



The illustration shows a supermarket shelf with two signs. The top sign reads 'CUSTOMERS WHO BOUGHT THIS ITEM:' and has a bunch of bananas below it. The bottom sign reads 'ALSO BOUGHT:' and has a muffin, a loaf of bread, and a carton of milk below it. A person is standing to the right of the shelf, holding a green shopping basket.

Рекомендательные системы (Recommender Systems): классические методы

Александр Дьяконов

19 апреля 2022 года

Системы рекомендаций: описание и назначение



КУРТ ВОННЕГУТ
МАТЕРЬ ТЬМА
Был, конечно же, мир. Ласково, объединяющее логично.

★★★★★ 5.0 из 5

Матерь Тьма ✓ В наличии

Скидка 20%

~~188 р.~~ **150 р.**

Добавить в корзину

Автор: [Воннегут К.](#)
Серия: [Эксклюзивная классика](#)
Жанр: [Классическая проза](#)
Издательство: [Издательство «АСТ»](#)
ISBN: 978-5-17-099474-8
Артикул: p1600862
Возрастное ограничение: 16+

Похожие товары



ДЖОРДЖ ОРУЭЛЛ
1984



КУРТ ВОННЕГУТ
ЗЕЛЕНАЯ ДОЛЖА ЧЕЛОВЕКА

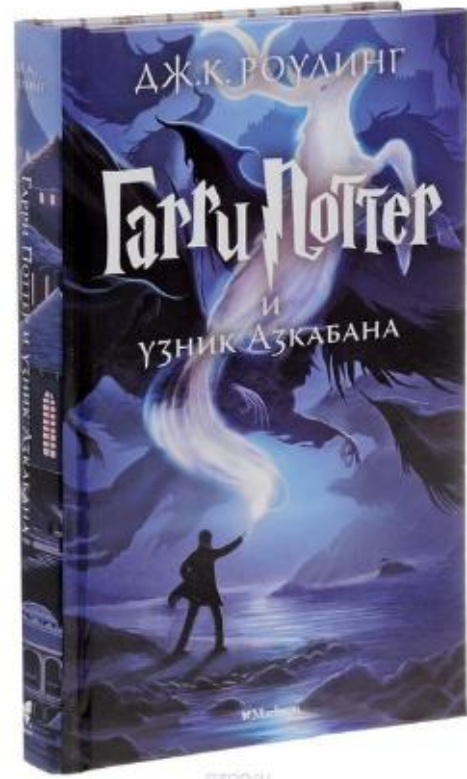


ДЖЕРОМ К. ДЖЕРОМ
ТРОИ В ЛОДКЕ. НЕ СЧИТАЙ СОБАКИ



УИЛЬЯМ ГОЛДИНГ
ПОВЕЛИТЕЛЬ МУХ

Системы рекомендаций: описание и назначение



Бestseller

Гарри Поттер и узник Азкабана

★★★★★ 14 отзывов В избранное Поделиться Код товара: 31275832

Твердый переплет (2)
от 414 ₽

Бумажное издание (2)
от 2 469 ₽

Нет в продаже
9 изданий

Ориг.название: Harry Potter and the Prisoner of Azkaban

Автор: Джован Кэтлин Роулинг

Формат издания: 130x200 мм (средний формат)

Количество страниц: 528

Год выпуска: 2015

[Показать все характеристики](#)

414 ₽

✓ В наличии

Курьер доставит завтра


Добавить в корзину

Продавец: OZON.ru

О книге


Книга, покори́вшая мир, эталон литературы для читателей всех возрастов, синоним успеха. Книга, сделавшая Дж.К.Роулинг самым читаемым писателем современности. [Читать далее](#)

Рекомендуем также




469 ₽

Гарри Поттер и Кубок Огня
Дж. К. Роулинг




1 160 ₽

Гарри Поттер и философский камень
Дж.К. Роулинг




414 ₽

Гарри Поттер и Тайная комната
Дж. К. Роулинг




509 ₽

Гарри Поттер и Орден Феникса
Дж. К. Роулинг



489 ₽


Гарри Поттер и Дары Смерти
Дж. К. Роулинг



414 ₽


Гарри Поттер и Принц-полукровка
Дж. К. Роулинг

Системы рекомендаций: описание и назначение




2 990 руб

Nike / Лонгслив спортивный W NK MILER TOP LS METALLIC




6 990 руб

Nike / Кроссовки Women's Nike Air Force 1 '07 Shoe



~~3 790 руб~~ 3 220 руб

Nike / Свитшот W NK TOP VERSA CREW



1 699 руб

Mango / Очки солнцезащитные - NAOMI

Системы рекомендаций: описание и назначение

Хоббит: Нежданное путешествие

The Hobbit: An Unexpected Journey

год

2012

В главных ролях:

Мартин Фриман

Иэн МакКеллен

Ричард Армитаж

Джеймс Несбитт

страна

США, Новая Зеландия

слоган

«From the smallest beginnings come the greatest legends»

режиссер

Питер Джексон

Рейтинг фильма

★ ★ ★ ★ ★ ★ ★ ★ ★ ★

1 2 3 4 5 6 7 8 9 10

8.062 259 806

IMDb: 7.90 (696 099)

ожидание: 93% (82 839)

Топ250: 104

об оценках и Топ-250

Рейтинг кинокритиков

в мире

185 + 102 = 287

★ 6.5

64%

в России

27 + 3 = 30

90%

о рейтинге критиков

Что смотрят?

Data Science in 30 Minutes

FREE MONTHLY WEBINAR SERIES

Michael Li

Founder & CEO

The Data Incubator

Exploring the Frontiers of Machine Learning

December 19th, 2018

5:30pm ET / 2:30pm PT

Zoubin Ghahramani

Chief Scientist

Uber

EVENTBRITE.COM

👍

❤️

😂

😮

😞

😡

Chief Scientist Explores AI

Нравится

Комментарий

Поделиться

Написать комментарий...

😊

📷

GIF

🗨️

Можно давать обратную связь

19 апреля 2022

«Машинное обучение»

4 слайд из 61

Рекомендательная система: цели

**информирует пользователя о товаре / услуге,
которые ему могут быть интересны**

**не обязательно зарабатываем на продаже товара / услуги
важна лояльность пользователя (удобство сервиса)**

Использование рекомендаций

- **предложение пользователю (товары / услуги)**
- **ранжирование товаров услуг (ex: карточек на сайте)**
 - **предложить дополнительную покупку (cross-sell)**
 - **информировать о новом товаре / услуге (up-sell)**

Разница между информационным поиском и рекомендательными системами

IR

«Я знаю, что я ищу»

RecSys

«Я не уверен, что мне надо»

Виды рекомендаций

по контенту Content-based	Рекомендация похожих по описанию товаров: текст, изображения, видео, ...
коллаборативная фильтрация Collaborative Filtering	Рекомендация по статистике покупок explicit feedback предыдущие рейтинги implicit feedback история просмотров Проблема холодного старта: новый товар, новый пользователь
гибридная Hybrid	
context-aware	
knowledge-based	

Данные для рекомендаций

Описание пользователя

+ лог пользователя (поиск, ожидания и т.п.)

Описание товара

Взаимодействие (пользователь, товар)

Взаимодействие (пользователь, пользователь)

Взаимодействия (товар, товар)

Объекты рекомендаций (что рекомендуют)

заменители (alternative)

сопутствующие товары (cross sell)

бандлы

аксессуары (up sell)

популярные товары (best sellers)

персональные / неперсональные

оффлайн / онлайн

демографические (demographic)

Как рекомендуют / цели бизнеса

- **max вероятность покупки**
увеличить удовлетворение пользователя (satisfaction, fidelity)
Понять, что нужно людям
- **max матожидание прибыли**
Продать больше (\$)
не стоимость, а маржа + расходы на упаковку, доставку и т.п.
- **товары из категорий (long-tail, новинки и т.п.)**
продать большой ассортимент / распродать

Сбор данных

явный (explicit)

- оценка объекта
- ранжирование группы объектов
 - выбор одного товара из двух
- создание списка любимых объектов

неявный (implicit)

- что искал, смотрел, клал в корзину, купил
 - лог поведения
- анализ содержимого компьютера

История исследований в рекомендательных системах

199х – первые алгоритмы (GroupLens)

1995-2000 – внедрение в бизнес

2006 – Netflix prize

2007 – первая конференция

Соревнование Netflix

2006 год

~ 100.5 миллионов оценок 1,2,...,5

~ 480 000 пользователей

17 770 фильмов

RMSE

Netflix = 0.9514

надо = 0.8563

~ 20 000 участников

RBM = 0.8990

SVD = 0.8914

Для бизнеса > 0.88

Рекомендации по контенту (content based methods)

**Если есть хорошие признаковые описания пользователей и объектов (и только они),
тогда**

$$\begin{aligned}u &\sim f_u \\ i &\sim f_i\end{aligned}$$

Можно решать как обычную задачу обучения с учителем

$$\{([f_u, f_i], r_{ui})\}$$

Цель: $u \rightarrow i_1, \dots, i_k : \hat{r}_{ui_1} \geq \hat{r}_{ui_2} \geq \dots$

Рекомендации по контенту (content based methods)**+****решает проблему холодного старта (cold start)****что новым пользователям / какие новые товары****может начать работать «прямо сейчас» – без статистики****рекомендация не зависит от других пользователей****(хм...)****ясность (transparency) можно объяснить****можно много где использовать****–****если есть хороший контент****описания пользователей часто примитивные / товаров ???****извлечение описаний часто отдельная задача****пример: музыка, видео****однообразные рекомендации (overspecialization)****контент же похожий...****при наличии статистики хуже CF****см. дальше**

Коллаборативная фильтрация

Если известна лишь статистика:

$$\{(u, i, r_{ui})\}$$

нет содержательных признаков!

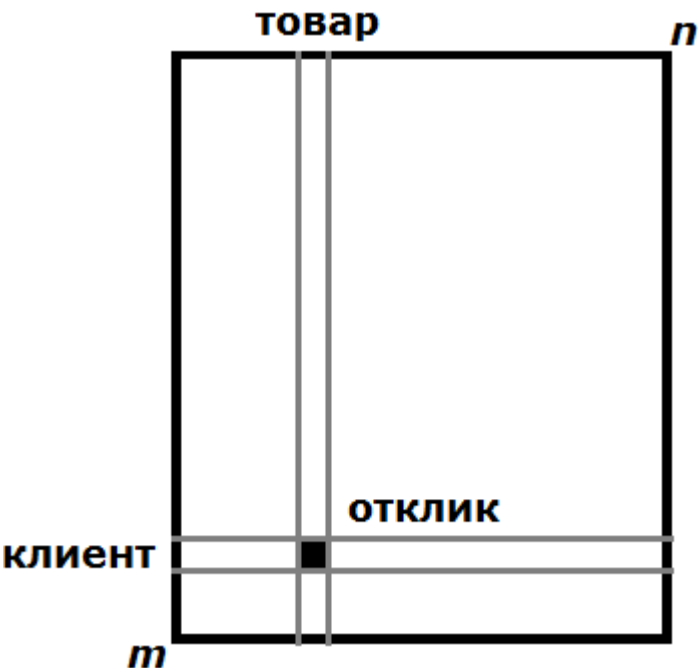
**Решение на статистике поведения лучше,
чем на описаниях!**

статья «Recommending new movies: even a few ratings
are more valuable than metadata» (context: Netflix)

Колаборативная фильтрация

- **memory based / nearest neighbors**
 - **model based**
 - **latent factors**
- **matrix factorization**

Статистика



	item1	item2	item3	item4
user1	1	2	5	
user2		2		5
user3	3	3	5	
user4		4		5
user5	5		3	

Матрица «пользователь – товар» (utility matrix)
разреженная матрица

Цель: фактически уметь дозаполнять матрицу...

GroupLens-алгоритм

По пользователям (User-based)

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_v \text{sim}(u, v)(r_{vi} - \bar{r}_v)}{\sum_v \text{sim}(u, v)}$$

По товарам (Item-based)

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_j \text{sim}(i, j)(r_{uj} - \bar{r}_j)}{\sum_j \text{sim}(i, j)}$$

Идея: как скорректировать простейшие baseline

Проблема холодного старта


Плохие предсказания, если мало статистики

Долгие вычисления (нужен пересчёт)

Похожесть пользователей и товаров**корреляция Пирсона в user-based CF**

$$\text{sim}(u, v) = \frac{\sum_i (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_i (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_i (r_{vi} - \bar{r}_v)^2}}$$

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1



sim = 0,85

sim = 0,00

sim = 0,70

sim = -0,79

Похожесть

- корреляция Пирсона в user-based CF
 - косинусная мера сходства
 - мера Жаккара

но м.б.

похожесть по описанию,
похожесть по кластерам,
...

**Не обязательно такую близость...
выбор k самых близких**

Нет теоретических предпосылок для выбора определённой метрики!

Алгоритм «YouTube»

у видео-роликов мало мета-данных (сравни: книги, фильмы)!

видео-ролики мало живут (сравни: ...)

видео-роликов много, они короткие, шумный отклик (сравни: ...)

- **YouTube video recommendation system (2010)**
- **random walks through the view graph (2008)**
 - **DL for youtube recommendations (2016)**

$$\text{sim}(i, j) = \frac{\text{view}(\{i, j\})}{\text{view}(\{i\}) \cdot \text{view}(\{j\})}$$

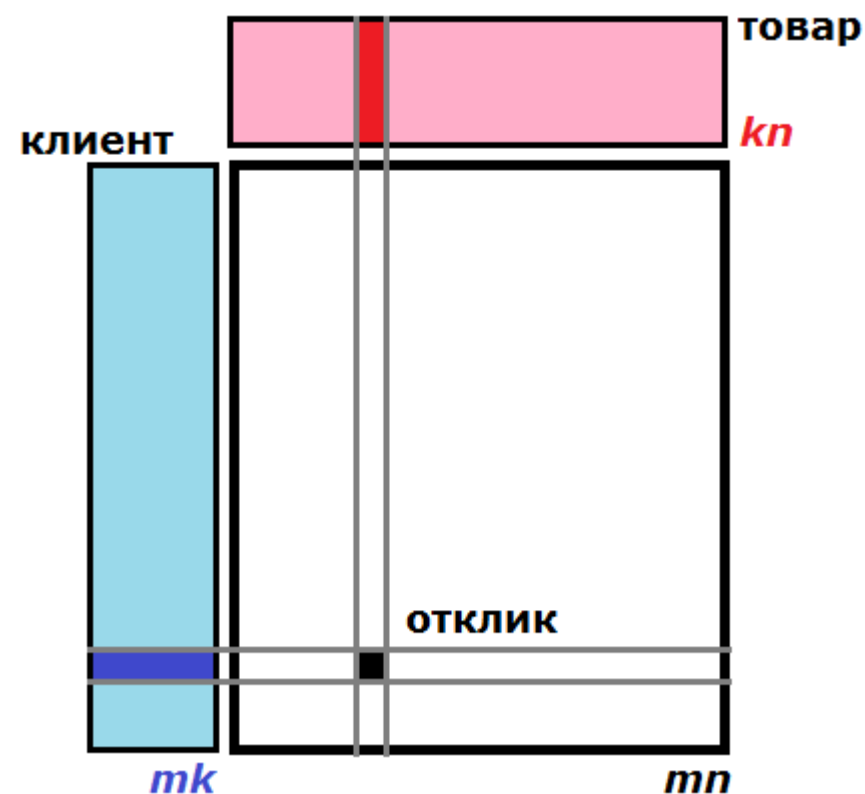
здесь – просмотры за последние 24 часа

Пусть S – просмотренные, понравившиеся, добавленные,

$R(S)$ – похожие на них

рекомендации из $R(S) \cup R(R(S)) \cup \dots$

Рекомендация на основе матричных разложений (SVD)



$$R = U \cdot \Lambda \cdot V^T$$

$$R_{m \times n} \approx U_{m \times k} \cdot \Lambda_{k \times k} \cdot V_{n \times k}^T$$

SVD = сингулярное матричное разложение

Сингулярное разложение матрицы (SVD)

любая $m \times n$ -матрица ранга k представляется в виде произведения

$$X_{m \times n} = U_{m \times k} \cdot \Lambda_{k \times k} \cdot V_{n \times k}^T$$

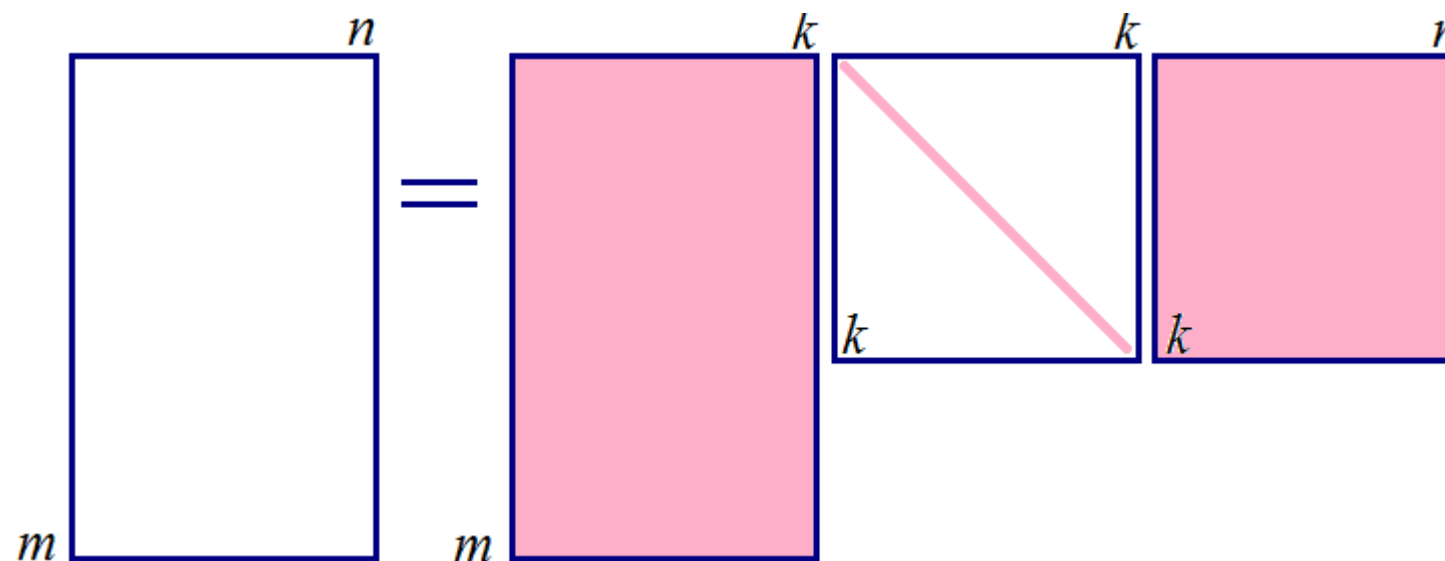
$$X = \sum_{i=1}^k \lambda_i u_i v_i^T$$

где $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_k)$

$$\lambda_1 \geq \dots \geq \lambda_k > 0$$

$$U^T U = I$$

$$V^T V = I$$



Сингулярное разложение матрицы (SVD)

$$X^T X = (U \Lambda V^T)^T U \Lambda V^T = V \Lambda^T U^T U \Lambda V^T = V \Lambda^2 V^T$$

поэтому

$$X^T X V = V \Lambda^2$$

**и матрица V состоит из с.в. матрицы $X^T X$,
которым соответствуют с.з. $\lambda_1^2 \geq \dots \geq \lambda_k^2 > 0$**

$\lambda_1 \geq \dots \geq \lambda_k > 0$ – сингулярные числа

аналогично матрица U состоит из с.в. матрицы XX^T с теми же с.з.

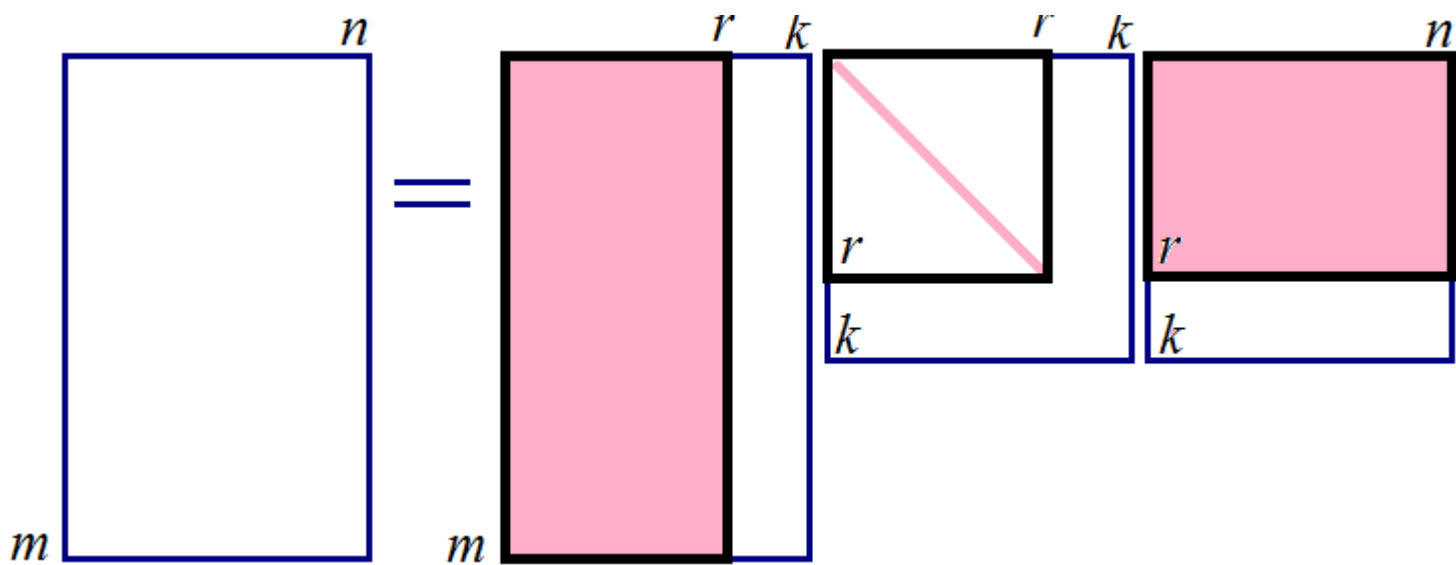
Усечённое сингулярное разложение матрицы (Truncated SVD)

что будет если

$$X = U \Lambda V^T$$

$$\Lambda' = \text{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$$

$$U \Lambda V^T = \sum_{i=1}^r \lambda_i u_i v_i^T \equiv X' = \arg \min_{H: \text{rank } H=r} \| X - H \|_2^2 = \sum_{i=r+1}^k \lambda_i^2$$



Применение SVD

- Для матриц малого ранга – экономное хранение
- Для произвольных матриц – приближение и сжатие
 - Регуляризация
 - Основа некоторых методов рекомендаций
- Основа некоторых методов тематического моделирования
 - Основа некоторых методов сокращения размерности

Реконструкция и сжатие изображений с помощью SVD



k=5



k=10



k=20



k=50



k=100



k=200

Изначальный размер изображения $300 \times 451 = 135\,300$
 $300 \times 50 + 50 + 50 \times 451 = 37\,600$

Минутка кода

```
from numpy.linalg import svd
U, L, V = svd(image)
k = 5
plt.imshow(U[:, :k].dot(np.diag(L[:k])).dot(V[:k, :]),
            cmap=plt.cm.gray)
```

Устойчивость к шумам



k=20



k=50



исходное

Рекомендация на основе матричных разложений (SVD)

$$R \approx U' \cdot V'$$

$$\hat{r}_{ui} = \langle p_u, q_i \rangle$$

SVD также метод CF (Simon Funk)

Рекомендация на основе матричных разложений (SVD)

$$r_{u,i} \approx \langle p_u, q_i \rangle$$

$$J = \sum_{(u,i)} (\langle p_u, q_i \rangle - r_{u,i})^2 + \lambda_1 \sum_u \|p_u\|^2 + \lambda_2 \sum_i \|q_i\|^2$$

Одновременно получили признаковое описание пользователей и товаров $\lambda_t \sim 0.02$

Минимизация

- **градиентный спуск** ($\eta \sim 0.005$)
- **ALS (Alternating Least Squares)**
 - **хорошо параллелится**

$$p_u(t+1) = \left(\sum_{i:r_{u,i}>0} (\langle q_i, q_i \rangle + \lambda_1 I) \right)^{-1} \left(\sum_{i:r_{u,i}>0} r_{u,i} q_i \right)$$

Улучшения модели

$$r_{u,i} \approx r + r_u + r_i + \langle p_u, q_i \rangle$$

Учитываем смещения
«добрый/злой» пользователь
«плохой/хороший» товар

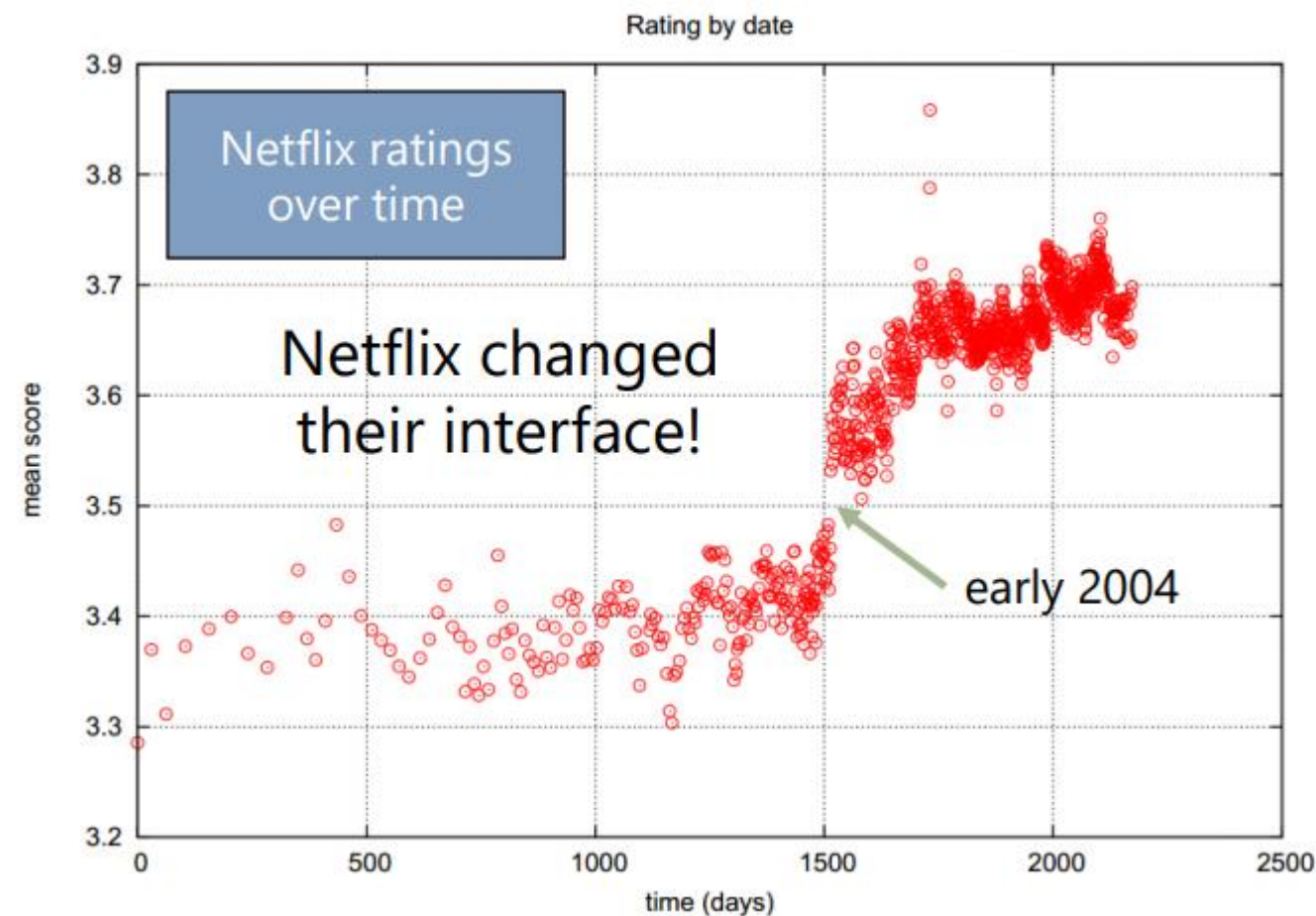
SVD++

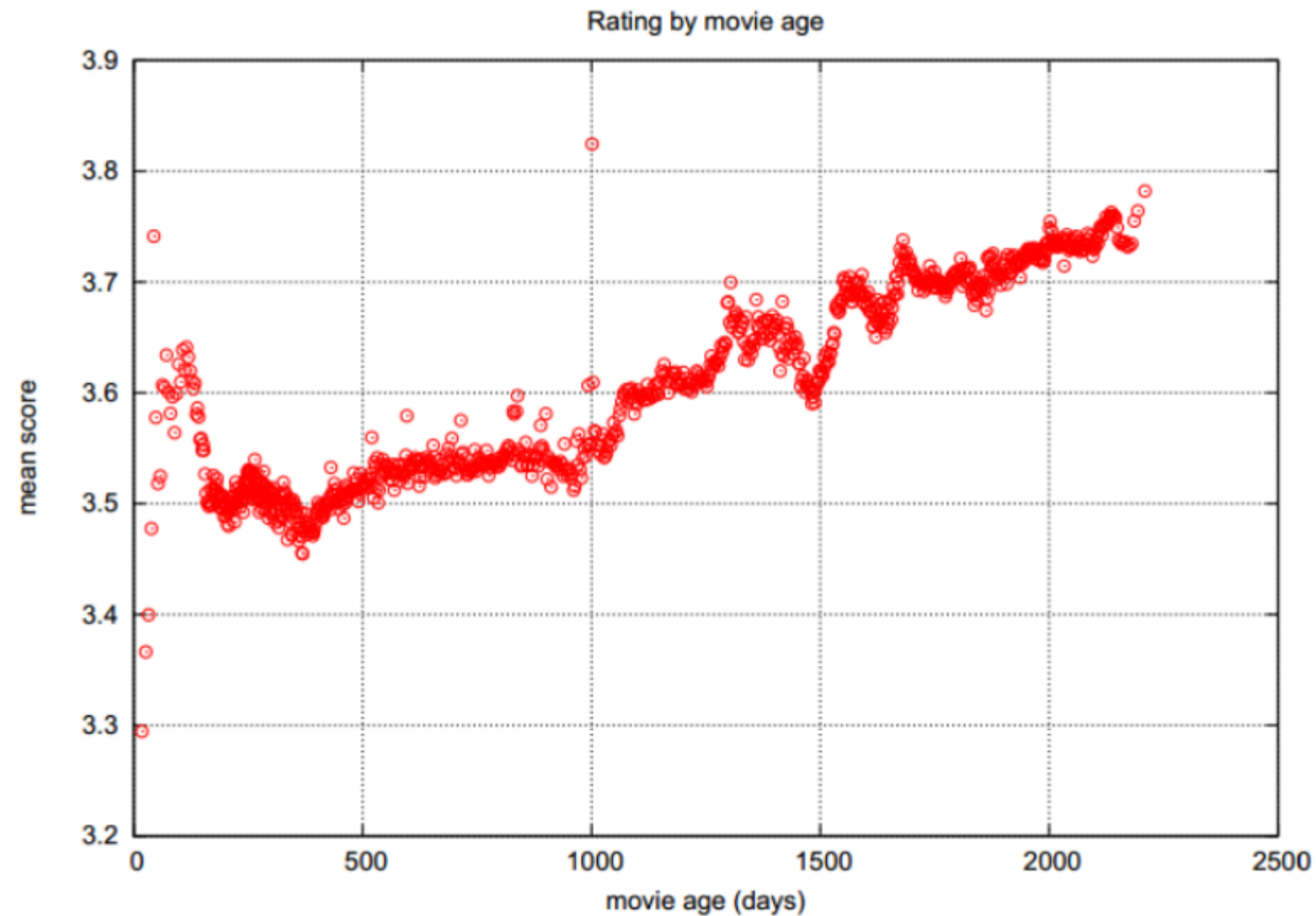
$$r_{u,i} \approx r + r_u + r_i + \left\langle p_u + \frac{1}{\sqrt{|\text{view}(u)|}} \sum_{j \in \text{view}(u)} y_j, q_i \right\rangle$$

+ что просматривал, но не покупал пользователь

Легко обобщать на разное число факторов:
(пользователь, канал, товар)

Simon Funk статья в блоге во время конкурса Netflix

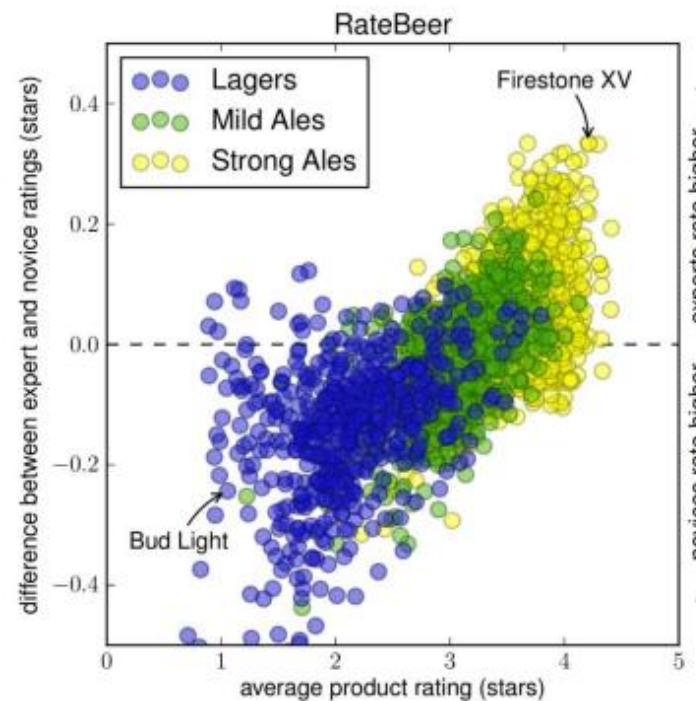
timeSVD++**Неизвестные зависят от времени...****Koren «Collaborative Filtering with Temporal Dynamics» KDD 2009**

timeSVD++

Люди склонны завышать рейтинги старых фильмов
есть много подобных эффектов – вывод: учитывайте время

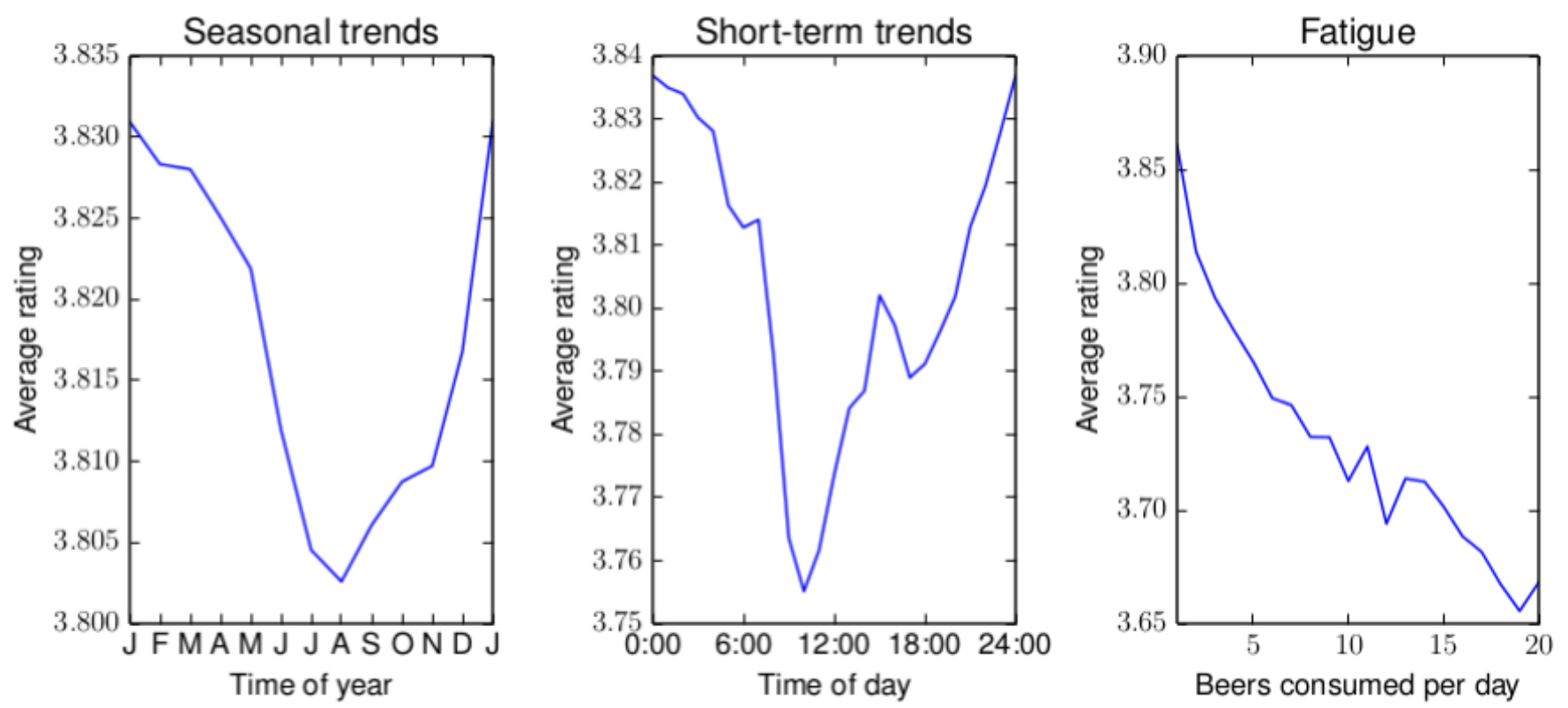
Что происходит со временем

- **меняется интерфейс** [Koren, 2009]
- **начинаем любить ретро** [Koren, 2009]
- **предпочтения меняются** [Godes, Silva, 2012]
- **пользователи меняются** (аккаунт стал семейным [Xiang et al., 2010])
 - **аномалии** (в каникулы смотрел сериал [Xiang et al., 2010])
 - **сезонность, мнение толпы и т.п.** [McAuley, Leskovec, 2013]



Differences between
"beginner" and "expert"
preferences for different
beer styles

Что происходит со временем



timeSVD++

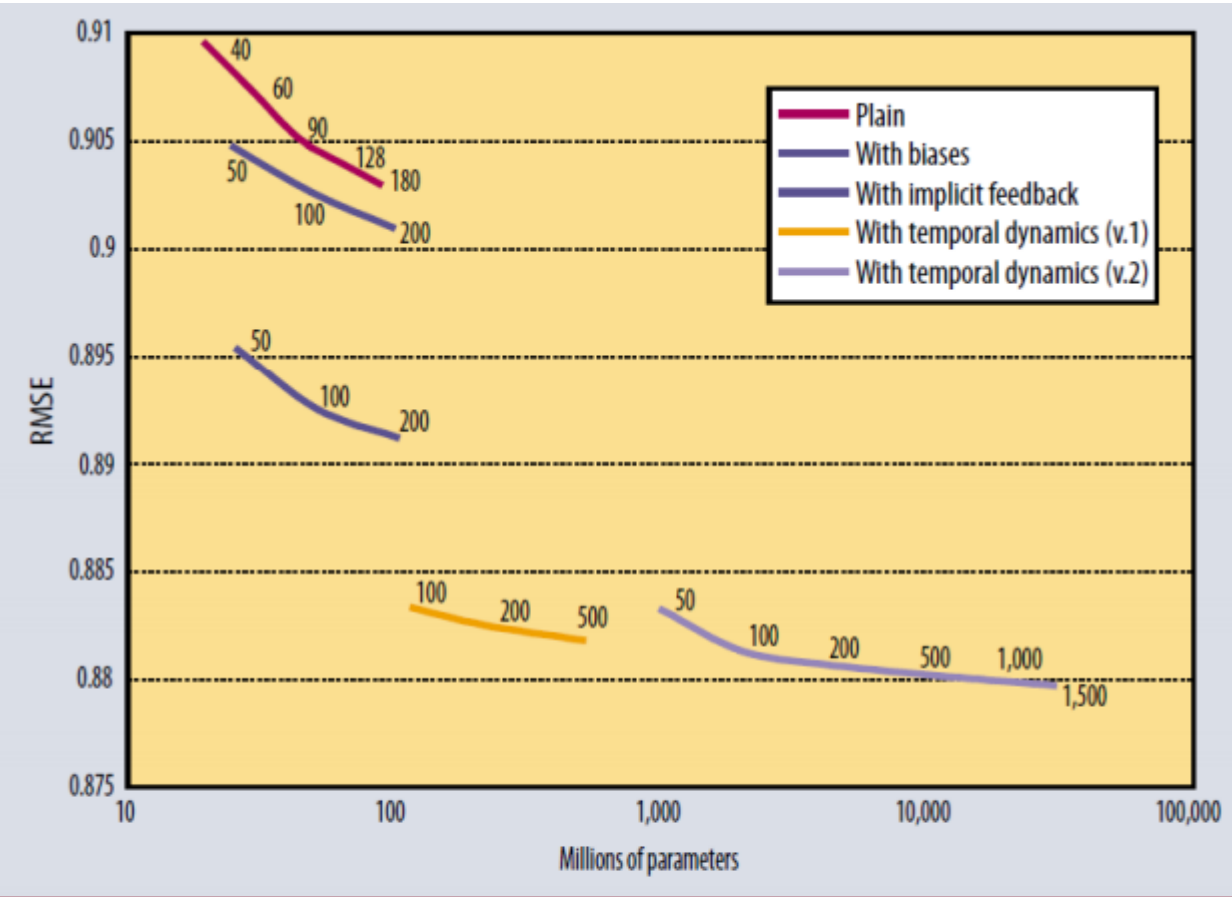


Figure 4. Matrix factorization models' accuracy. The plots show the root-mean-square error of each of four individual factor models (lower is better). Accuracy improves when the factor model's dimensionality (denoted by numbers on the charts) increases. In addition, the more refined factor models, whose descriptions involve more distinct sets of parameters, are more accurate. For comparison, the Netflix system achieves $RMSE = 0.9514$ on the same dataset, while the grand prize's required accuracy is $RMSE = 0.8563$.

Адаптация SVD под социальные связи

$$\sum_{(u,i)} (\langle p_u, q_i \rangle - r_{u,i})^2 + \lambda \sum_u \left\| p_u - \frac{1}{|F(u)|} \sum_{v \in F(u)} p_v \right\|^2 + \\ + \lambda_1 \sum_u \|p_u\|^2 + \lambda_2 \sum_i \|q_i\|^2$$

$F(u)$ – множество друзей u

или (тут по-другому!)

$$+ \lambda \sum_u \sum_{v \in F(u)} \text{sim}(u, v) \|p_u - p_v\|^2$$

можно учитывать похожесть на друзей

<https://www.microsoft.com/en-us/research/wp-content/uploads/2011/01/wsdm10.pdf>

Когда нет явного отклика

Если оценки даны не в шкале,
а перечислены только отклики на услугу...

$$\{(u, i, 1)\}$$

(покупка, скачивание, просмотр и т.п.)

– **более честно** (Netflix ex: highly rated vs watched)!

иногда решение вырождается в константное

выход: пропуски = нули

На практике:

часто знаем, что видел пользователь...

и почему-то не отреагировал

содержание рассылки

баннеры на странице

сбор информации (оценки, лайки) – дополнительные усилия!

One-class recommendation / или неявный отклик

Если есть «лайки» и «дизлайки» (или отсутствие реакции)

$$\{(u, i, +1)\} \cup \{(u, i, -1)\}$$

Можно строить модель «один товар лучше другого»

$$P(i \succ j) \approx \sigma(p_u^T q_i - p_u^T q_j)$$

Стохастический градиентный спуск

~ случайно выцепляем пары сравнимых товаров

+ регуляризация, как всегда

**Steffen Rendle et al «BPR: Bayesian Personalized Ranking from Implicit Feedback» //
Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence
(2009), с. 452—461. <https://arxiv.org/pdf/1205.2618.pdf>**

<https://implicit.readthedocs.io/en/latest/quickstart.html>

CF для Implicit Feedback

$$J = \sum_{(u,i)} c_{u,i} (p_u^T q_i - I[r_{u,i} \geq 0])^2 + \lambda \left(\sum_u \|p_u\|^2 + \sum_i \|q_i\|^2 \right)$$

Здесь сумма по всем(!) парам «пользователь-товар»

Если есть оценка взаимодействия, то она влияет лишь на множитель:

$$c_{u,i} = \begin{cases} 1 + \alpha r_{u,i} \\ 1 + \alpha \log(1 + \beta r_{u,i}) \end{cases}$$

Минимизация методом ALS

Yifan Hu, Yehuda Koren и Chris Volinsky «Collaborative Filtering for Implicit Feedback Datasets» // IEEE International Conference on Data Mining, 2008, с. 263—272 <http://yifanhu.net/PUB/cf.pdf>

<https://implicit.readthedocs.io/en/latest/quickstart.html>

Logistic Matrix Factorization

$$P(u \leftrightarrow i) \approx \sigma(p_u^T q_i + \alpha_u + \beta_j)$$

**Вероятность взаимодействия пользователя и товара записываем в виде сигмоиды,
не взаимодействия:**

$$P(u \nleftrightarrow i) \approx 1 - \sigma(p_u^T q_i + \alpha_u + \beta_j)$$

**Выписываем правдоподобие и максимизируем
+ регуляризация**

Christopher C. Johnson «Logistic matrix factorization for implicit feedback data» //
Distributed Machine Learning and Matrix Computations (2014)

<https://web.stanford.edu/~rezab/nips2014workshop/submits/logmat.pdf>

<https://implicit.readthedocs.io/en/latest/quickstart.html>

Коллаборативная фильтрация – минусы

- **проблема холодного старта (cold start)**
другая техника: по контенту, не персональные и т.п.
система рейтинга (обратная связь), костыли (по умолчанию)
- **популярные становятся популярнее (popularity bias)**
- **условия шума (семейные аккаунты, случайные покупки и т.п.)**
 - **возможны «атаки» на систему**

Факторизационные машины



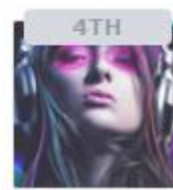
1st/239



4th/657



4th/163



4th/133

Steffen Rendle**libFM: Factorization Machine Library**<http://www.libfm.org/>

**Супермодель, иммитирует
SVD, SVD++, FPMC, Pairwise interaction tensor factorization,
SVM с полином. ядром и т.п.**

Ask Peter Norvig

Q5: What, say, 3 recent papers in machine learning do you think will be influential to directing the cutting edge of research these days? (41 Up-votes, 26.08.2014)

I've never been able to pick lasting papers in the past, so don't trust me now, but here are a few:

Rendle's "Factorization Machines"

Wang et al. "Bayesian optimization in high dimensions via random embeddings"

Dean et al. "Fast, Accurate Detection of 100,000 Object Classes on a Single Machine"

Факторизационные машины

Feature vector \mathbf{x}																Target \mathbf{y}						
$\mathbf{x}^{(1)}$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...	5	$y^{(1)}$
$\mathbf{x}^{(2)}$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...	3	$y^{(2)}$
$\mathbf{x}^{(3)}$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...	1	$y^{(2)}$
$\mathbf{x}^{(4)}$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...	4	$y^{(3)}$
$\mathbf{x}^{(5)}$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...	5	$y^{(4)}$
$\mathbf{x}^{(6)}$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...	1	$y^{(5)}$
$\mathbf{x}^{(7)}$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...	5	$y^{(6)}$
	A	B	C	...	TI	NH	SW	ST	...	TI	NH	SW	ST	...	Time	TI	NH	SW	ST	...		
	User				Movie					Other Movies rated						Last Movie rated						

$$r_{ui} \sim w_0 + w_u + w_i + v_u^T v_i$$

модель второго порядка:

$$w_0 + \sum_{i=1}^n w_i x_i + \sum_{1 \leq i < j \leq n} v_i^T v_j x_i x_j \sim w_0 + w^T x + x^T \underbrace{W}_{\sim \text{rg}=k} x$$

«факторизация» – в предположении, какая у нас матрица весов,
иначе была бы просто «модель второго порядка»

Факторизационные машины

Что ещё...

- факторизация отдельных блоков (FFM – field-aware factorization machine)
 - эффективное блочное хранение

Факторизационная машина с полями FFM – field-aware factorization machine

criteo labs

Avazu

«RecSys 2015»

outbrain



Линейная модель

$$w^T x = \sum_{i=1}^n w_i x_i$$

Полиномиальная модель (Poly2)

$$x^T W x = \sum_{1 \leq i < j \leq n} w_{ij} x_i x_j$$

Факторизационная машина

$$x^T V^T V x = \sum_{1 \leq i < j \leq n} v_i^T v_j x_i x_j$$

Факторизационная машина с полями

$$\sum_{1 \leq i < j \leq n} v_{i,f(j)}^T v_{j,f(i)} x_i x_j$$

$f(i)$ – поле для i

Оптимизационная задача

$$\sum_{t=1}^m \left(\log(1 + \exp(-y_t \varphi(w, x_t))) + \lambda \|w\|^2 \right) \rightarrow \min$$

$$\varphi(w, x) = \sum_{1 \leq i < j \leq n} w_{i, f(j)}^T w_{j, f(i)} x_i x_j$$

LogLoss + регуляризация

Что такое поля...

Field name		Field index
User	→	field 1
Movie	→	field 2
Genre	→	field 3
Price	→	field 4

Что ещё?

- неотрицательные матричные разложения
 - вероятностные разложения
 - специальные регуляризаторы
 - локальная низкоранговость
 - бикластеризация
- тензоры (тензорное разложение)

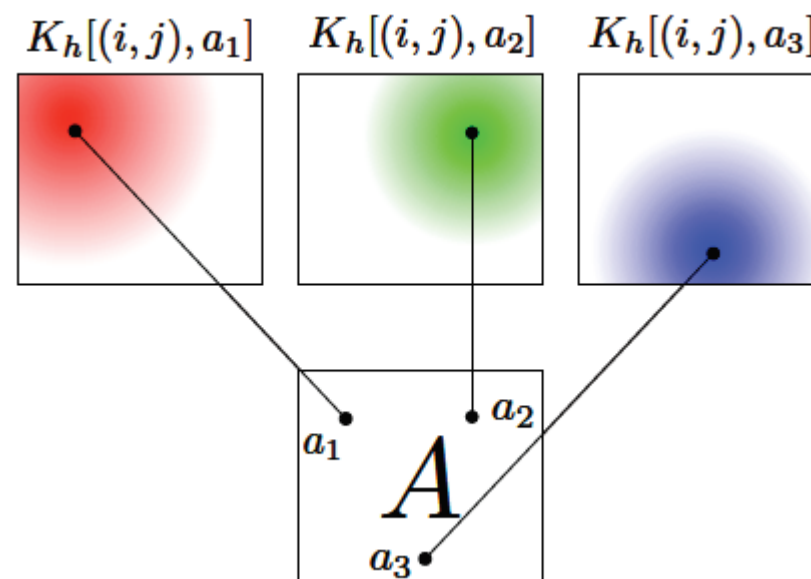
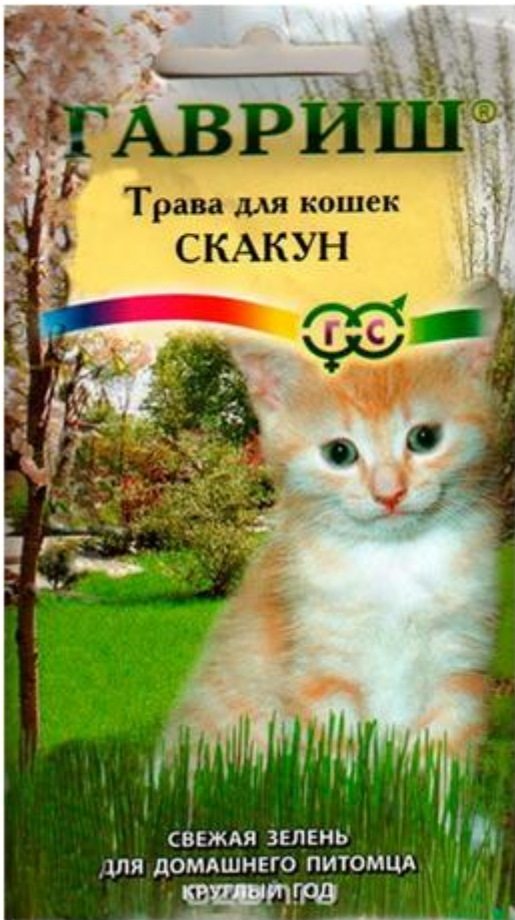


рис. из дипломной работы М.Трофимова

Простые методы рекомендаций



Трава для кошек Скакун, 10 г

Тип	Комнатные растения
Вид	Разнообразные комнатные
Время посадки в грунт	Январь, Февраль, Март, Апрель, Май, Июнь, Июль, Август, Сентябрь, Октябрь, Ноябрь, Декабрь
Время урожая	Январь, Февраль, Март, Апрель, Май, Июнь, Июль, Август, Сентябрь, Октябрь, Ноябрь, Декабрь
Назначение	Для контейнеров

15 р

Добавить в корзину

Вместе с этим товаром покупают

☒ 
Трава для кошек Скакун, 10 г

+

☒ 
Фикус Притупленный, 3 шт.

+

☒ 
Нолина (бокарня отогнутая) Бутылочное дерево, 3 шт.

= 85 р

В корзину

Бандлы ~ по статистике

Простые методы рекомендаций

FPM – Frequent Pattern Mining

- Ассоциативные правила (Association Rule Mining)

если $\{A, B, C\} \Rightarrow D$ (были в одной сессии)

- Sequential Pattern Mining

если $A \rightarrow \dots \rightarrow B \rightarrow \dots \rightarrow C \Rightarrow D$ (были до)

Contiguous Sequential Pattern Mining

если $A \rightarrow B \rightarrow C \Rightarrow D$ (были последовательно перед)

Кластеризация пользователей / товаров

(+ стандартные рекомендации)

есть и автоматические кластеры

(интересы, любимые театры / жанры, актёры и т.п.)

Методы на основе случайных блужданий

Laknath Semage «Recommender Systems with Random Walks: A Survey» //

<https://arxiv.org/pdf/1711.04101.pdf>

Случайные блуждания в RecSys

Node2Vec

идея skip-gram для кодирования вершин графа, ходим по графу – генерируем посл-ть вершин, используем посл-ти для получения вложения

[Grover and Leskovec, 2016]

аугментация данных

хождение по двудольному графу с целью пополнения выборки
(user, item) + близкие вершины из хождения

Функционалы качества

Уже было...

- **RMSE (Netflix)**
- **Precision, Recall**
- **NDCG**

Желаемые свойства рекомендаций

это всё сложно оценить!

Разнообразие (diversity) ~ непохожие на другие товары из списка

Плохо: к ноутбуку только ноутбуки того же производителя

Новизна (novelty) ~ для пользователя

Плохо: каждый день одно и то же

Серендипность (Serendipity) ~ неожиданная, но полезная находка

Хорошо, если пользователь открывает для себя новые товары

Доверие ~ обосновать рекомендацию

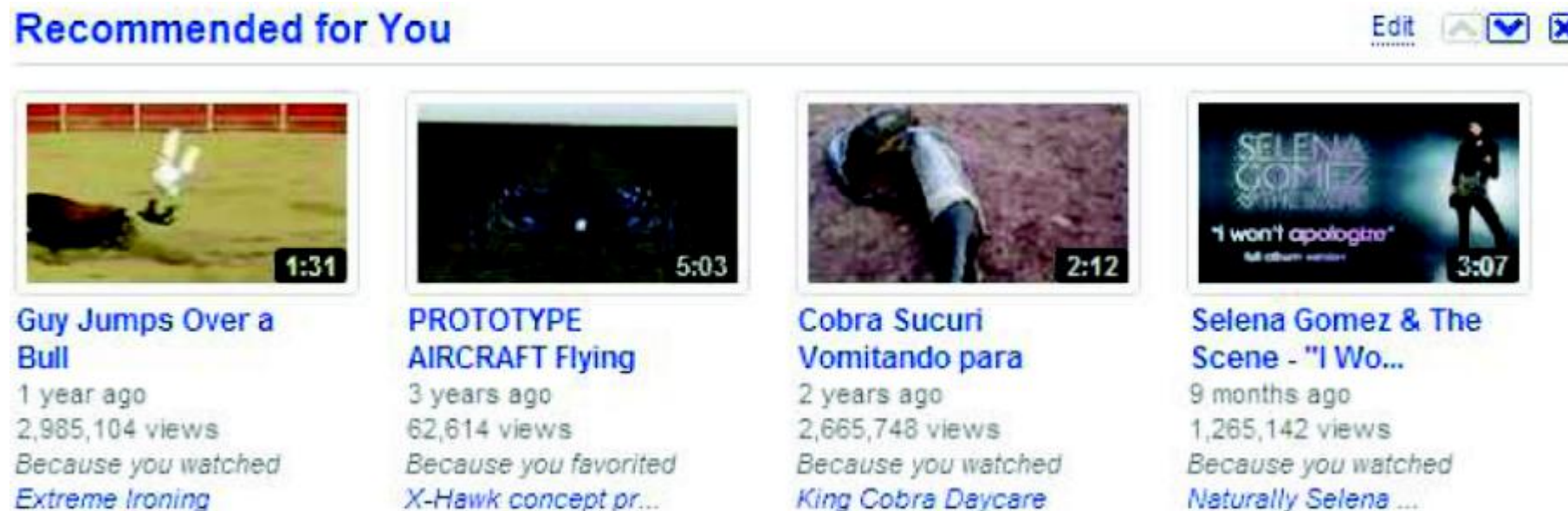
«с товаром покупают», «скидка за комплект», ...

+ лёгкость внедрения / эффективность и удобство эксплуатации

Желаемые свойства рекомендаций

Как делают (пример YouTube)

- **глобальный рейтинг** (просмотры, оценки, комментарии, пересылка)
 - **предпочтения пользователя** (текущее видео + история)
- **лимиты** (на видео одного автора, последовательности видео и т.д.)



Неидентифицируемые пользователи + новые товары

многорукие бандиты

тут RL

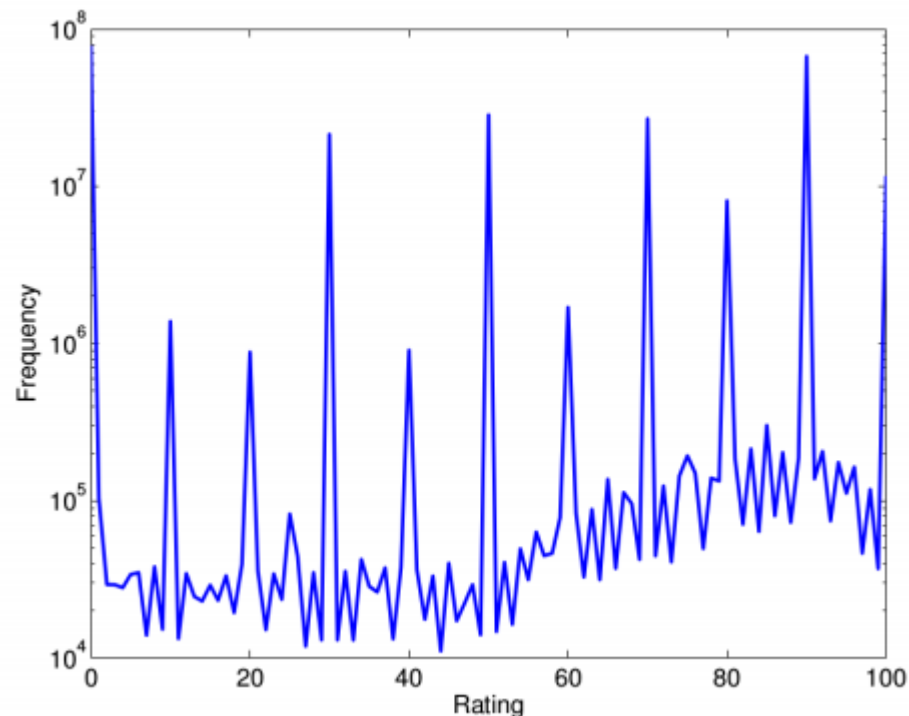
даже в неперсональных рекомендациях

исследование (exploration) – сбор статистики

использование (exploitation) – рекомендация топовых

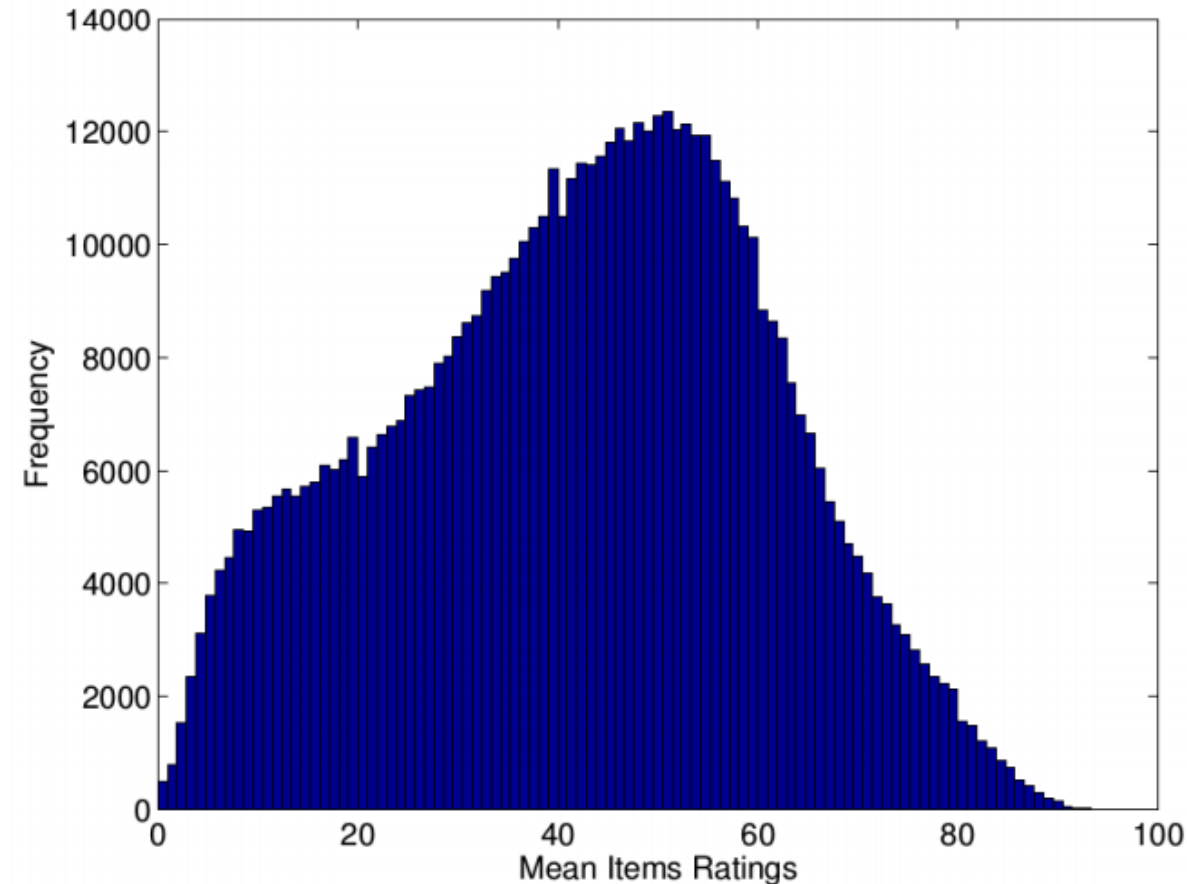
Немного о реакции пользователей

Yahoo! Music Recommendations: Modeling Music Ratings with Temporal Dynamics and Item Taxonomy (2011)



Распределение проставляемых рейтингов

Есть приложения, где только звёздочки



**Распределение средних рейтингов
композиций**

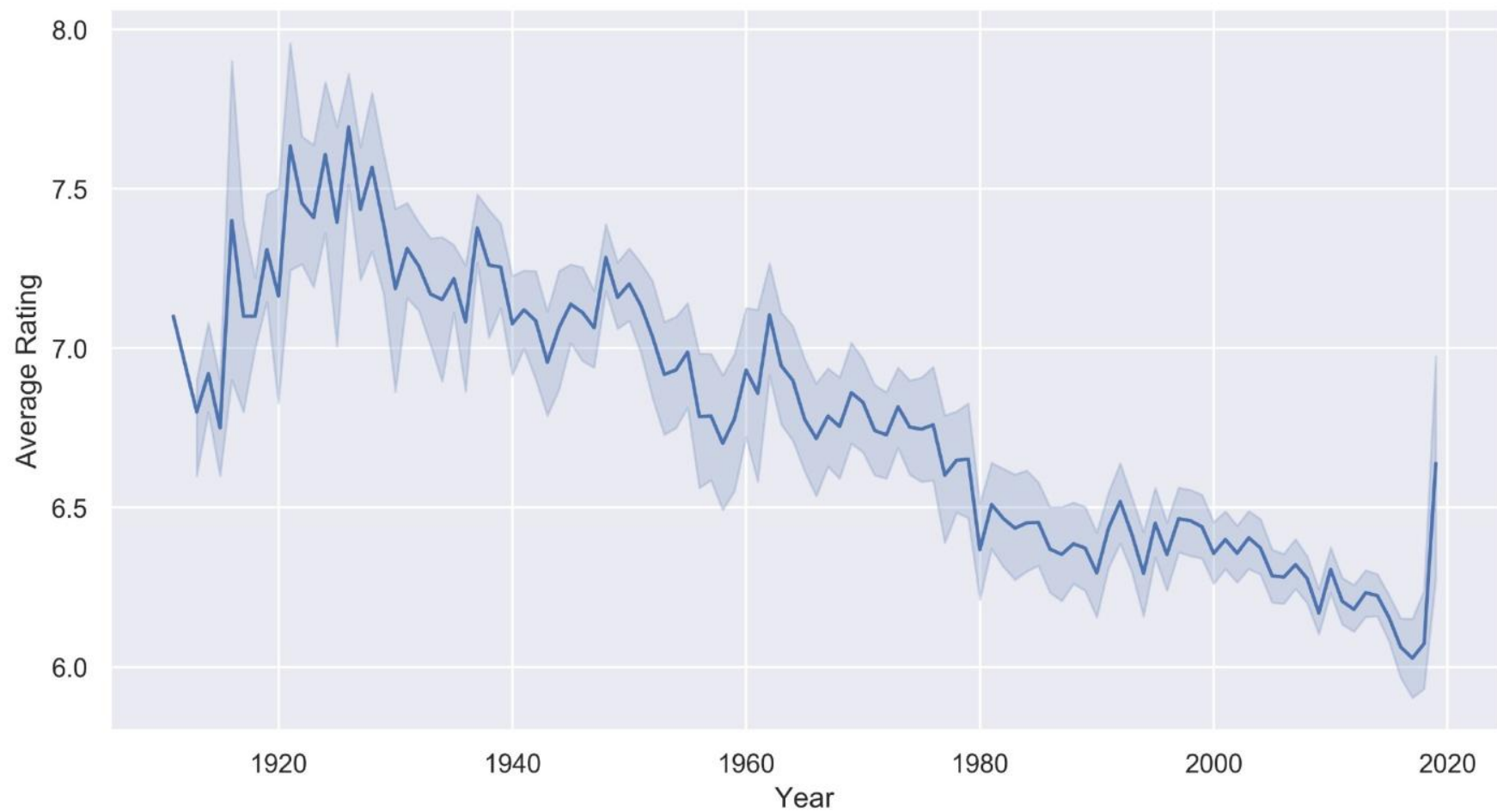
Немного о реакции пользователей

Рейтинг не репрезентативен!



Пользователи оставляют отзыв в специальных случаях

Немного о реакции пользователей



Рейтинг а зависимости от года выпуска фильма

https://www.reddit.com/r/dataisbeautiful/comments/cymkx6/relationship_between_imdb_movie_rating_and/

Литература

Дьяконов А.Г. Алгоритмы для рекомендательной системы: технология LENKOR // Бизнес-Информатика, 2012, №1(19), С. 32–39.

[https://bijournal.hse.ru/2012--1\(19\)/53535879.html](https://bijournal.hse.ru/2012--1(19)/53535879.html)

Y. Koren, R.M. Bell, C. Volinsky Matrix Factorization Techniques for Recommender Systems // IEEE Computer 42(8): 30-37 (2009).

S. Funk Netflix Update: Try This at Home // <http://sifter.org/~simon/journal/20061211.html>

libFM: Factorization Machine Library // <http://www.libfm.org/>

FFM – field-aware factorization machine (слайды) // <http://www.csie.ntu.edu.tw/~r01922136/slides/ffm.pdf>

Литература

Книга по коллаборативной фильтрации

Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan «Collaborative Filtering Recommender Systems»

<https://md.ekstrandom.net/pubs/cf-survey.pdf>

Курс по RS: PV254 Recommender Systems

<https://www.fi.muni.cz/~xpelanek/PV254/>

список ресурсов

https://github.com/grahamjenson/list_of_recommender_systems

<https://gist.github.com/entaroadun/1653794>

Хорошая презентация

<https://www.slideshare.net/MassimoQuadrana/personalizing-sessionbased-recommendations-with-hierarchical-recurrent-neural-networks>