

# Метрики и базовые подходы

Николай Анохин

7 октября 2021 г.

Сбор данных  
oooooooooo

Релевантность  
oooooooooo

Покрытие  
oo

Разнообразие  
oo

Удачность  
oooo

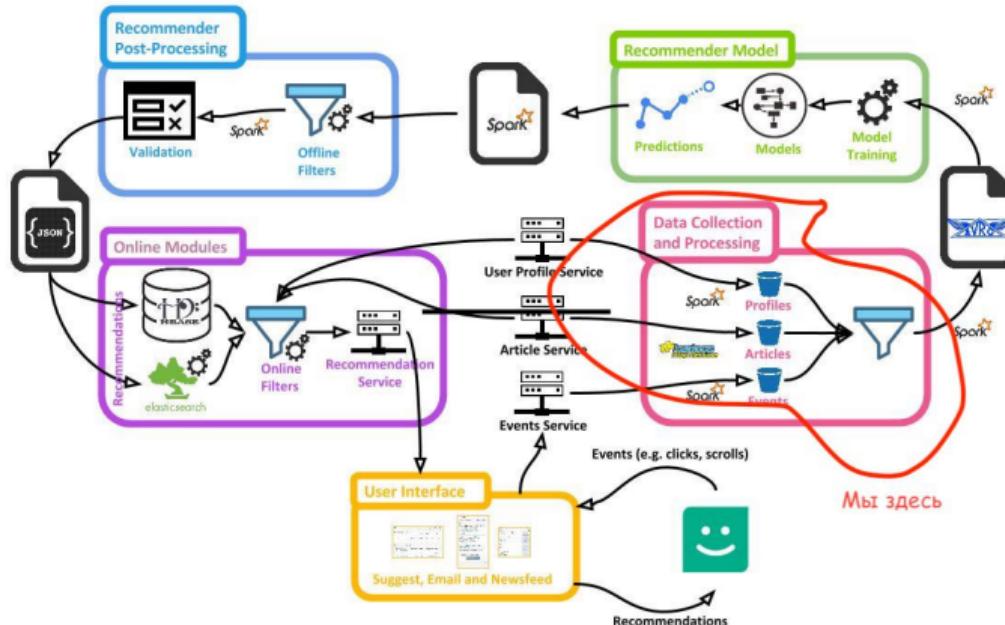
Бейзлайны  
о

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

Сбор данных  
оо●oooooo

Релевантность  
oooooooo

Покрытие  
оо

Разнообразие  
оо

Удачность  
oooo

Бейзлайны  
о

Итоги  
оо

## A/B эксперимент [RRSK10]

### Плюсы

- Надежная оценка эффекта на любую метрику

### Минусы

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

Сбор данных  
ооо●ооооо

Релевантность  
оооооооооо

Покрытие  
оо

Разнообразие  
оо

Удачность  
оооо

Бейзлайны  
о

Итоги  
оо

## 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

- Проверка большого числа гипотез
- Нельзя сломать прод

### Минусы

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

Сбор данных  
ооооо●ooo

Релевантность  
оооооооооо

Покрытие  
оо

Разнообразие  
оо

Удачность  
оооо

Бейзлайны  
о

Итоги  
оо

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

## Техники выбора тестовых данных

- Семплировать случайные пары user-item

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

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

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

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
- Написать симулятор системы

## Бизнесовая метрика

напрямую интересует бизнес

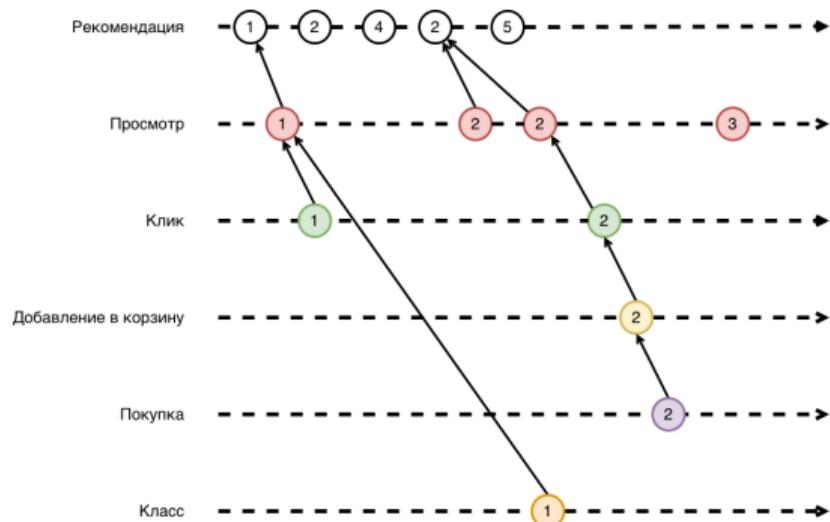
- сложно оптимизировать
- сложно понять, как компоненты системы влияют на метрику
- сложно мерить офлайн

## Техническая метрика

отражает один аспект системы

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

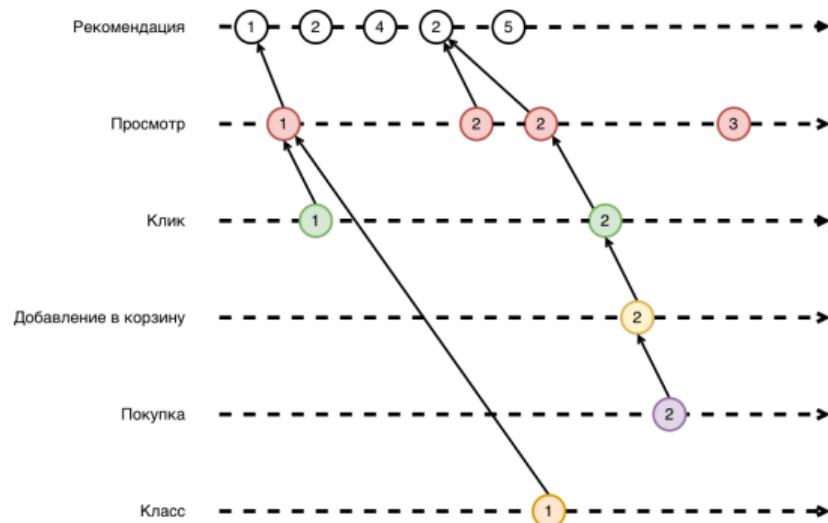
## Какой бывает фидбэк



### Техническая метрика

- Явный/explicit

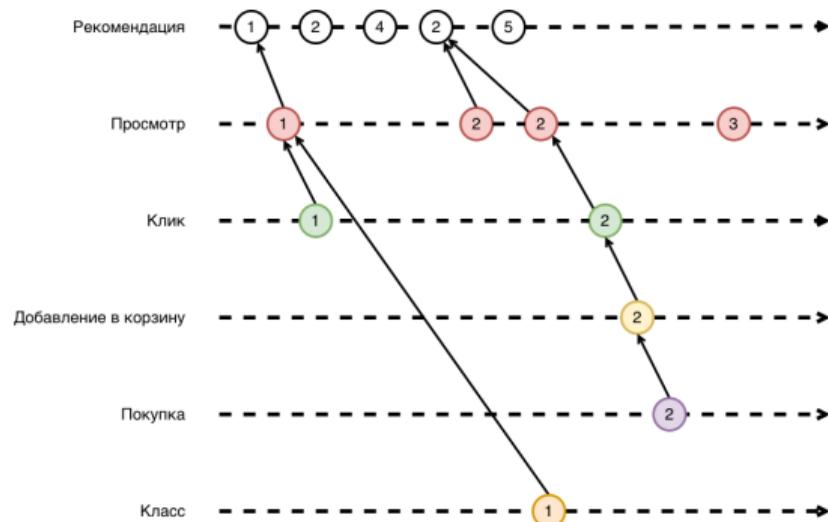
## Какой бывает фидбэк



### Техническая метрика

- Явный/explicit
- Неявный/implicit

## Какой бывает фидбэк



### Техническая метрика

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

Сбор данных  
oooooooo●

Релевантность  
oooooooo

Покрытие  
oo

Разнообразие  
oo

Удачность  
oooo

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.

Сбор данных  
ooooooooo

Релевантность  
●ooooooooo

Покрытие  
oo

Разнообразие  
oo

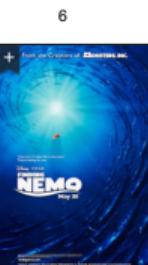
Удачность  
oooo

Бейзлайны  
o

Итоги  
oo

## Релевантность

Насколько рекомендации соответствуют вкусам пользователя?



Сбор данных  
oooooooo

Релевантность  
●●oooooooo

Покрытие  
oo

Разнообразие  
oo

Удачность  
oooo

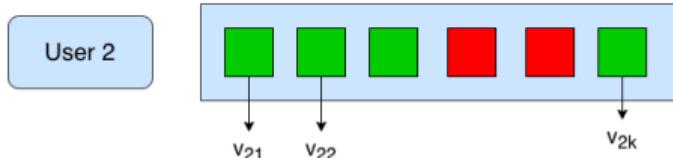
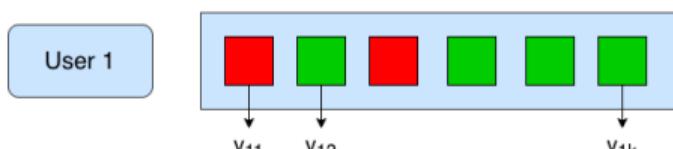
Бейзлайны  
o

Итоги  
oo

## Метрики точности

 Non-relevant item

 Relevant item



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

Сбор данных  
oooooooo

Релевантность  
oo●ooooo

Покрытие  
oo

Разнообразие  
oo

Удачность  
oooo

Бейзлайны  
o

Итоги  
oo

## Метрики ранжирования

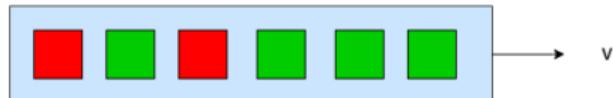


Non-relevant item

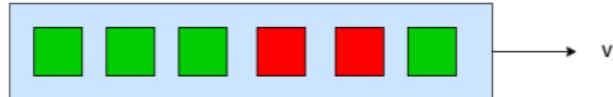


Relevant item

User 1



User 2



Сбор данных  
ooooooooo

Релевантность  
ooo●ooooo

Покрытие  
oo

Разнообразие  
oo

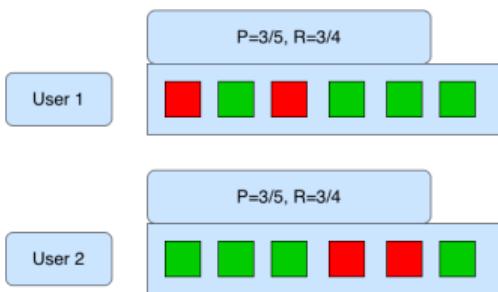
Удачность  
oooo

Бейзлайны  
o

Итоги  
oo

## Precision@k, Recall@k

- Non-relevant item
- Relevant item

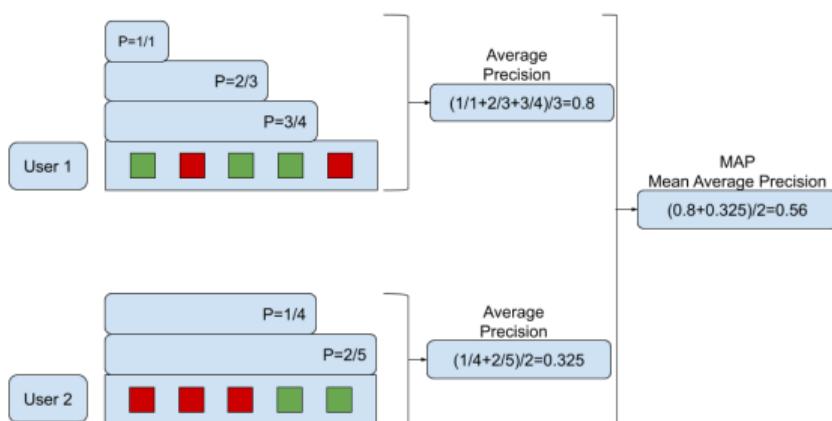


- Легко интерпретировать
- Легко реализовать

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

## Mean Average Precision [Tai19]

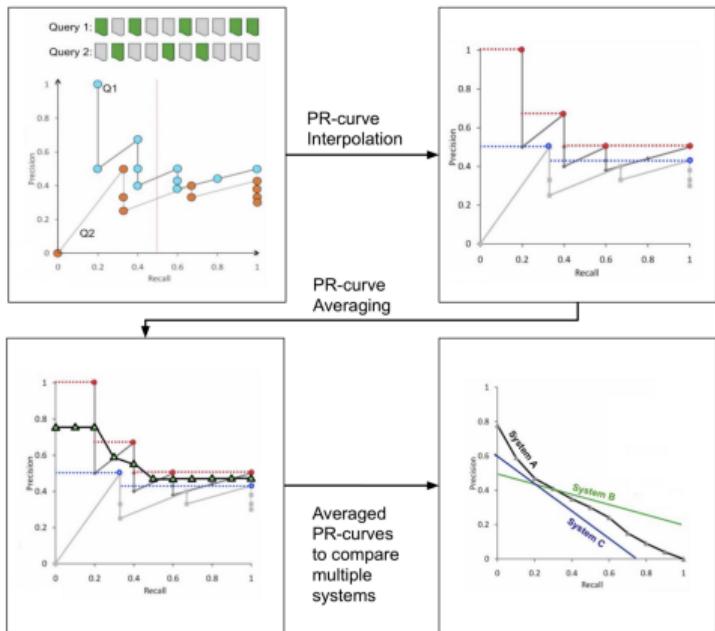
Relevant Item  
Non-Relevant Item



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

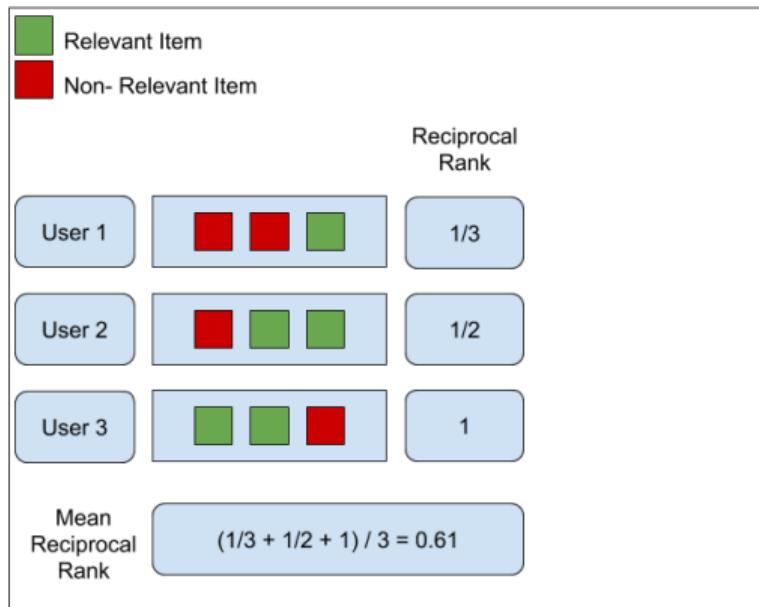
- Подходит только для бинарного фидбэка

## Area Under Precision-Recall curve



Визуальное представление  
MAP

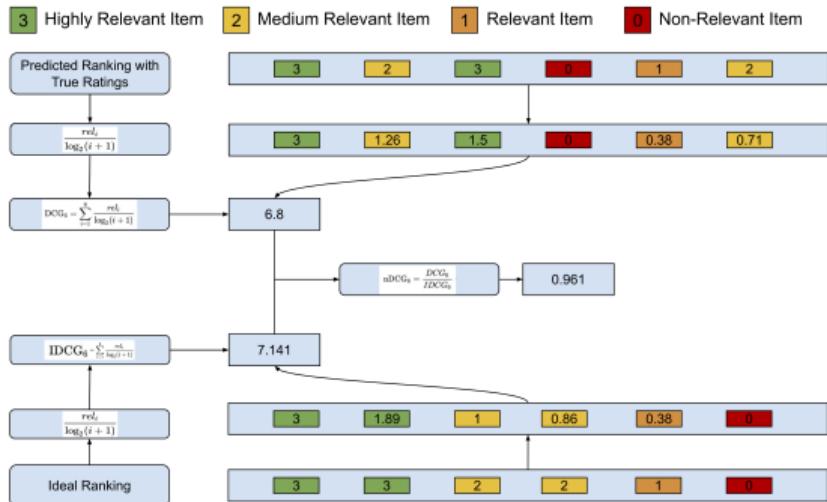
## MRR



- 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  
ooooooooo

Релевантность  
oooooooo●

Покрытие  
oo

Разнообразие  
oo

Удачность  
oooo

Бейзлайны  
о

Итоги  
oo

## 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

Сбор данных  
oooooooooo

Релевантность  
oooooooooo

Покрытие  
●○

Разнообразие  
oo

Удачность  
oooo

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  
oooooooooo

Релевантность  
oooooooooo

Покрытие  
oo●

Разнообразие  
oo

Удачность  
oooo

Бейзлайны  
o

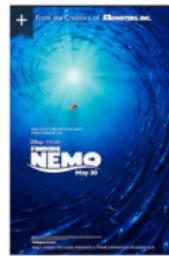
Итоги  
oo

## User space coverage

Доля пользователей, которые могут получить рекомендации

## Разнообразие [KP17]

[diversity] Насколько разнообразные айтемы в списке рекомендаций пользователя?



Сбор данных  
oooooooooo

Релевантность  
oooooooooo

Покрытие  
oo

Разнообразие  
o●

Удачность  
oooo

Бейзлайны  
o

Итоги  
oo

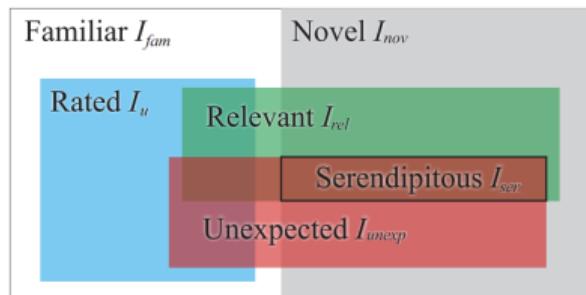
$$div(u) = \frac{\sum_{i=1}^n \sum_{j=1}^n (1 - similarity(i,j))}{n/2(n-1)}$$

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  
oooooooooo

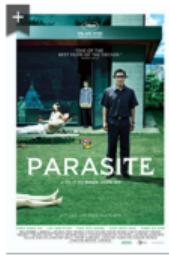
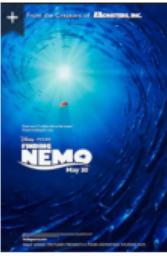
Покрытие  
oo

Разнообразие  
oo

Удачность  
o●○○

Бейзлайны  
o

Итоги  
oo



Сбор данных  
ooooooooo

Релевантность  
ooooooooo

Покрытие  
oo

Разнообразие  
oo

Удачность  
oo●o

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?

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

## 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.