

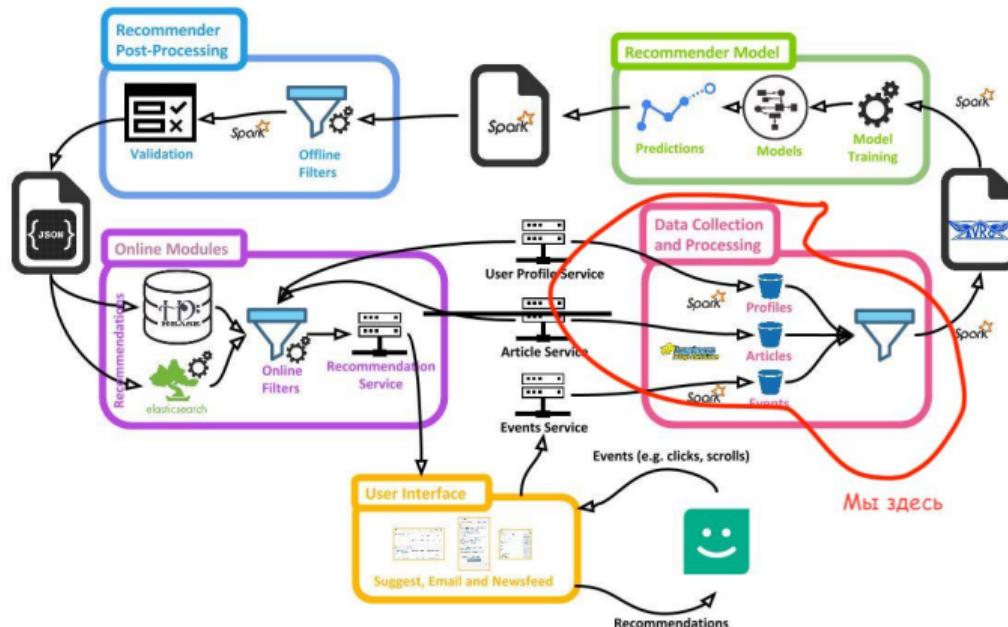
Метрики и базовые подходы

Николай Анохин

22 мая 2022 г.



Контекст



Оценка успешности идей
●ooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
oooo

Удачность
oooooo

Бейзлайны
oo

Итоги
ooo

Оценка успешности идей

Оценка успешности идей
о●oooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
oooo

Удачность
oooooo

Бейзлайны
oo

Итоги
ooo

Научный метод



Чем быстрее проходим все этапы, тем быстрее улучшаем сервис

A/B эксперимент [RRSK10]

Плюсы

- Надежная оценка эффекта на любую метрику

Минусы

- Риск необратимо расстроить пользователей
- Риск финансовых потерь
- Дорого заводить
- Долго ждать результат
- Метрик не всегда достаточно

Опрос пользователей

Плюсы

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

Минусы

- Дорогой сбор данных
- Смещение аудитории
- Нечестный фидбэк

Оффлайн эксперимент

Плюсы

- Большая скорость проверки гипотез
- Нельзя сломать прод

Минусы

- Не все метрики доступны офлайн
- Смещение выборки
- Результат не обязан обобщаться

Оценка успешности идей
ооооо●

Оффлайн эксперимент
оооо

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
оооо

Удачность
оооооо

Бейзлайны
oo

Итоги
ooo

Миссия социальной сети

Helping users feel closer to their friends and family.

Q: Какую метрику вы бы предложили измерять в A/B?

Оценка успешности идей
oooooooo

Оффлайн эксперимент
●ooo

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
oooo

Удачность
oooooo

Бейзлайны
oo

Итоги
ooo

Оффлайн эксперимент

Бизнесовая метрика

напрямую интересует бизнес

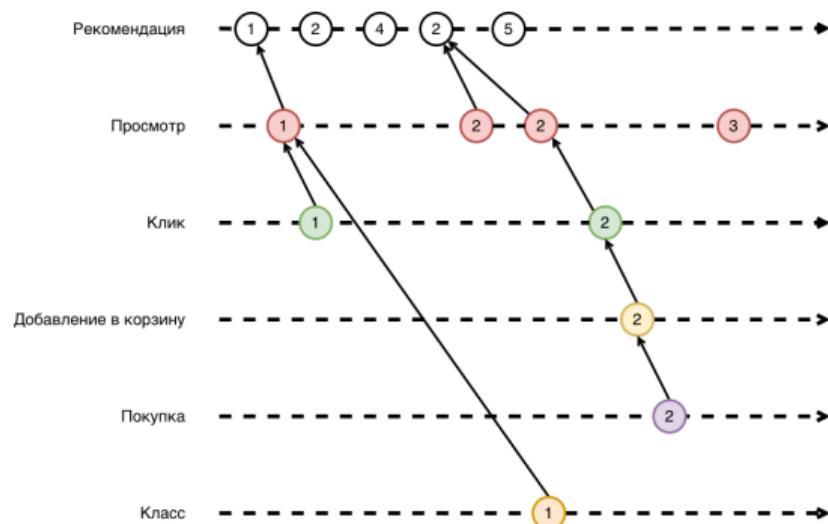
- сложно оптимизировать
- сложно понять, как компоненты системы влияют на метрику
- сложно мерить онлайн

Техническая метрика

отражает один аспект системы

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

Какой бывает фидбэк



Техническая метрика

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

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

Техники выбора данных для оффлайн оценки

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

Оценка успешности идей
оооооо

Оффлайн эксперимент
оооо

Релевантность
●oooooooooooo

Покрытие
ooo

Разнообразие
ооооо

Удачность
оооооо

Бейзлайны
oo

Итоги
ooo

Релевантность

Оценка успешности идей
оооооо

Оффлайн эксперимент
оооо

Релевантность
о●oooooooooooo

Покрытие
ooo

Разнообразие
ооооо

Удачность
оооооо

Бейзлайны
oo

Итоги
ooo

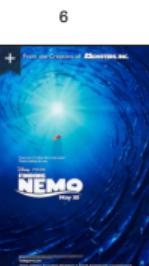
Pinkamena Diane Pie



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

Релевантность

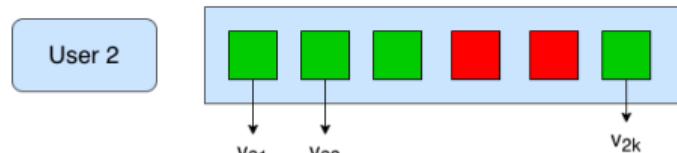
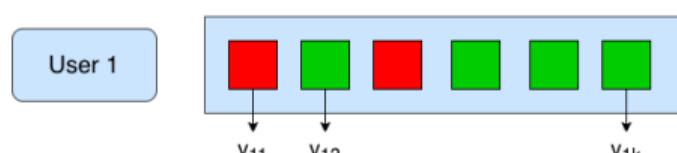
Насколько рекомендации соответствуют вкусам пользователя?



Метрики точности

 Non-relevant item

 Relevant item



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

Метрики ранжирования



Non-relevant item



Relevant item

User 1

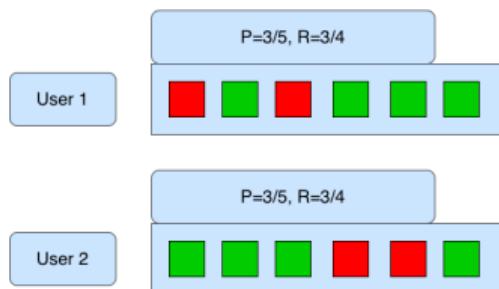


User 2



Precision@k, Recall@k

- Non-relevant item
- Relevant item

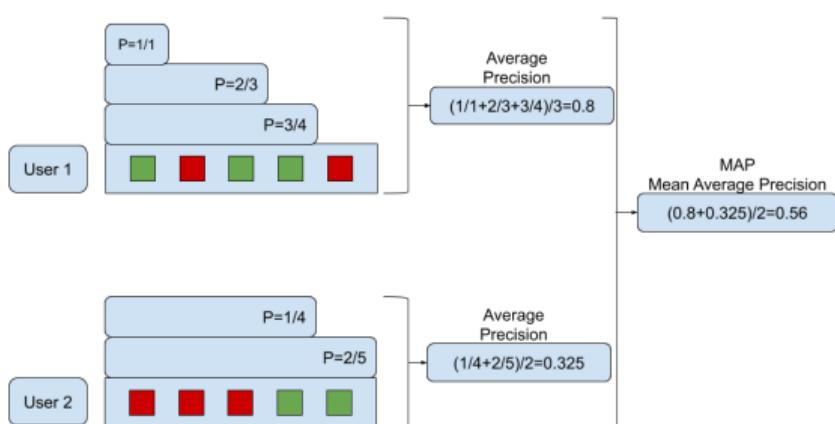


- Легко интерпретировать
- Легко реализовать

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

Mean Average Precision [Tai19]

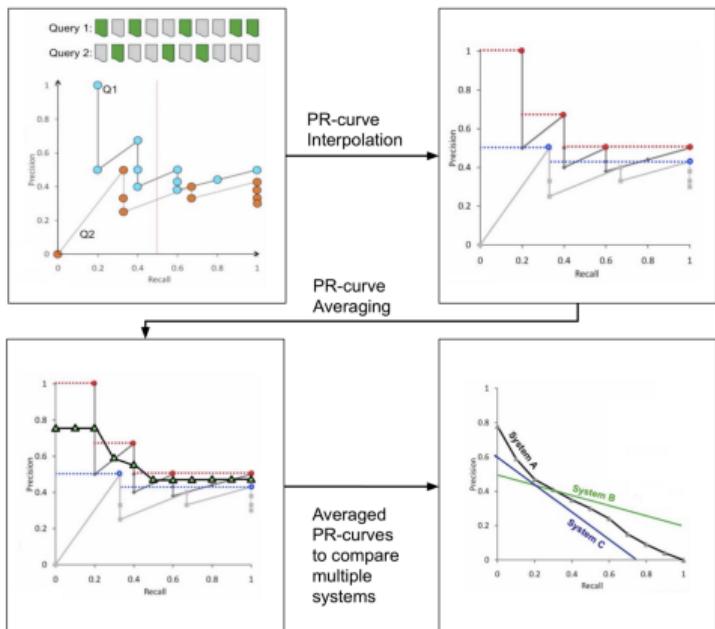
■ Relevant Item
■ Non-Relevant Item



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

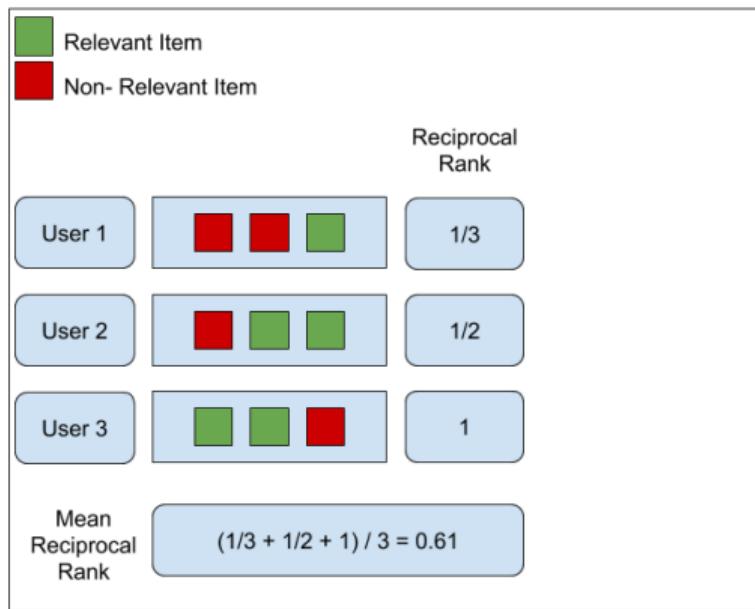
- Подходит только для бинарного фидбэка

Area Under Precision-Recall curve



Визуальное представление
MAP

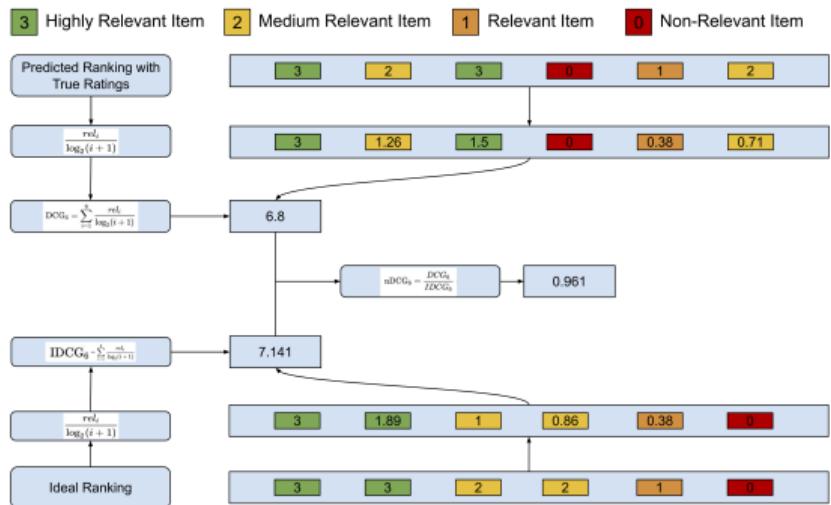
MRR



- 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

Оценка успешности идей
оооооо

Оффлайн эксперимент
оооо

Релевантность
oooooooooooo●

Покрытие
ooo

Разнообразие
оооо

Удачность
оооооо

Бейзлайны
oo

Итоги
ooo

Гайд по выбору метрик Николая Анохина

1. Находим метрики релевантности, которые подходят к задаче
2. Выбираем в качестве основной самую интерпретируемую
3. Усложняем метрику, если оказалось, что она не отражает реальность

Оценка успешности идей
oooooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

Покрытие
●○○

Разнообразие
oooo

Удачность
oooooo

Бейзлайны
oo

Итоги
ooo

Покрытие

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

Айтемы отсортированы по возрастанию $p(I_j)$

Оценка успешности идей
оооооо

Оффлайн эксперимент
оооо

Релевантность
оооооооооооооо

Покрытие
оо●

Разнообразие
оооо

Удачность
оооооо

Бейзлайны
оо

Итоги
ооо

User space coverage

Доля пользователей, которые могут получить рекомендации

Оценка успешности идей
oooooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
●ooo

Удачность
oooooo

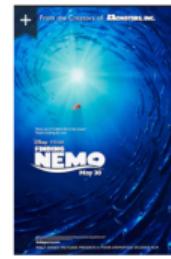
Бейзлайны
oo

Итоги
ooo

Разнообразие

Разнообразие [KP17]

[diversity] Насколько разнообразные айтемы в списке рекомендаций пользователя?



Оценка успешности идей
oooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
ooo●o

Удачность
oooooo

Бейзлайны
oo

Итоги
ooo

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

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

Оценка успешности идей
oooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
ooo●

Удачность
oooooo

Бейзлайны
oo

Итоги
ooo

Maximal Marginal Relevance [CG98]

$$MMR = \max_j \left[\lambda \text{similarity}(j, U) + (1 - \lambda) \max_{k < j} \text{similarity}(k, j) \right]$$

Оценка успешности идей
oooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
oooo

Удачность
●ooooo

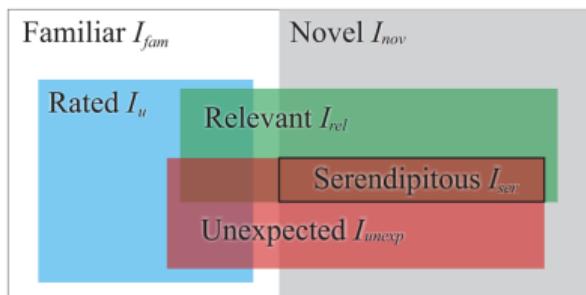
Бейзлайны
oo

Итоги
ooo

Удачность

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]



Оценка успешности идей
ooooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

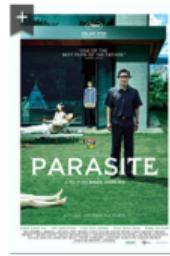
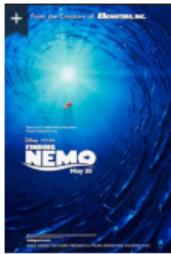
Покрытие
ooo

Разнообразие
oooo

Удачность
oo●ooo

Бейзлайны
oo

Итоги
ooo



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?

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

Неожиданность

[unexpectedness] Насколько пользователь ожидает увидеть в рекомендациях айтем?

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

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

Оценка успешности идей
ооооооОффлайн эксперимент
ооооРелевантность
ооооооооооооПокрытие
оооРазнообразие
ооооУдачность
ооооо●Бейзлайны
ооИтоги
ооо

Цель	rel	cov	div	ser	poll
Бизнесу					
Увеличить продажи	✓			✓	
Продвигать более разнообразные айтемы		✓	✓	✓	
Улучшить пользовательский опыт	✓		✓	✓	✓
Добиться большей лояльности					✓
Лучше понимать пользователей					✓
Пользователям					
Найти лучший товар	✓			✓	✓
Найти все подходящие товары	✓	✓			✓
Найти последовательность или набор товаров			✓		✓
Залипнуть			✓	✓	✓
Найти рекомендер, которому можно доверять					✓
Помочь другим сделать выбор					✓

Оценка успешности идей
оооооо

Оффлайн эксперимент
оооо

Релевантность
оооооооооооо

Покрытие
ооо

Разнообразие
оооо

Удачность
оооооо

Бейзлайны
●○

Итоги
ооо

Бейзлайны

Простые бейзлайны

- позволяют определить нижнюю границу качества системы
- позволяют быстро стартануть

- Живительный рандом
- TopPopular
- Эвристики
- Редакторская подборка

Оценка успешности идей
oooooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooooooo

Покрытие
ooo

Разнообразие
oooo

Удачность
oooooo

Бейзлайны
oo

Итоги
●oo

Итоги

Итоги

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

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

Don't be a hero: не связываемся со сложными алгоритмами, пока не заведем простые бейзлайны

Литература I

-  Jaime G. Carbonell and Jade Goldstein, *The use of MMR, diversity-based reranking for reordering documents and producing summaries*, Research and Development in Information Retrieval, 1998, pp. 335–336.
-  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.