

Метрики и базовые подходы

Николай Анохин

6 октября 2021 г.

Сбор данных
oooooooooo

Релевантность
ooooooo

Покрытие
oo

Разнообразие
oo

Удачность
ooo

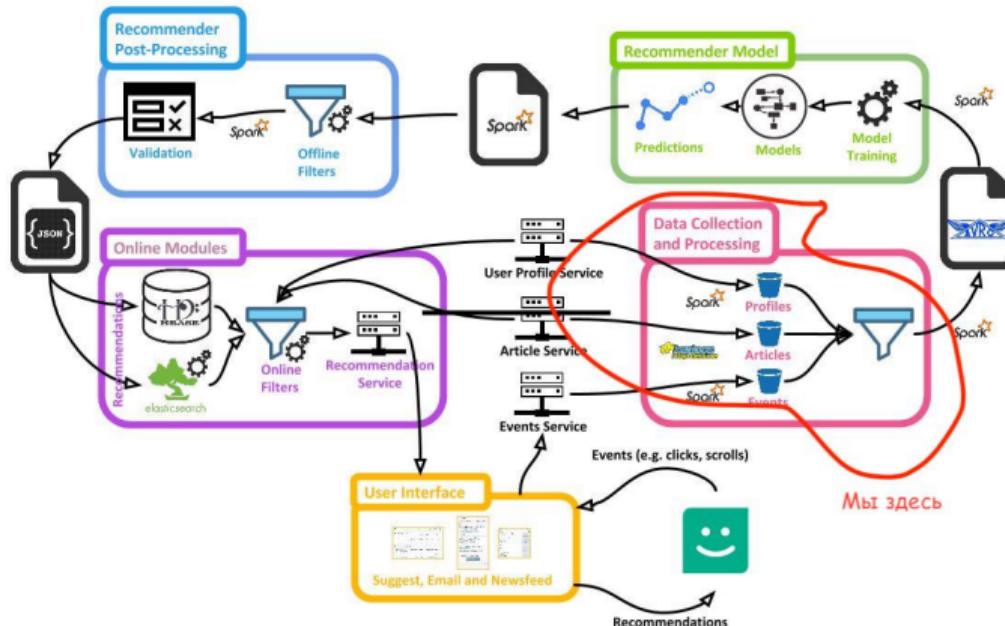
Бейзлайны
o

Итоги
oo

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	✓	

Контекст



Сбор данных
○●○○○○○○

Релевантность
○○○○○○○○

Покрытие
○○

Разнообразие
○○

Удачность
○○○

Бейзлайны
○

Итоги
○○

Научный метод



Чем быстрее делаем
оборот, тем быстрее
улучшаем сервис

A/B эксперимент [RRSK10]

+ Надежная оценка эффекта на любую метрику

- Риск необратимо расстроить пользователей
- Риск финансовых потерь
- Дорого заводить
- Ограниченный трафик

Опрос пользователей

- + Полный контроль над экспериментом
- + Оценка эффекта на любую метрику
- + Собрать фидбэк напрямую

- Дорогой сбор данных
- Смещение аудитории
- Нечестный фидбэк

Оффлайн эксперимент

- + Проверка большого числа гипотез
- + Нельзя сломать прод

- Нужно подбирать метрики
- Смещение выборки
- Результат не обязан обосноваться

При оффлайн оценке нужно стремиться к тому, чтобы данные были максимально похожи на реальность

Техники выбора тестовых данных

- Семплировать случайные пары user-item
- Семплировать случайные item у каждого пользователя
- Семплировать тестовых пользователей
- Тестовые данные после обучающих по времени

Бизнесовая метрика

напрямую интересует бизнес

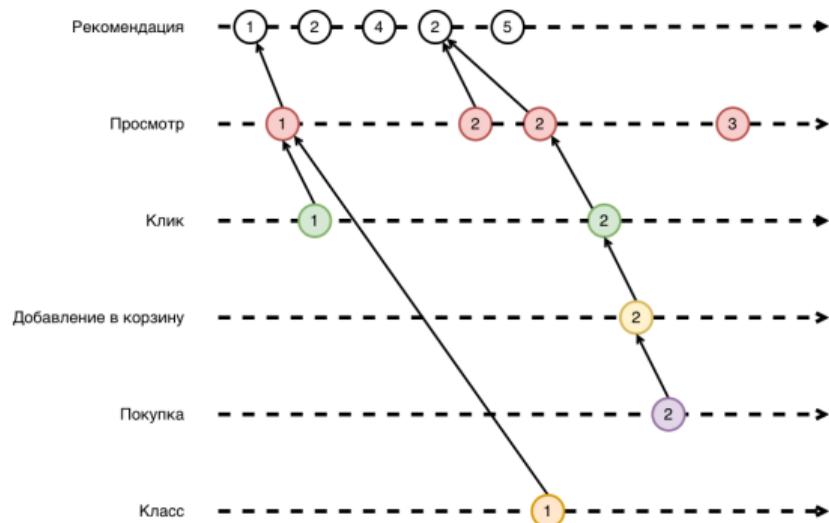
- сложно оптимизировать
- сложно понять, как компоненты системы влияют на метрику
- сложно мерить офлайн

Техническая метрика

отражает один аспект системы

- можно оптимизировать
- можно померить офлайн
- не интересует бизнес :(

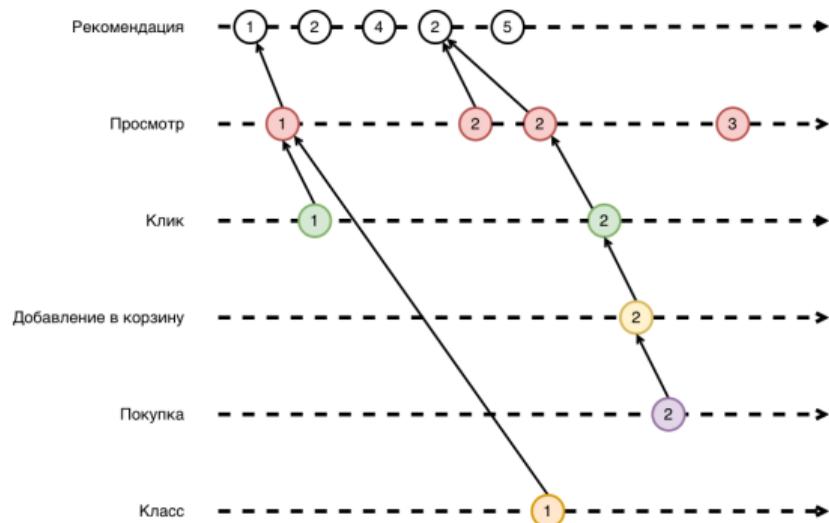
Какой бывает фидбэк



Техническая метрика

- Явный/explicit

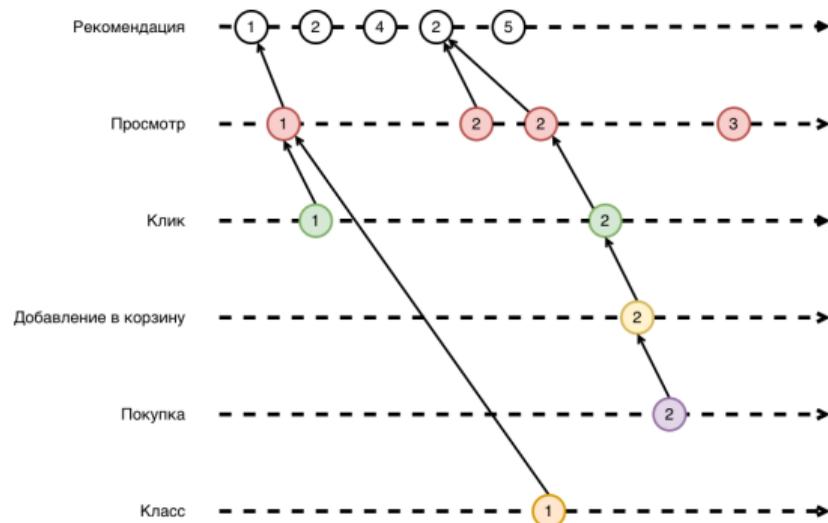
Какой бывает фидбэк



Техническая метрика

- Явный/explicit
- Неявный/implicit

Какой бывает фидбэк



Техническая метрика

- Явный/explicit
- Неявный/implicit
- Отложенный/delayed

Сбор данных
oooooooo●

Релевантность
ooooooo

Покрытие
oo

Разнообразие
oo

Удачность
ooo

Бейзлайны
o

Итоги
oo

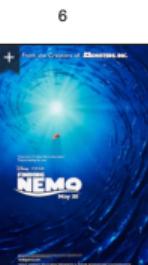
Pinkamena Diane Pie



A comic relief character [...] appears to be the naive party animal of the group, she also displays admirable skill in science and engineering.

Релевантность

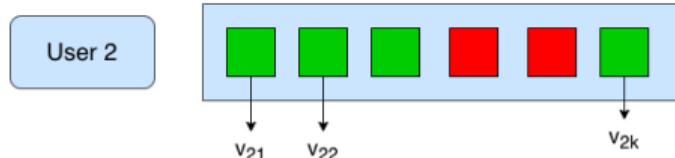
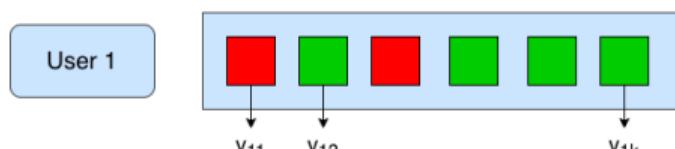
Насколько рекомендации соответствуют вкусам пользователя?



Метрики точности

 Non-relevant item

 Relevant item



RMSE, MAE, accuracy, precision, recall, auc, ...

Сбор данных
oooooooo

Релевантность
oo●ooooo

Покрытие
oo

Разнообразие
oo

Удачность
ooo

Бейзлайны
o

Итоги
oo

Метрики ранжирования



Non-relevant item



Relevant item

User 1



v_1

User 2



v_2

Сбор данных
ooooooooo

Релевантность
ooo●ooo

Покрытие
oo

Разнообразие
oo

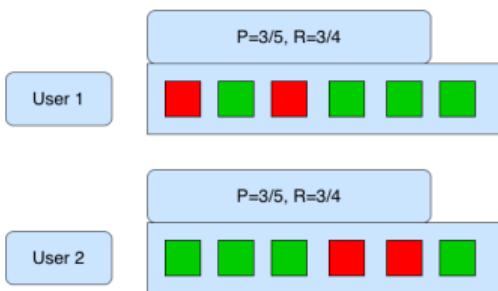
Удачность
ooo

Бейзлайны
o

Итоги
oo

Precision@k, Recall@k

- Non-relevant item
- Relevant item

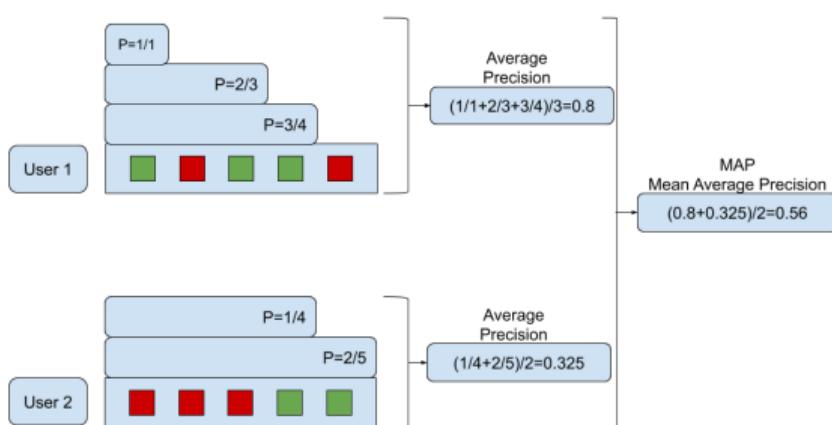


- + Легко интерпретировать
- + Легко реализовать

- Нечувствительны к порядку внутри k
- Не дают общей картины для любого k

Mean Average Precision [Tai19]

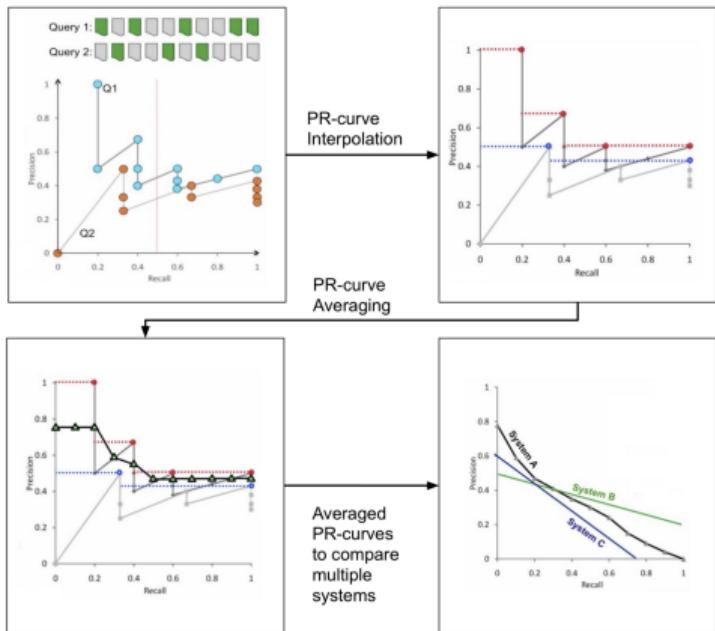
Relevant Item
Non-Relevant Item



- + Дают общую картину качества
- + Больше внимания айтемам в голове списка

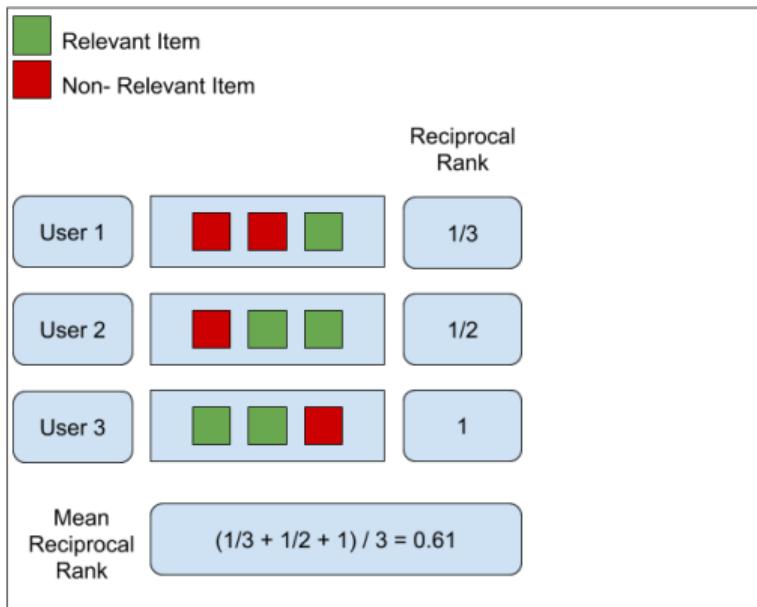
- Подходит только для бинарного фидбэка

Area Under Precision-Recall curve



Визуальное представление
MAP

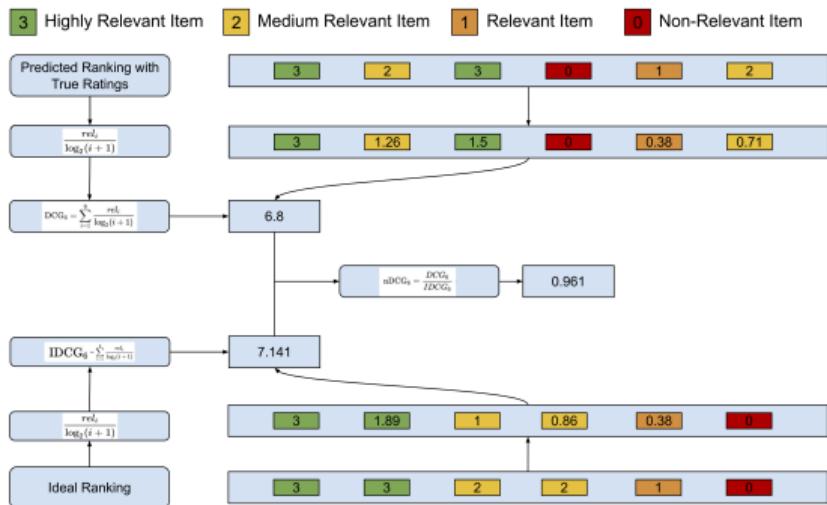
MRR



- + Легко интерпретировать
- + Легко реализовать
- + Удобна для задач, где имеет значение первый результат

- Учитывает только первый результат
- Быстро убывает

[N]DCG



- + Учитывает не только бинарный фидбэк
- + Хорошо учитывает позицию

- Сложно интерпретировать

Сбор данных
oooooooo

Релевантность
ooooooo

Покрытие
●○

Разнообразие
oo

Удачность
ooo

Бейзлайны
o

Итоги
oo

Item space coverage

Какую долю из всех возможных айтемов умеет рекомендовать сервис?

$$cov = \frac{|I_p|}{|I|}$$

$$gini = \frac{1}{|I|-1} \sum_{j=1}^{|I|} (2j - |I| - 1)p(I_j)$$

$p^1(I_j)$ – частота, с которой пользователи выбирают айтем I_j
 $p^2(I_j)$ – частота, с которой рекомендер показывает айтем I_j

Сбор данных
oooooooo

Релевантность
ooooooo

Покрытие
oo

Разнообразие
oo

Удачность
ooo

Бейзлайны
o

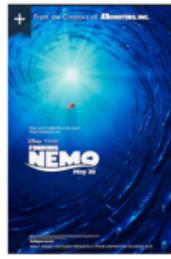
Итоги
oo

User space coverage

Доля пользователей, которые могут получить рекомендации

Разнообразие [KP17]

[diversity] Насколько разнообразные айтемы в списке рекомендаций пользователя?



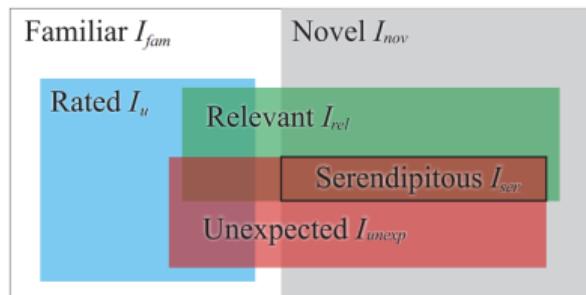
$$div(u) = \frac{\sum_{i=1}^n \sum_{j=1}^n (1 - similarity(i,j))}{n/2(n-1)}$$

With 1% precision loss, percentage of rec. long-tail items increases from 16 to 32, with 5% loss perc. increases to 58.

Метрика сильно зависит от того, как определить сходство

Удачность

The term **serendipity** has been recognized as one of the most untranslatable words. The first known use of the term was found in a letter by Horace Walpole to Sir Horace Mann on January 28, 1754. The author described his discovery by referencing a Persian fairy tale, “The Three Princes of Serendip”. The story described a journey taken by three princes of the country Serendip to explore the world. In the letter, Horace Walpole indicated that the princes were “always making discoveries, by accidents and sagacity, of things which they were not in quest of”. [KWV16]



Сбор данных
oooooooooo

Релевантность
ooooooooo

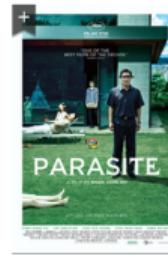
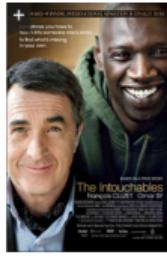
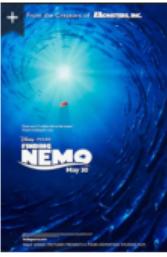
Покрытие
oo

Разнообразие
oo

Удачность
ooo

Бейзлайны
o

Итоги
oo



Новизна

[novelty] Насколько айтем неизвестен пользователю?

Идея 1: Насколько айтемы близки к айтэмам из истории пользователя?

$$nov^1(u, i) = \min_{j \in I_u} dist(j, i)$$

Идея 2: Насколько айтэмы близки к популярным?

$$nov^2(u, i) = 1 - \frac{|U_i|}{|U|}$$

Сбор данных
ooooooooo

Релевантность
ooooooo

Покрытие
oo

Разнообразие
oo

Удачность
ooo

Бейзлайны
o

Итоги
oo

Неожиданность

[unexpectedness] Насколько пользователь ожидает увидеть в рекомендациях айтем?

$$nPMI(i, j) = -\log \frac{p(i, j)}{p(i)p(j)} / \log p(i, j)$$

$$unexp(u, i) = \max_{j \in I_u} (-nPMI(i, j))$$

Простые бейзлайны

- позволяют определить нижнюю границу качества системы
- позволяют быстро стартануть

> Живительный рандом

> TopPopular

> Эвристики

Итоги

При выборе подхода к проверке гипотез, нужно иметь в виду компромисс надежности и скорости

Технические метрики отражают разные аспекты рекомендаций: релевантность, разнообразие, удачность

Не обмазываемся сложными алгоритмами, пока не заведем простые бейзлайны

Литература I

-  Matevz Kunaver and Tomaz Pozrl, *Diversity in recommender systems - a survey*, Knowl. Based Syst. **123** (2017), 154–162.
-  Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen, *A survey of serendipity in recommender systems*, Knowledge-Based Systems **111** (2016).
-  Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, *Recommender systems handbook*, 1st ed., Springer-Verlag, Berlin, Heidelberg, 2010.
-  Moussa Taifi, *Mrr vs map vs ndcg: Rank-aware evaluation metrics and when to use them*, Nov 2019.