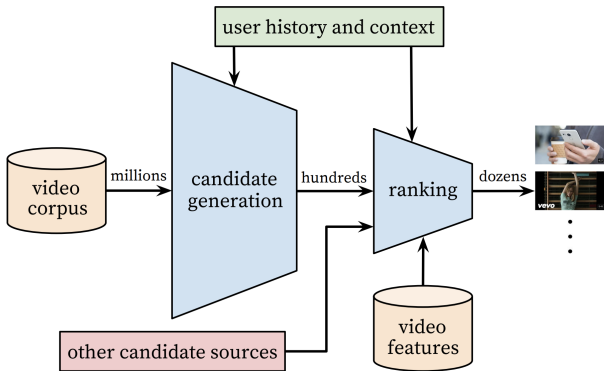
Машинное обучение – небольшая часть рекомендательного сервиса

Подходы к обработке данных / Netflix (2013) [NN13]



Чем ближе вычисления к real-time, тем больше ограничений и компромиссов

Двухступенчатая архитектура / Youtube (2016) [CAS16]



Айтемов так мно-
го, что учесть
полный контекст
не может даже
Google

Загадка

Что общего между

- населением городов
- количеством друзей у пользователей в социальной сети
- размерами лесных массивов
- количеством прослушиваний песен в Spotify

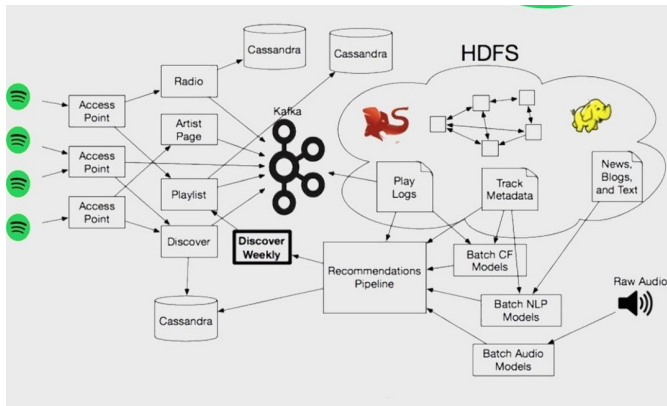
Power law

$$p(x) = \frac{C}{x^\alpha}, \quad x > x_{min}$$



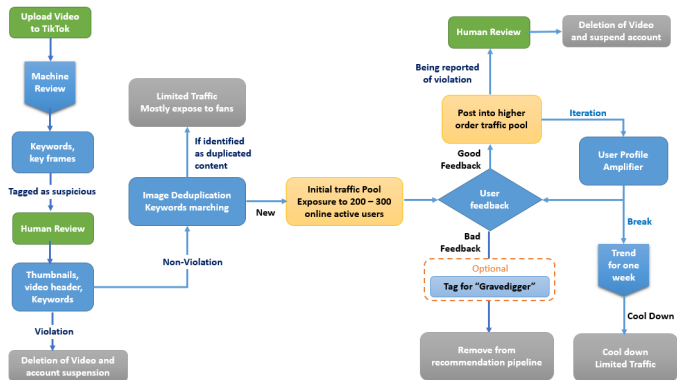
Правило 80/20

Холодный старт и длинный хвост / Spotify (2016) [Spo16]



Холодные айтемы
и пользователи
будут всегда: ду-
маем, что с ними
делать

Работа с контентом / TikTok (2020) [Wan20]



Потребности
людей нельзя упа-
ковать в удобную
метрику: вокруг
МЛ нужен пре- и
пост-процессинг

Как в действительности выглядит архитектура RS



Метрики и эксперименты

Разрабатываем инструменты для анализа метрик

Задача

Какой эффект на распределение целевой метрики окажет выбранное воздействие T ?

Фундаментальная Проблема Causal Inference

Для конкретного объекта невозможно вычислить causal effect напрямую, потому что нельзя пронаблюдать значение целевой переменной при более чем одном значении T^a

^aБез дополнительных предположений эту проблему не решить [GH07]

Фреймворк Potential Outcomes

Воздействие на i пользователя:

$$T_i = \begin{cases} 0, & \text{если показываем control} \\ 1, & \text{если показываем treatment} \end{cases}$$

Соответствующие потенциальные исходы:

$$y_i^0 \text{ и } y_i^1$$

Требуется оценить:

Average Treatment Effect

$$ATE = E[y_i^1 - y_i^0]$$

Randomized Controlled Experiment

Схема эксперимента

Все доступные пользователи независимо друг от друга случайным образом распределяются в control либо treatment с одинаковой вероятностью

Предположение 1:

Можно оценить значение некоторой характеристики для всей популяции, имея выборку из этой популяции.

Предположение 2: Stable Unit Treatment Value Assumption

Потенциальные исходы для каждого пользователя зависят только от свойств этого пользователя, но не свойств и исходов других пользователей.

Оцениваем АТЕ в RCE

$$ATE = E[y_i^1 - y_i^0] = E[y_i^1] - E[y_i^0] \sim \text{avg}_{i \in T}(y_i^1) - \text{avg}_{i \in C}(y_i^0) = \bar{y}_1 - \bar{y}_0$$

- нужно оценить две характеристики – $E[y_i^0]$ и $E[y_i^1]$, поэтому используем выборки C и T
- проще всего сделать оценку, если выборка несмещенная
- чем больше данных, тем точнее оценка

Доверительный интервал на АТЕ

Доверительный интервал (L, U) с уровнем доверия α :

$$P(L < \theta < U) = 1 - \alpha$$

Формула Уэлча:

$$\bar{y}_1 - \bar{y}_0 \pm t_{\alpha/2, r} \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}, \quad r = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_0^4}{n_0^2(n_0-1)}}$$

Где:

- n_1 и n_0 – количество пользователей в treatment и control
- s_1^2 и s_0^2 – оценки дисперсии метрики в treatment и control
- $t_{\alpha/2, r}$ – табличное значение для r степеней свободы

На практике





- Метрики распределены по-разному: нужно подбирать подходящие тесты
- Используются методы снижения дисперсии оценок (cuped, diff-in-diff)
- Собираются тысячи метрик: часто для интерпретации нужны специалисты

Если вы попали в компанию, в которой есть культура принятия решений на основе данных – сохраняйте ее всеми силами. Если нет – пропагандируйте.

Итоги

A set of small navigation icons typically found in Beamer presentations, including symbols for back, forward, search, and other slide controls.

Литература I

-  Paul Covington, Jay Adams, and Emre Sargin, *Deep neural networks for youtube recommendations*, Proceedings of the 10th ACM Conference on Recommender Systems (New York, NY, USA), RecSys '16, Association for Computing Machinery, 2016, p. 191–198.
-  Andrew Gelman and Jennifer Hill, *Data analysis using regression and multilevel/hierarchical models*, vol. Analytical methods for social research, Cambridge University Press, New York, 2007.
-  Kris Jack, Ed Ingold, and Maya Hristakeva, *Mendeley suggest architecture*, Oct 2016.
-  Xavier Amatriain Netflix and Justin Basilico Netflix, *System architectures for personalization and recommendation*, Mar 2013.

