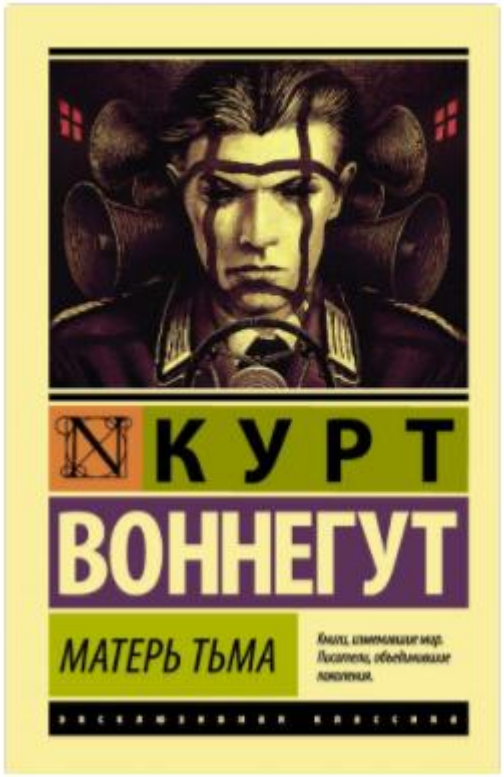
Прикладные задачи анализа данных

РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ
RECOMMENDER SYSTEMS

Дьяконов А.Г.

Московский государственный университет
имени М.В. Ломоносова (Москва, Россия)

Системы рекомендаций



★★★★★ 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»

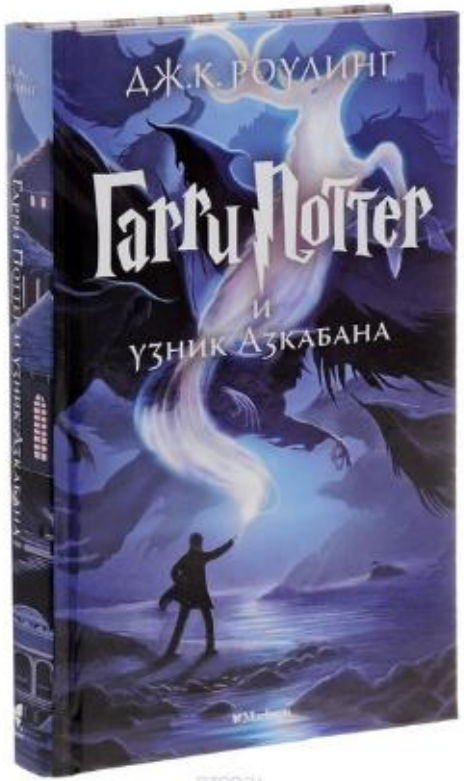
ISBN: 978-5-17-099474-8

Артикул: p1600862

Возрастное ограничение: 16+

Похожие товары

Системы рекомендаций



Бестселлер

Гарри Поттер и узник Азкабана

14 отзывов В избранное Поделиться Код товара: 31275832

Твердый переплет (2)	Букинист. издание (2)	Нет в продаже
от 414 ₽	от 2 469 ₽	9 изданий

Ориг.название	Harry Potter and the Prisoner of Azkaban
Автор	Джоан Кэтлин Роулинг
Формат издания	130x200 мм (средний формат)
Количество страниц	528
Год выпуска	2015

Показать все характеристики

О книге

Книга, покорившая мир, эталон литературы для читателей всех возрастов, синоним успеха. Книга, сделавшая Дж.К.Роулинг самым читаемым писателем совре... Читать далее

414 ₽

✓ В наличии

Курьер доставит завтра

Добавить в корзину

Продавец:
OZON.ru

Рекомендуем также

469 ₽

Гарри Поттер и Кубок Ог...
Дж. К. Роулинг

1 160 ₽

Гарри Поттер и философ...
Дж.К. Роулинг

414 ₽

Гарри Поттер и Тайная ко...
Дж. К. Роулинг

509 ₽

Гарри Поттер и Орден Фе...
Дж. К. Роулинг

489 ₽

Гарри Поттер и Дары См...
Дж. К. Роулинг

414 ₽

Гарри Поттер и Ф...
Дж. К. Роулинг

Системы рекомендаций

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

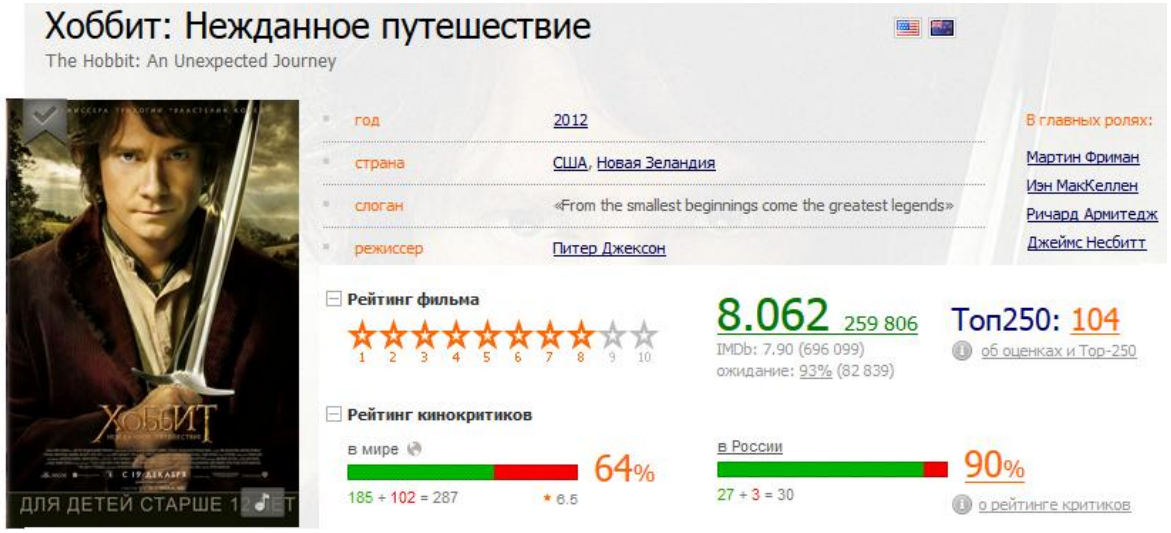
~~3 790 руб~~ 3 220 руб

Nike / Свитшот W NK TOP VERSA CREW

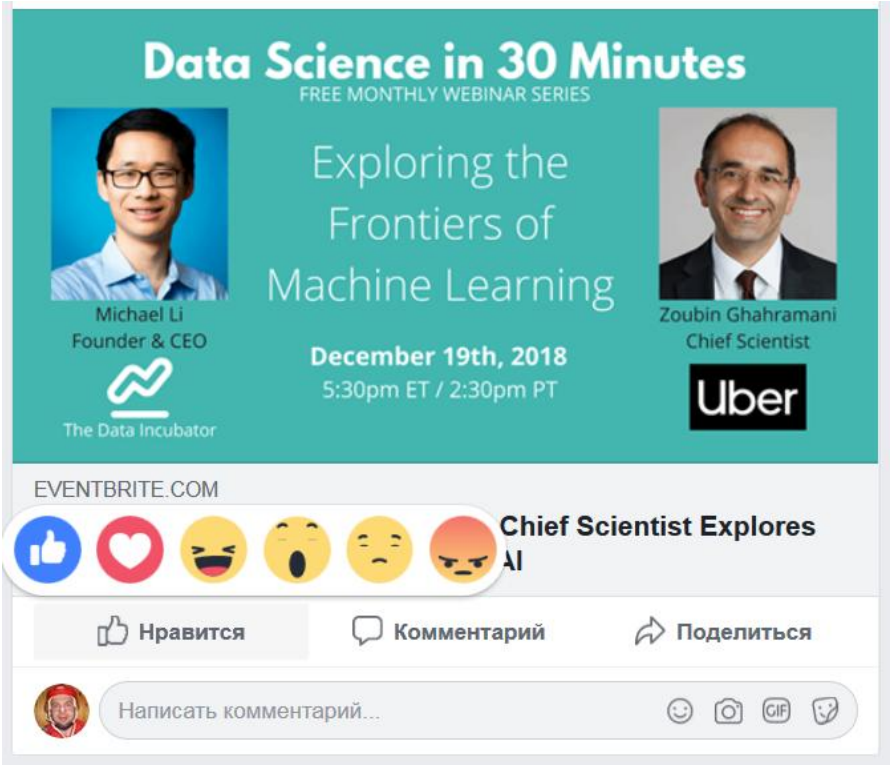
1 699 руб

Mango / Очки солнцезащитные - NAOMI

Системы рекомендаций



Что смотрят?



Можно давать обратную связь

Системы рекомендаций (с точки зрения пользователя)

«то, что мы любим»

что интересно данному пользователю

в данный момент времени

в данном контексте

«то, что подходит»

«что может понравится – что ищем»

~ моделирование предпочтений и поведения

Помощь в поиске товара / услуги!

Системы рекомендаций

товары	книги фильмы музыка игры приложения
контент	новости сайты статьи видео-курсы
досуг	рестораны отели театральные представления выставки туры
социальные связи	друзья группы
услуги	медосмотр

Виды рекомендаций

по контенту **Content-based**	**Рекомендация похожих по описанию товаров**
коллаборативная фильтрация **Collaborative Filtering**	**Рекомендация по статистике покупок** **Проблема холодного старта:** **новый товар** **новый пользователь**
гибридная **Hybrid**	
non-personalized	
demographic	
knowledge-based	

Информация

Описание пользователя

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

Описание товара

Взаимодействие (пользователь, товар)

Взаимодействие (пользователь, пользователь)

Взаимодействия (товар, товар)

Что рекомендуют

заменители (alternative)

сопутствующие товары (cross sell)

бандлы

аксессуары (up sell)

популярные товары (best sellers)

персональные / неперсональные

оффлайн / онлайн

Как рекомендуют / цели бизнеса

- **max вероятность покупки**

Увеличить удовлетворение пользователя (satisfaction, fidelity)

Понять, что нужно людям

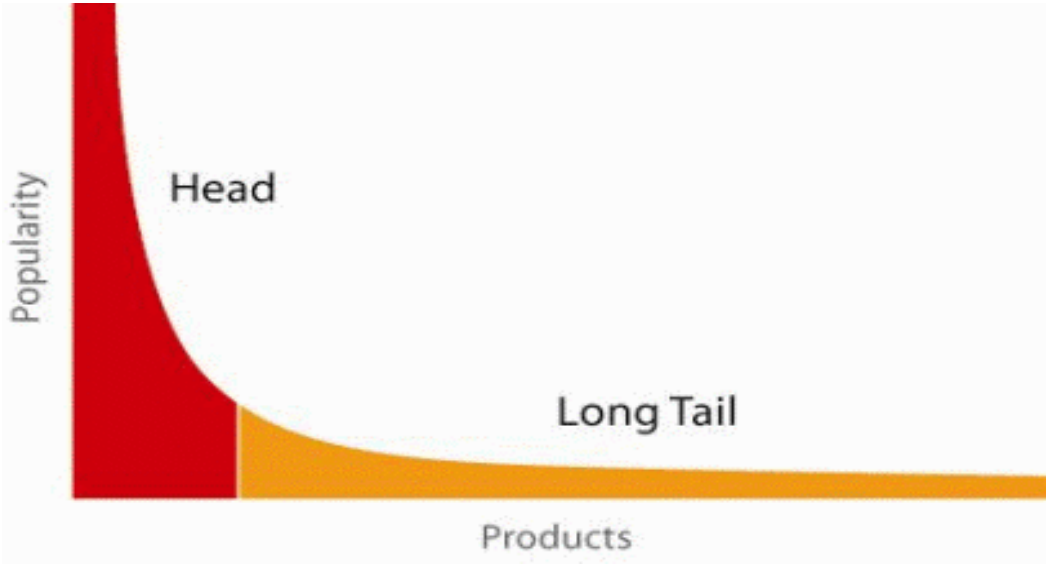
- **max матожидание прибыли**

Продать больше ($)

- **товары из категории (long-tail)**

Продать больший ассортимент / распродать

Пример: рекомендация long-tail



Амазон: 20% товаров – 76% продаж

Разные каналы рекомендаций

сайт

почта

смс, приложения

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

Мифы о рекомендательных системах

Если улучшить ленту на главной странице,
то покупки с неё увеличатся на 30%

Меньше 10% смотрят на ленту главной страницы

70% всех покупок совершаются по рекомендациям (Amazon – 35%)
2/3 всех фильмов смотрят по рекомендациям (Netflix)

Настоящая эффективность рекомендаций (сколько покупок только благодаря им) меньше 10%

После внедрения на 38% больше кликов (Google news)

Насколько здесь ответственна рекомендации,
а не наполненность страницы

Подводные камни

Внедрение РС, как правило, нетривиально

- **Будет ли ценность?**
- **Достаточно ли товаров / пользователей?**
- **Знают ли пользователи, что ищут?**

Разница между информационным поиском и рекомендательными системами

IR	RecSys
«Я знаю, что я ищу»	«Я не уверен, что мне надо»

Подводные камни

Какие цели системы?

Как её оценивать?

Что такое «хорошая» рекомендация?

Пример – рекомендации в обучении

История

199x – 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

По контенту (content based methods)

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

$$u \sim f_u$$
$$i \sim f_i$$

Можно решать как обычную задачу обучения с учителем

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

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

По контенту (content based methods)

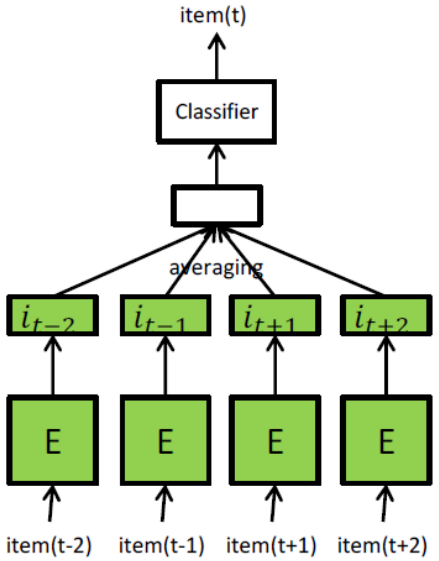
Хорошие признаки

- **даны**

пол, возраст, рейтинг, число лайков и т.п.

- **word2vec, GloVe**
- **матричные разложения**
- **Deep Walk (графовые)**
- **Автокодировщики**

word2vec → prod2vec



По контенту (content based methods)

+

решает проблему холодного старта (cold start)

что новым пользователсм / 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!

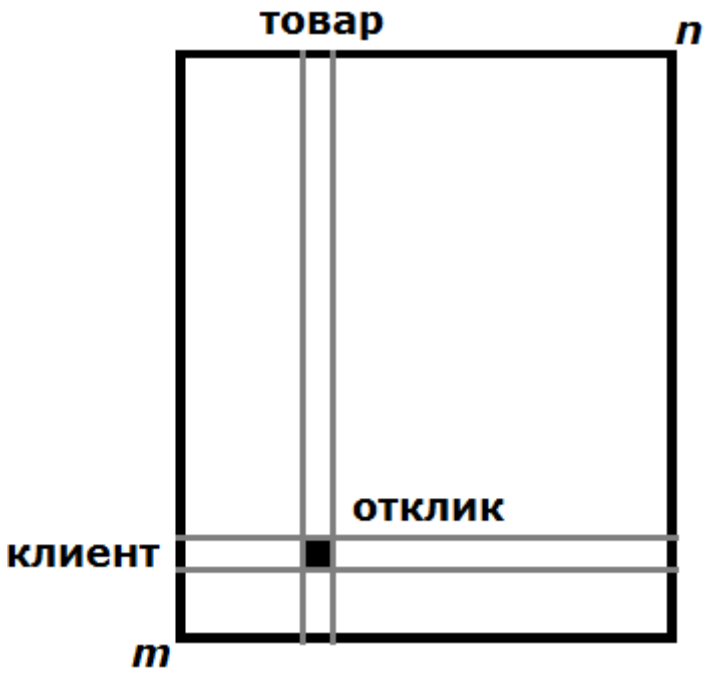
Решение на статистике поведения лучше, чем на описаниях!

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

корреляция Пирсона в user-based CF

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

	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

но м.б.
похожесть по описанию,
похожесть по кластерам,
…

Не обязательно такую близость…
выбор k самых близких

Нет теоретических предпосылок для выбора определённой метрики!

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)**

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

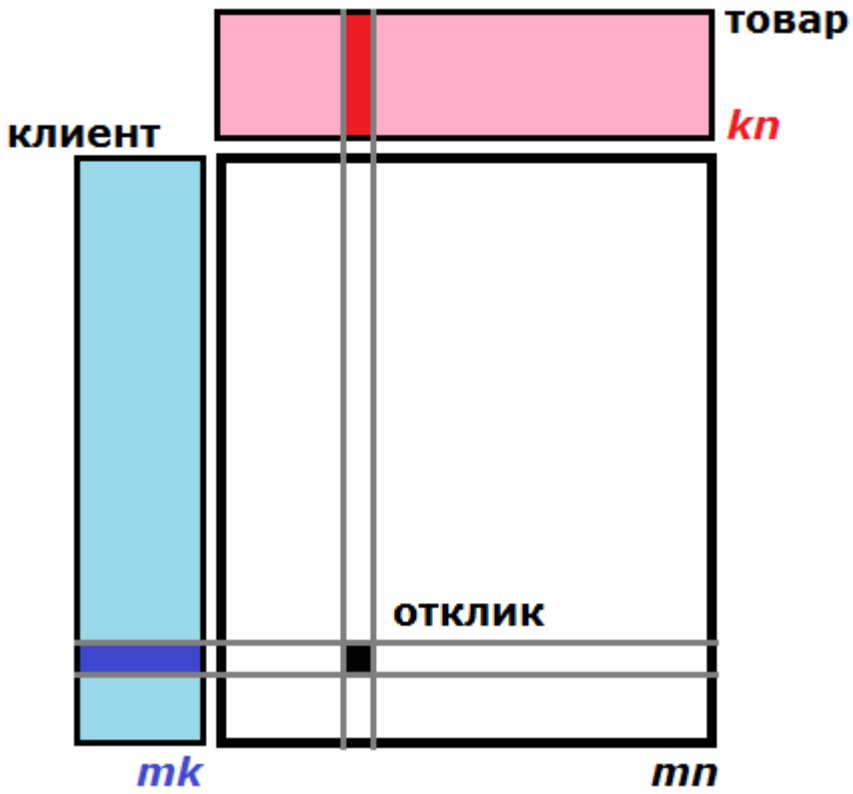
здесь – просмотры за последние 24 часа

Пусть S **– просмотренные, понравившиеся, добавленные,**

$R(S)$ **– похожие на них**

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

SVD



$$R = U \cdot \Lambda \cdot V^{\mathrm{T}}$$

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

SVD = сингулярное матричное разложение

SVD

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

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

SVD также метод CF (Simon Funk)

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)**

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

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++

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

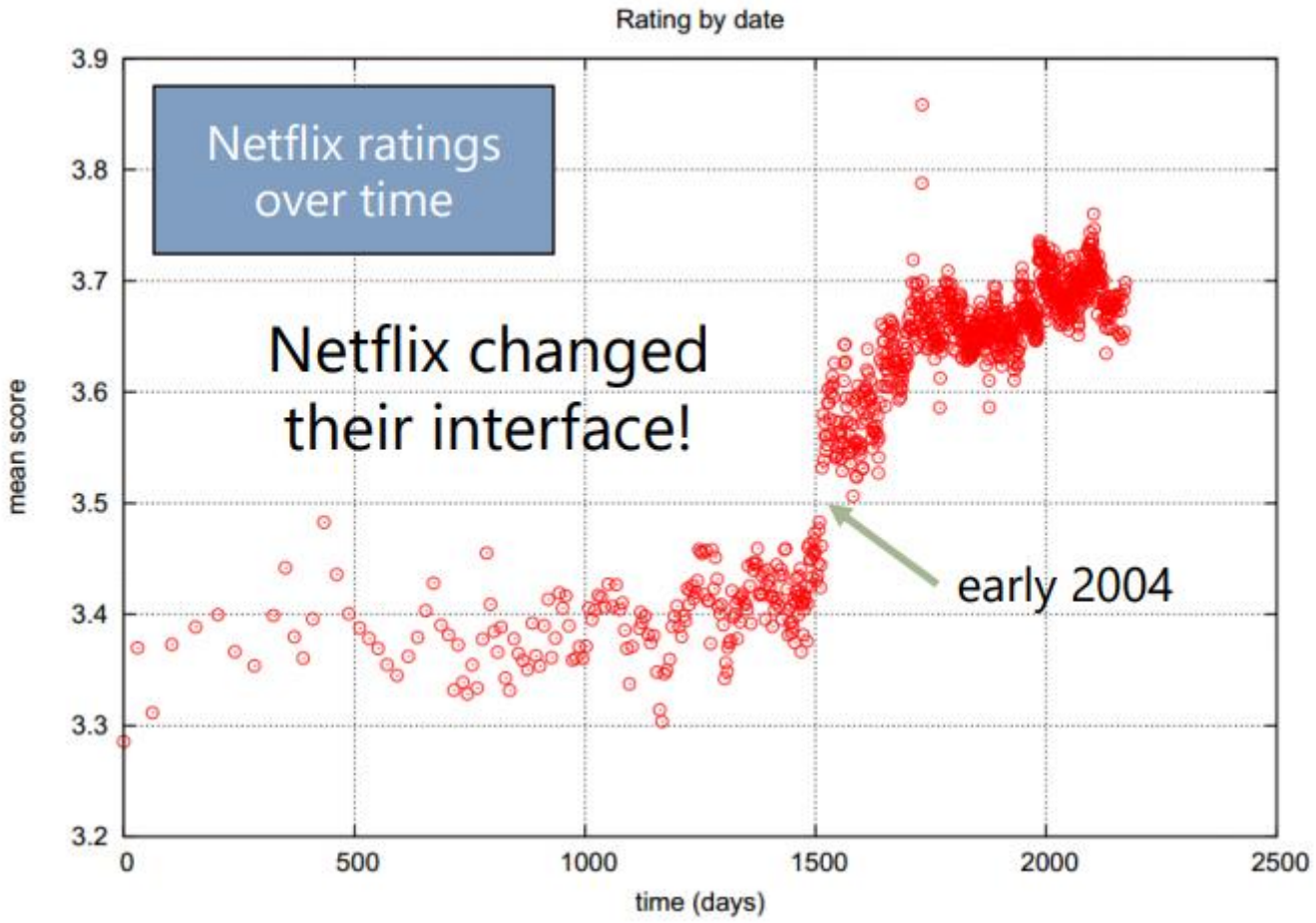
+ что просматривал, но не покупал пользователь

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

Simon Funk статья в блоге во время конкурса Netflix

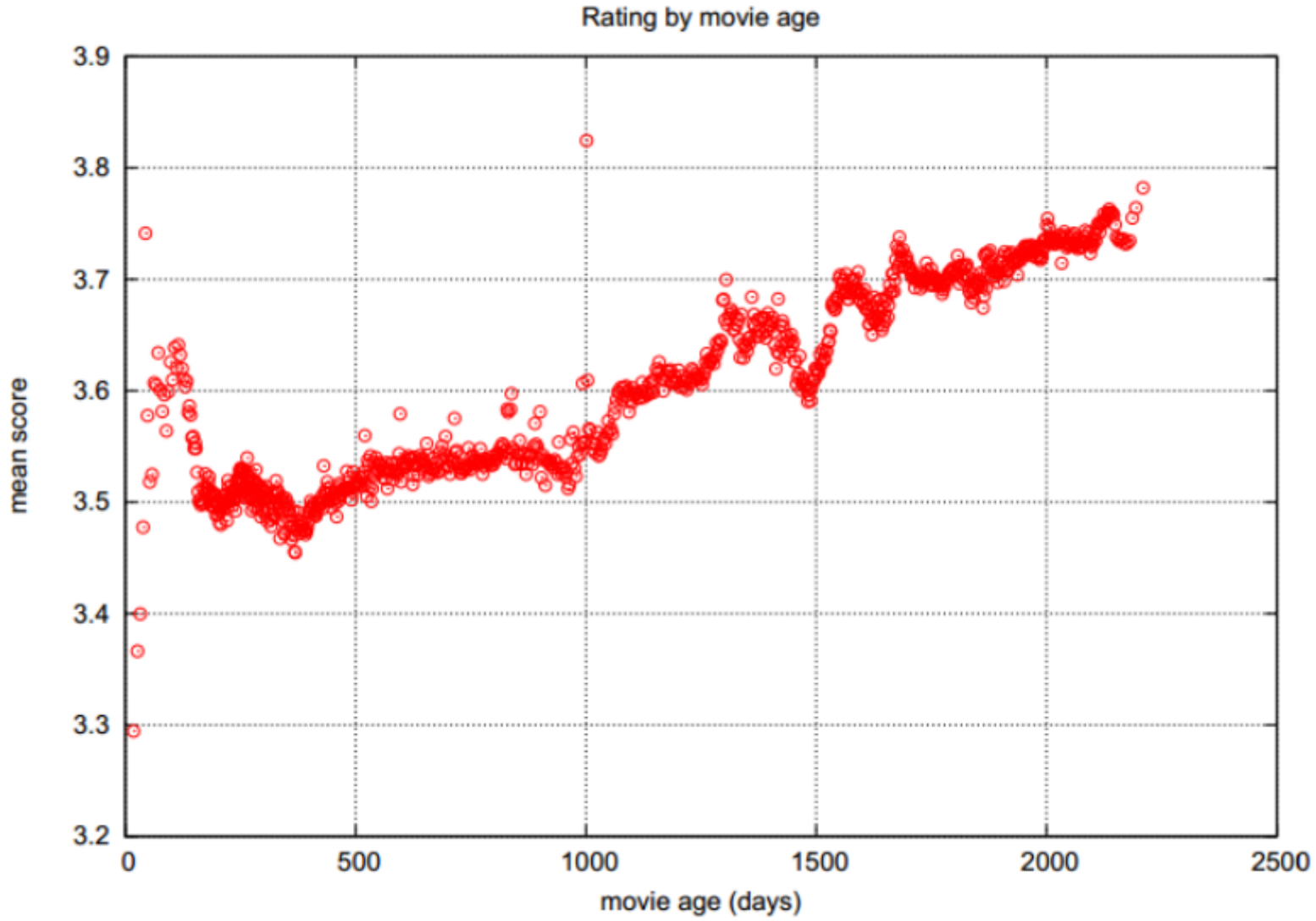
timeSVD++

Неизвестные зависят от времени…



Koren «Collaborative Filtering with Temporal Dynamics» KDD 2009

timeSVD++ (??)

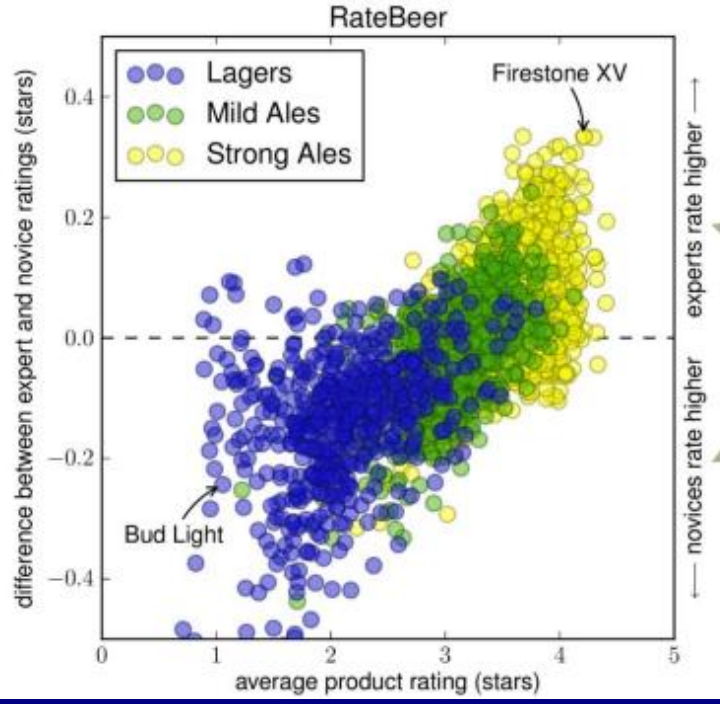


Люди склонны завышать рейтинги старых фильмов

есть много подобных эффектов – вывод: учитывайте время

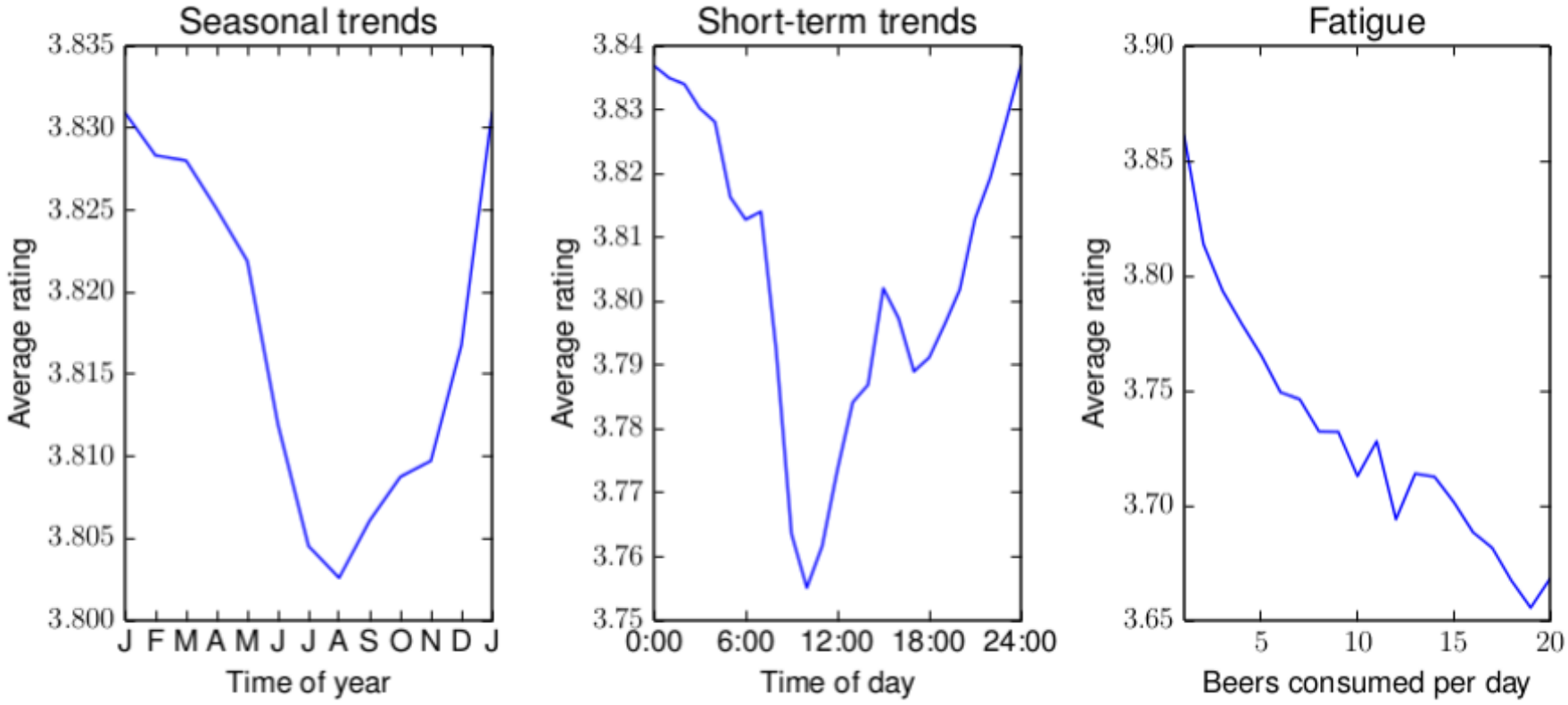
Что происходит со временем

- **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++ (??)

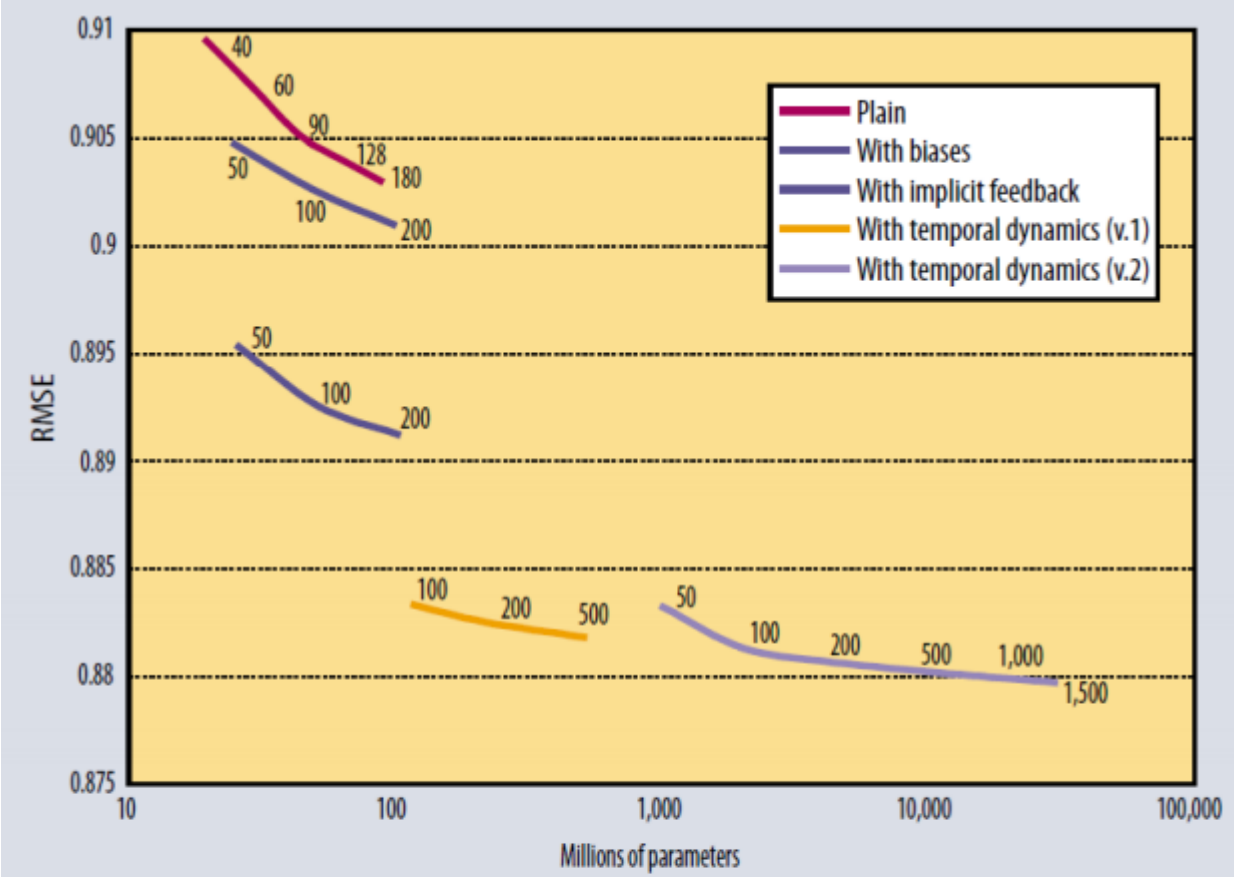


Figure 4. Matrix factorization models' accuracy. The plots show the root-mean-square error of each of four individual factor models (lower is better). Accuracy improves when the factor model's dimensionality (denoted by numbers on the charts) increases. In addition, the more refined factor models, whose 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.

timeSVD++ (??)

Регуляризация по времени

$$\ldots + \lambda \,||\, w(t) - w(t+\delta) \,||$$

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

Если есть «лайки» и «дизлайки»

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

Можно строить модель «один товар лучше другого»

$$P(i \succ j) = \sigma(w^{\mathrm{T}}\gamma_i - w^{\mathrm{T}}\gamma_j)$$

Стохастический градиентный спуск

~ случайно выцепляем пары сравнимых товаров

Коллаборативная фильтрация – минусы

- **проблема холодного старта (cold start)**

другая техника: по контенту, не персональные и т.п.

система рейтинга (обратная связь), костыли (по умолчанию)

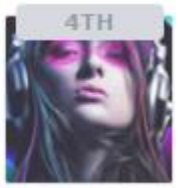
- **популярные становятся популярнее (popularity bias)**
- **условия шума (семейные аккаунты, случайные покупки и т.п.)**
- **возможны «атаки» на систему**

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

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

модель второго порядка:

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

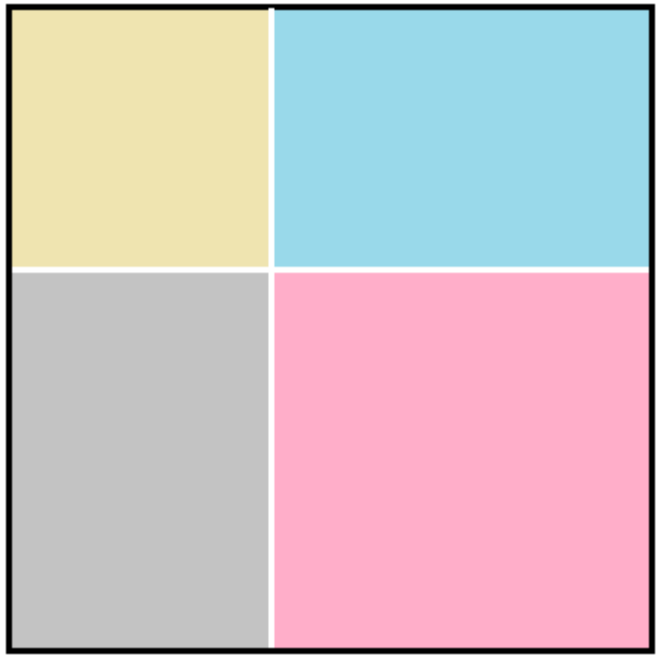
«факторизация» – в предположении, какая у нас матрица весов, иначе была бы просто «модель второго порядка»

Факторизационные машины

Что ещё…

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

FFM – field-aware factorization machine



Линейная модель

$$\phi(\mathbf{w},\mathbf{x}) = \mathbf{w}^T\mathbf{x} = \sum_{j\in C_1} w_j x_j$$

Полиномиальная модель (Poly2)

$$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2\in C_2} w_{j_1,j_2} x_{j_1} x_{j_2}$$

Факторизационная машина

$$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2\in C_2} \langle \mathbf{w}_{j_1}, \mathbf{w}_{j_2}\rangle x_{j_1} x_{j_2}$$

Факторизационная машина с полями

$$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2\in C_2} \langle \mathbf{w}_{j_1,f_2}, \mathbf{w}_{j_2,f_1}\rangle x_{j_1} x_{j_2}$$

Оптимизационная задача

$$\min_{\mathbf{w}} \sum_{i=1}^{L} \left(\log\left(1 + \exp(-y_i \phi(\mathbf{w}, \mathbf{x}_i)\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2 \right),$$

$$\phi(\mathbf{w}, \mathbf{x}) = \sum_{j_1, j_2 \in C_2} \langle \mathbf{w}_{j_1, f_2}, \mathbf{w}_{j_2, f_1} \rangle x_{j_1} x_{j_2},$$

LogLoss + регуляризация

Что такое поля...

Field name		Field index
User	$\rightarrow$	field 1
Movie	$\rightarrow$	field 2
Genre	$\rightarrow$	field 3
Price	$\rightarrow$	field 4

Что ещё?

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

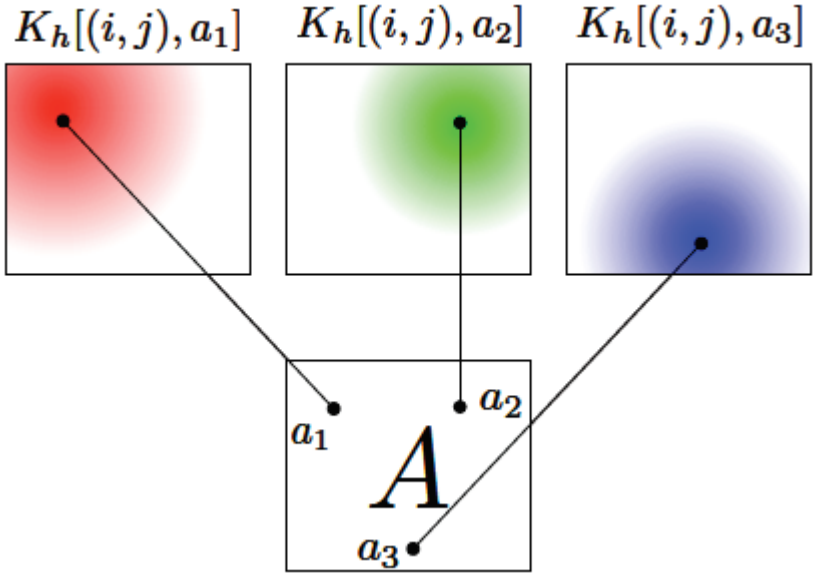
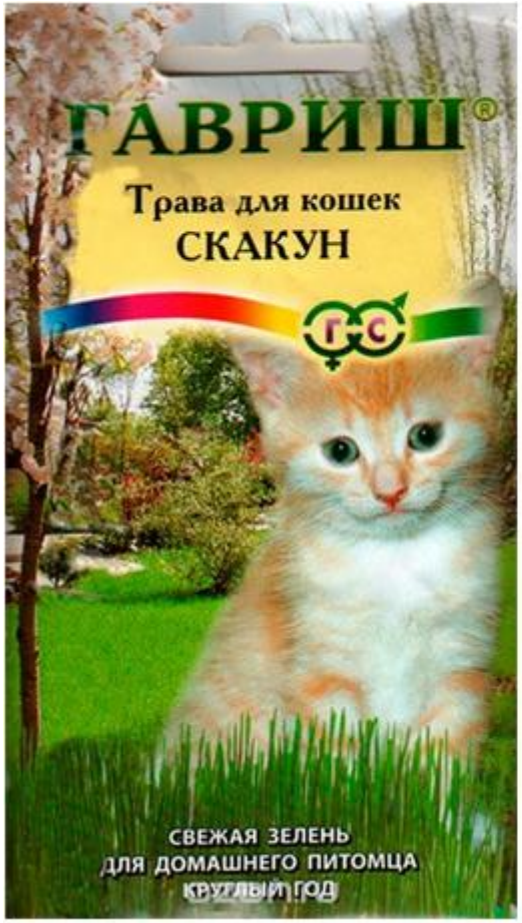


рис. из дипломной работы М.Трофимова

Простые методы



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 ₽

Добавить в корзину

Вместе с этим товаром покупают


Трава для кошек Скакун, 10 г

+


Фикус Притупленный, 3 шт.

+


Нолина (бокарнея отогнутая) Бутылочное дерево, 3 шт.

= 85 ₽

В корзину

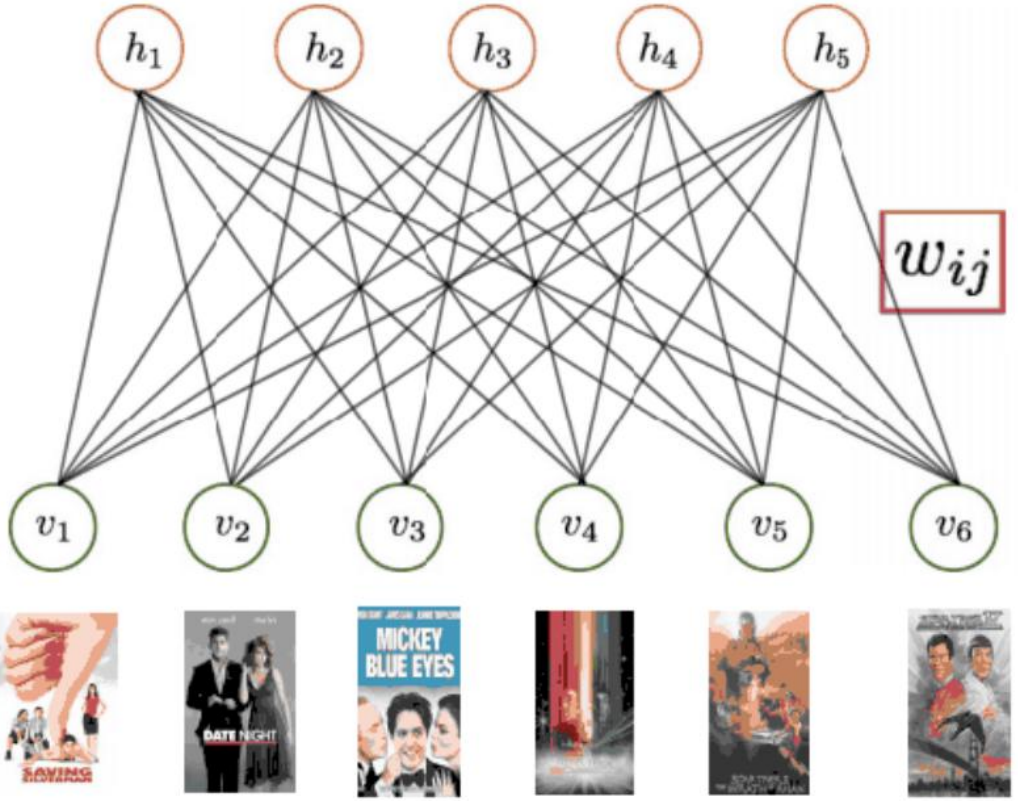
Бандлы ~ по статистике

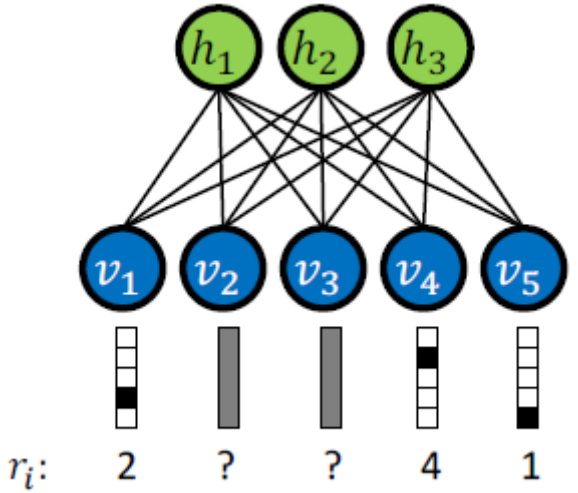
Простые методы

Ассоциативные правила

Кластеризация пользователей / товаров
(+ стандартные рекомендации)
есть и автоматические кластеры
(интересы, любимые театры / жанры, актёры и т.п.)

RBM


h_1
h_2
h_3
h_4
h_5
w_{ij}
v_1
v_2
v_3
v_4
v_5
v_6
MICKEY BLUE EYES
DATE NIGHT

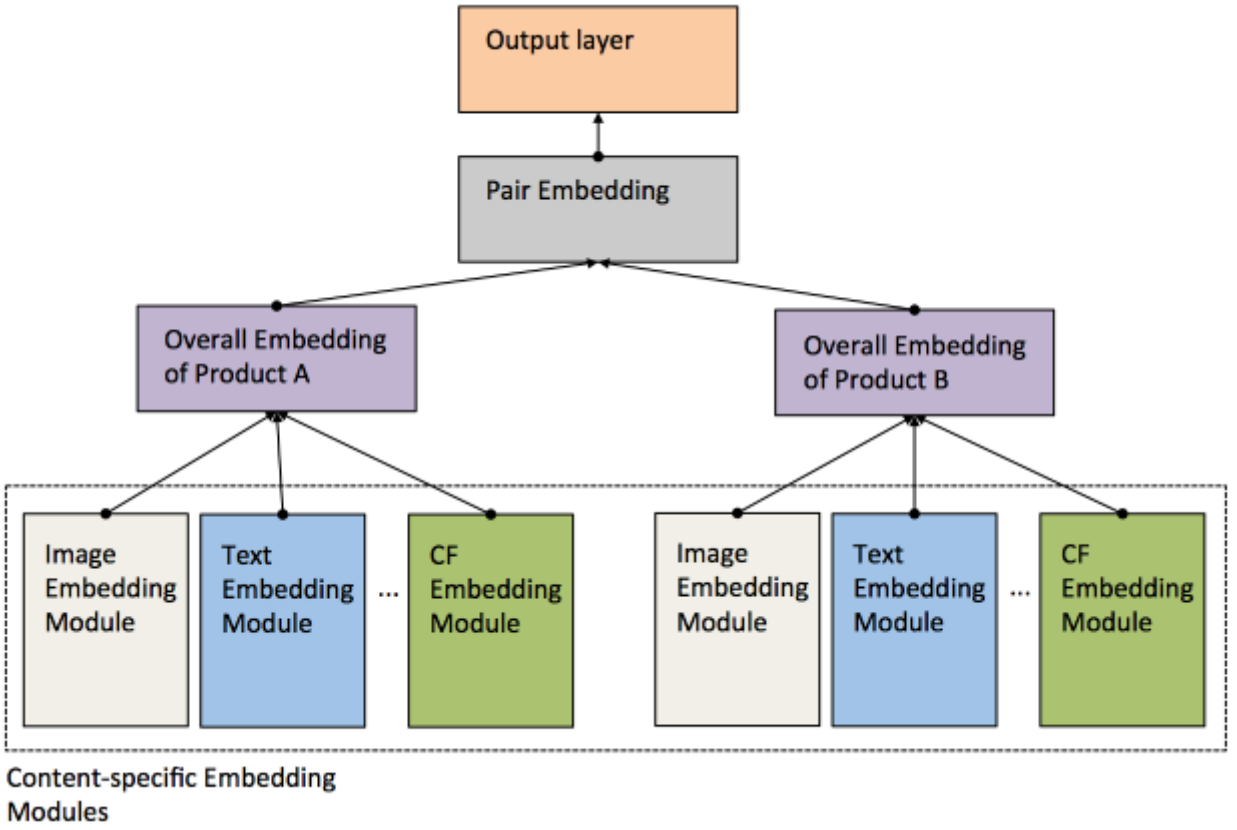

h_1
h_2
h_3
v_1
v_2
v_3
v_4
v_5
r_i: 2 ? ? 4 1

Spotify

Рекуррентные нейронные сети для предсказания последовательности треков

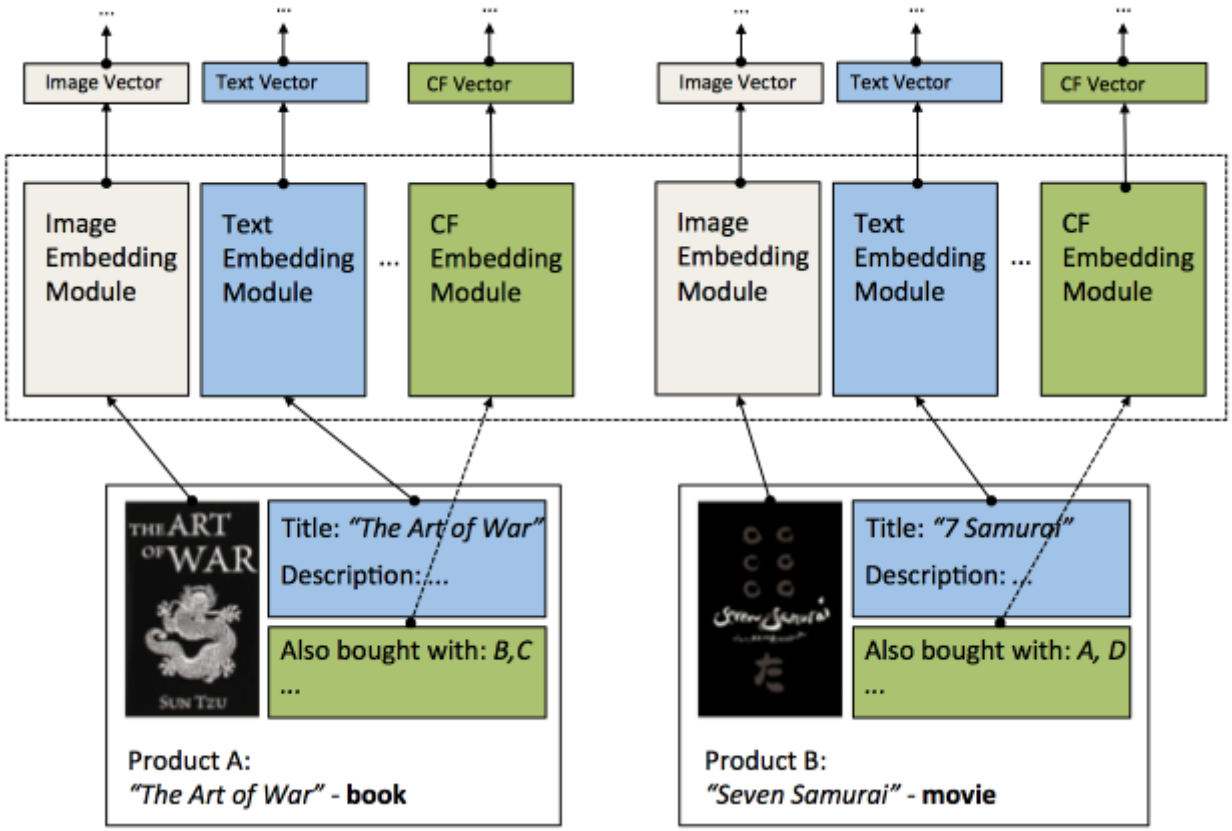
CONTENT2VEC

Thomas Nedelec, Elena Smirnova, Flavian Vasile Content2vec: specializing joint representations of product images and text for the task of product recommendation // https://openreview.net/pdf?id=ryTYxh5ll



Как вычислять расстояния между продуктами

CONTENT2VEC



Использование RNN

Сессия – последовательность товаров – предсказываем следующий

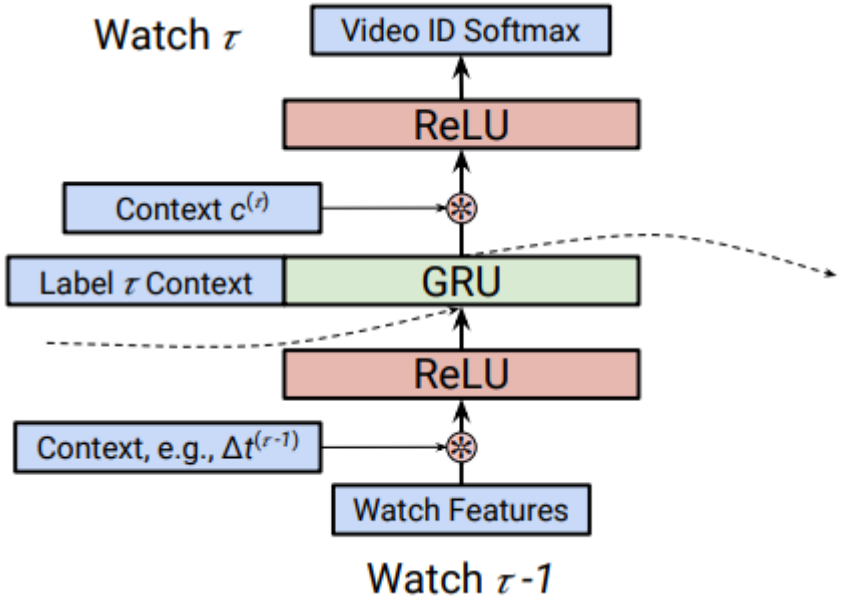
Balázs Hidasi, Alexandros Karatzoglou
«Recurrent Neural Networks with Top-k Gains for Session-based Recommendations» https://arxiv.org/pdf/1706.03847.pdf

Hidasi et al. Session-based Recommendations with Recurrent Neural Networks.

Использование RNN

Contextual data in neural recommender systems

добавляем контекст!



Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Dawei Yin, Yihong Zhao, Jiliang Tang Deep Reinforcement Learning for List-wise Recommendations https://arxiv.org/pdf/1801.00209.pdf

Вложения

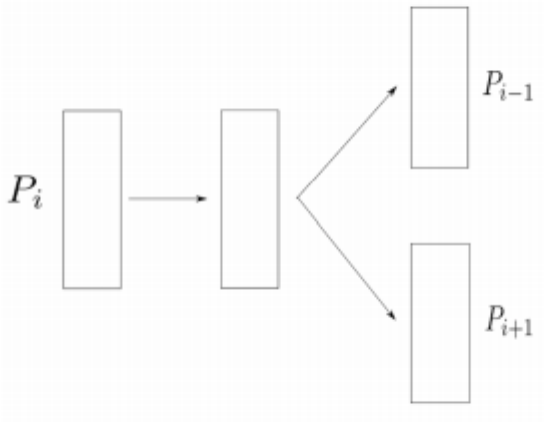


Figure 1: Prod2Vec Neural Net Architecture.

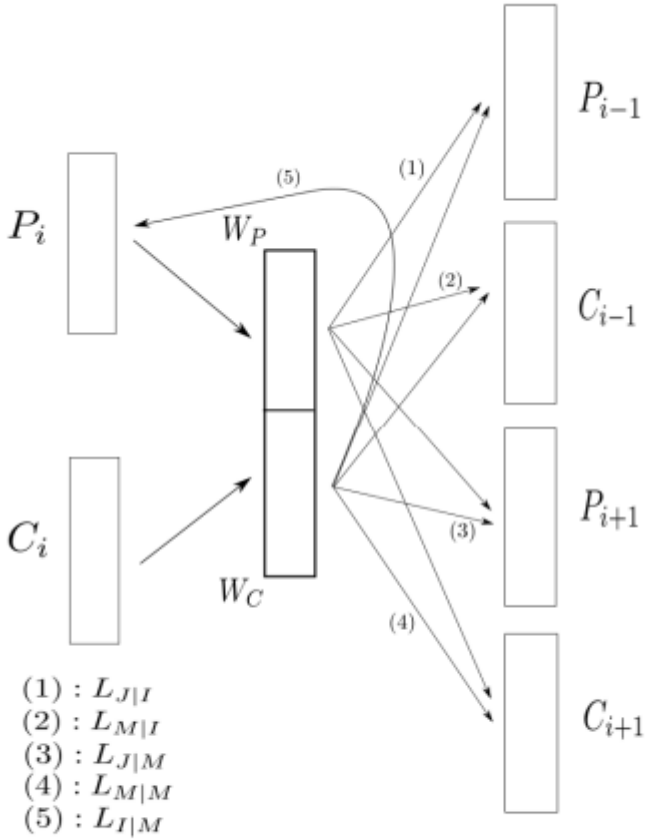


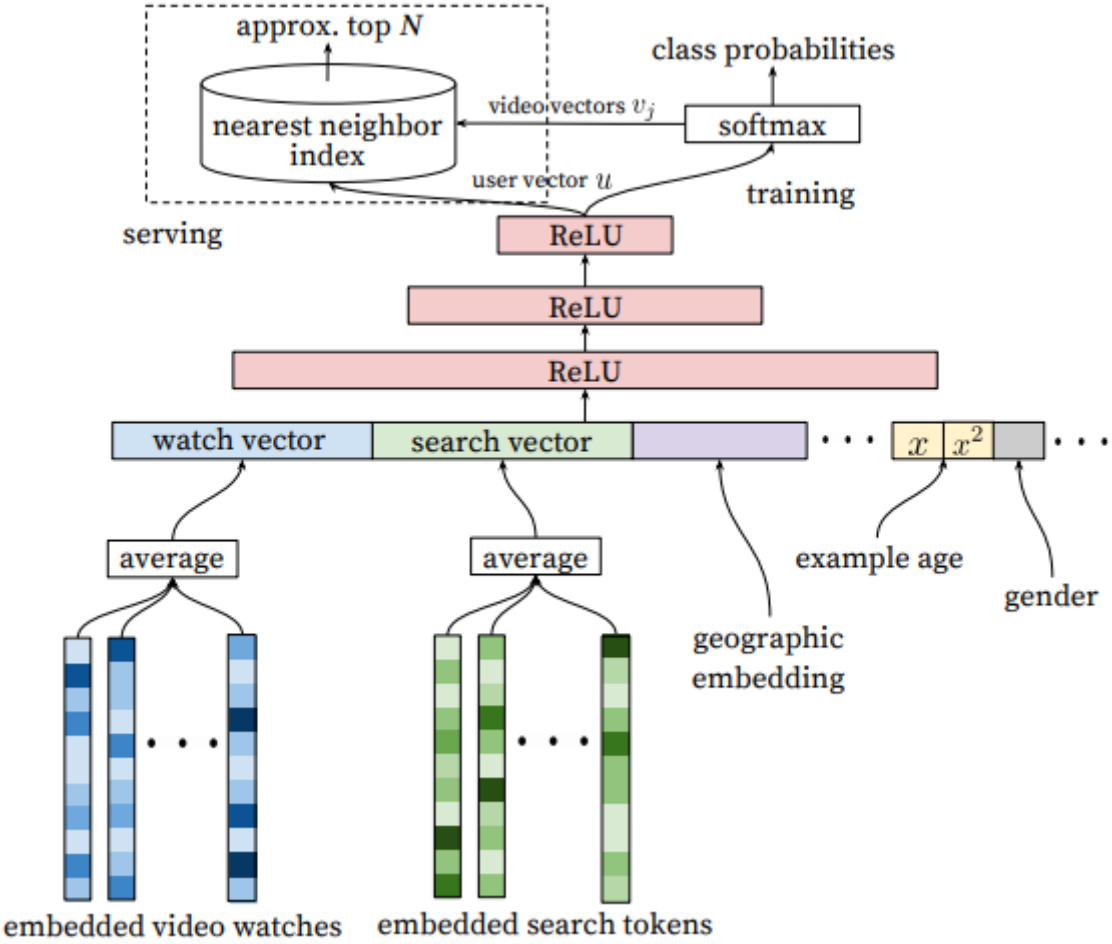
Figure 2: Meta-Prod2Vec Neural Net Architecture.

Flavian Vasile, Elena Smirnova, Alexis Conneau «Meta-Prod2Vec - Product Embeddings Using Side-Information for Recommendation»

Использование DL

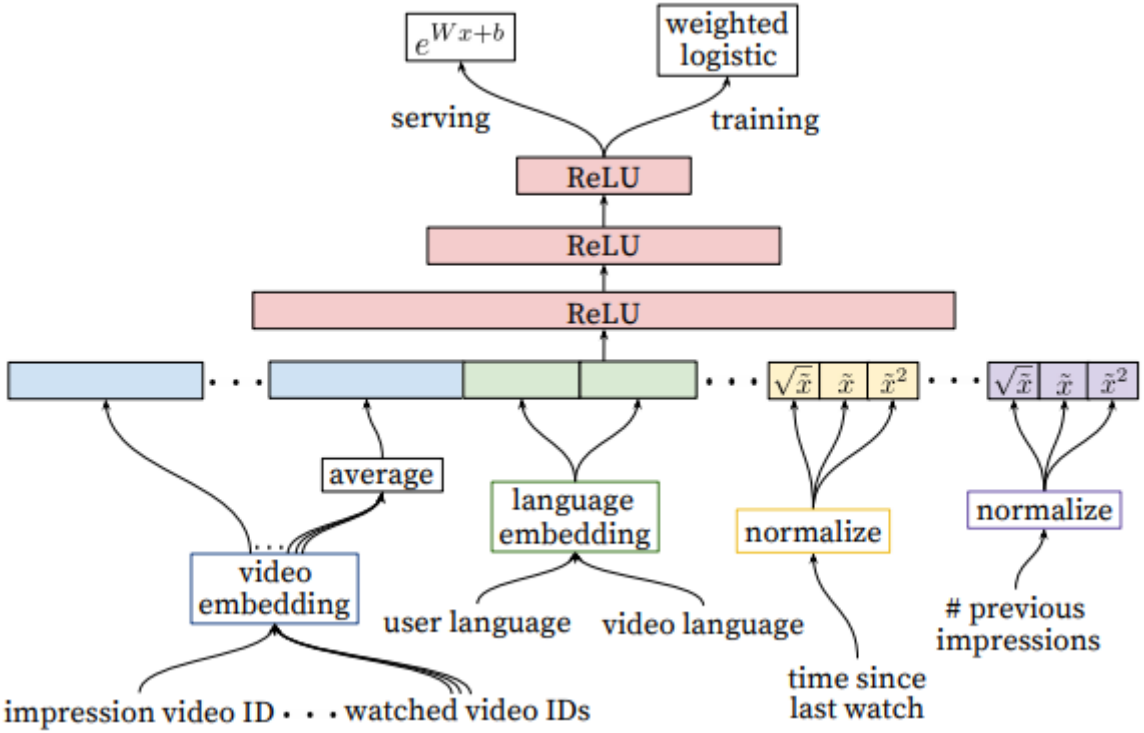
Paul Covington, Jay Adams, Emre Sargin «Deep Neural Networks for YouTube Recommendations» // https://cseweb.ucsd.edu/classes/fa17/cse291-b/reading/p191-covington.pdf

Сначала генерируем кандидатов:



Использование DL

Потом их ранжируем – оцениваем «impressions»



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.)

Контекст

канал захода / просмотра / покупки

состояние корзины / счёта / предыстория

география

время (года, суток)

погода и т.п.

Рекомендовать надо только то, что без рекомендации не купит…

Неидентифицируемые пользователи + новые товары

многорукие бандиты

тут RL

даже в неперсональных рекомендациях

исследование (exploration) – сбор статистики

использование (exploitation) – рекомендация топовых

Замечание

Тоже есть подготовка данных

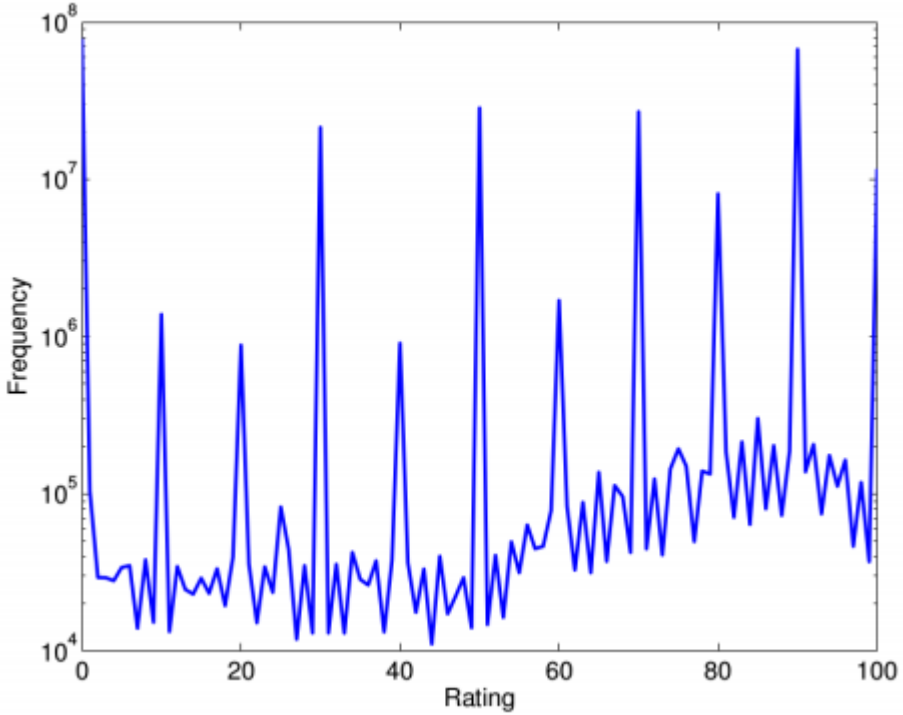
- **удаление выбросов**

слишком популярные товары /

активные пользователи (оптовики)

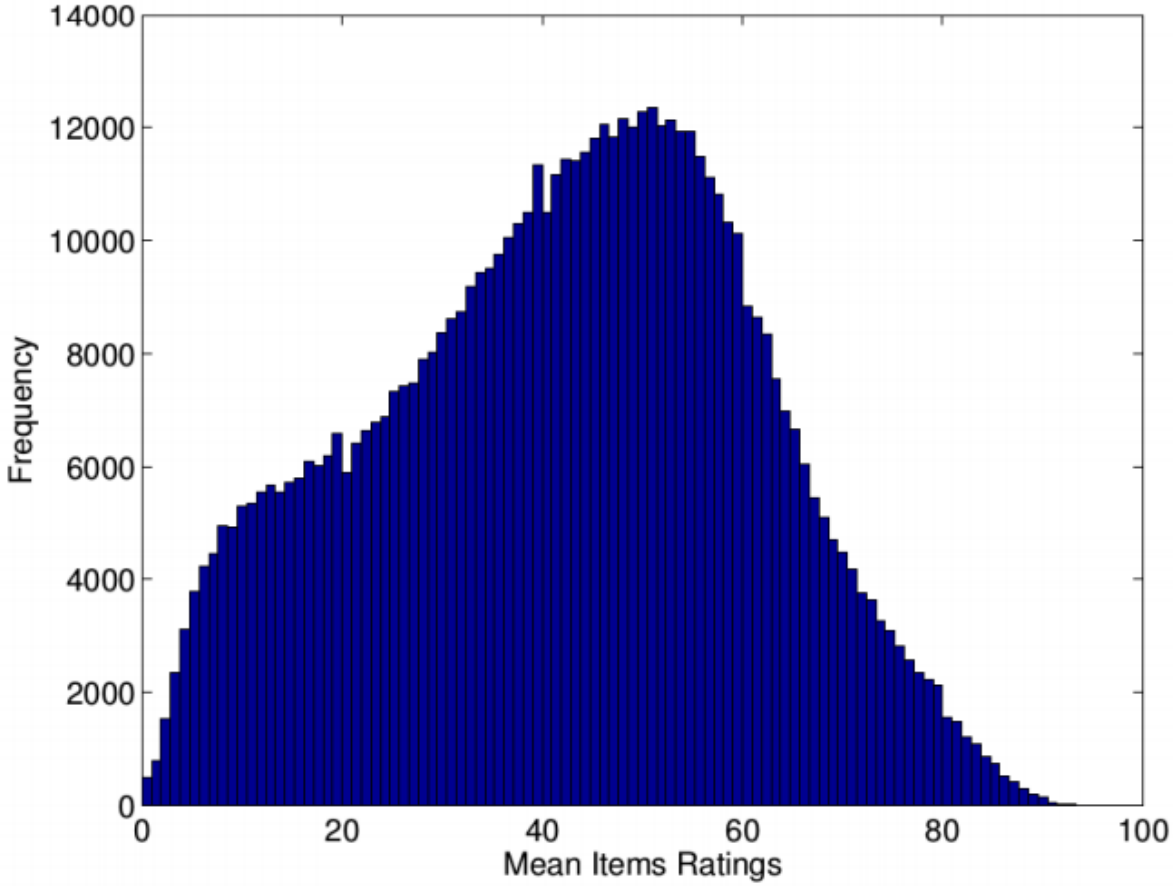
Немного о реакции пользователей

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



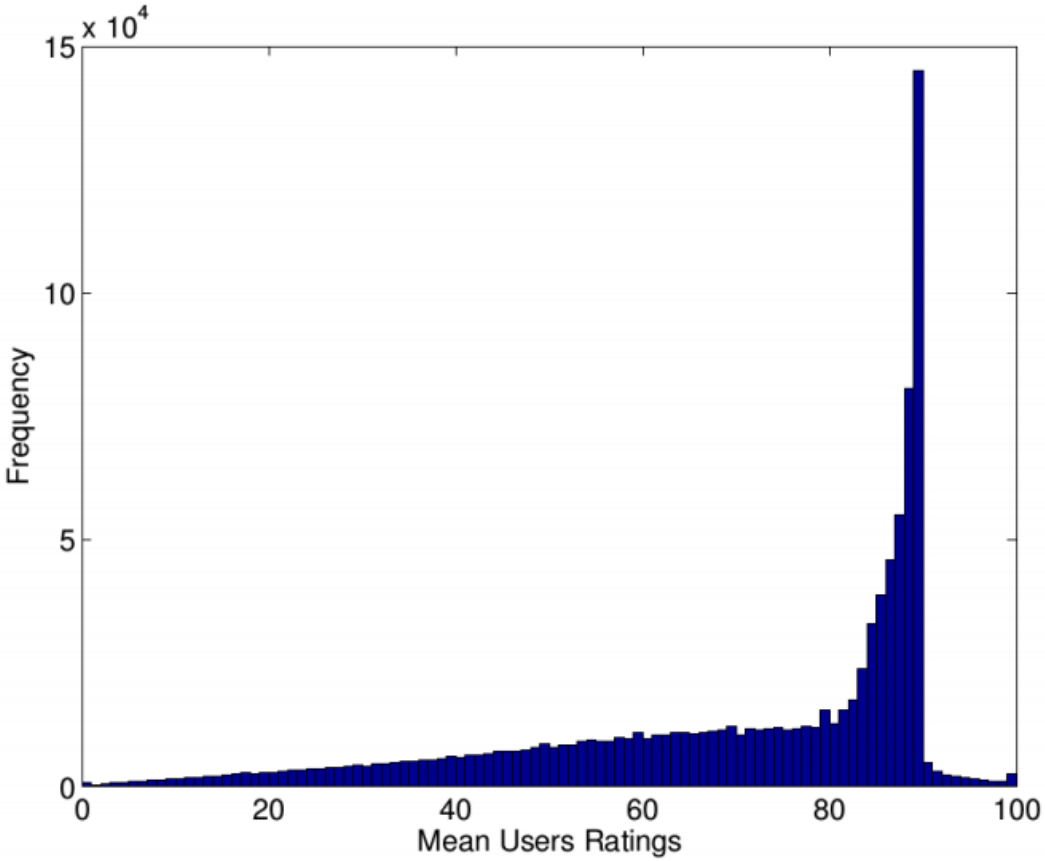
Распределение проставляемых рейтингов

Есть приложения, где только звёздочки

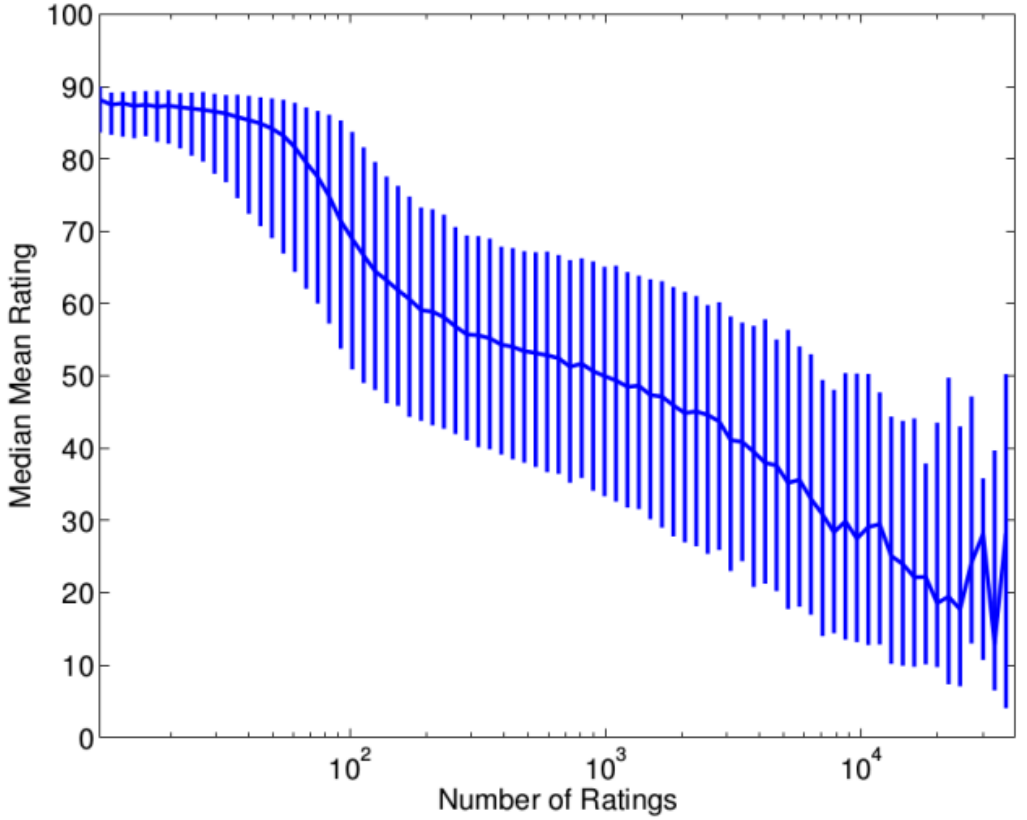


Распределение средних рейтингов композиций

Немного о реакции пользователей

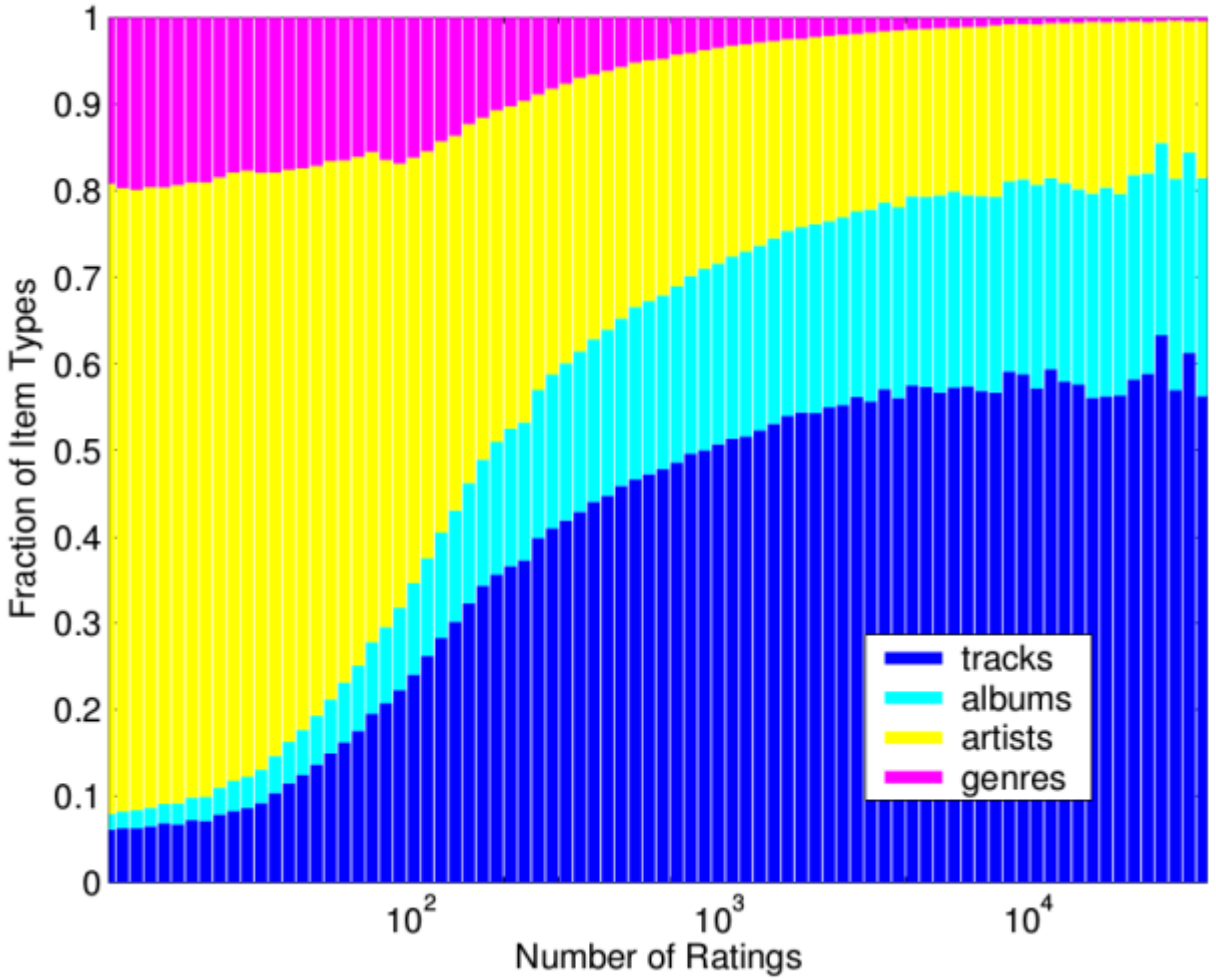


Распределение средних рейтингов пользователей



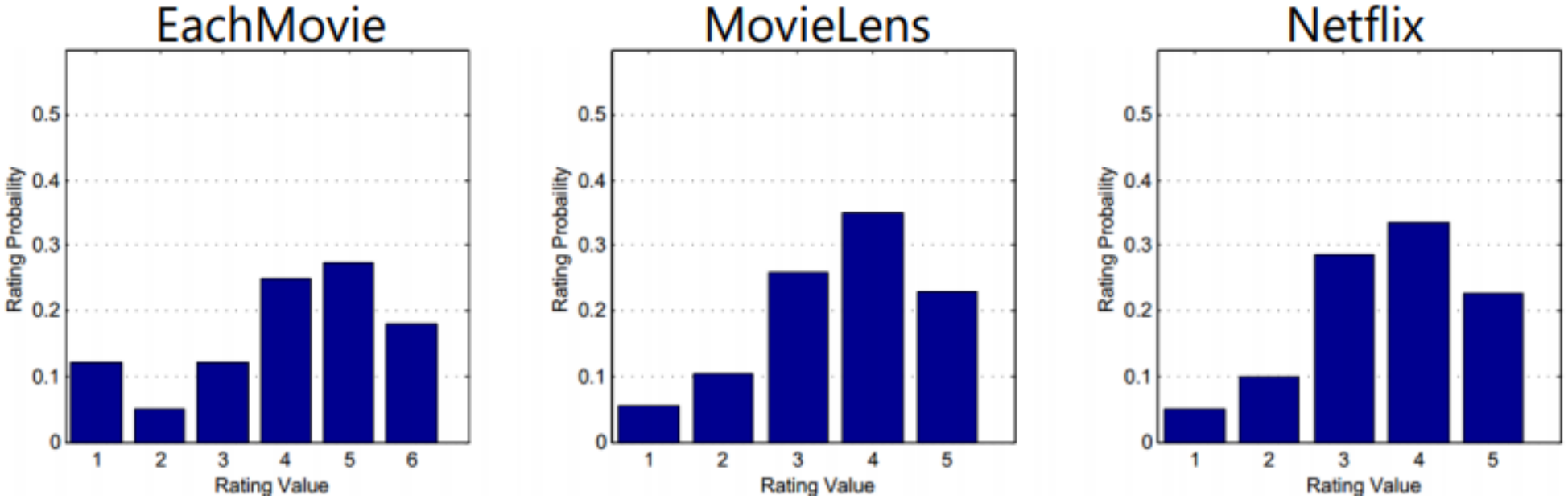
Кстати: 600 000 композиций, 1 000 000 пользователей 250 000 000 рейтингов

Немного о реакции пользователей



Немного о реакции пользователей

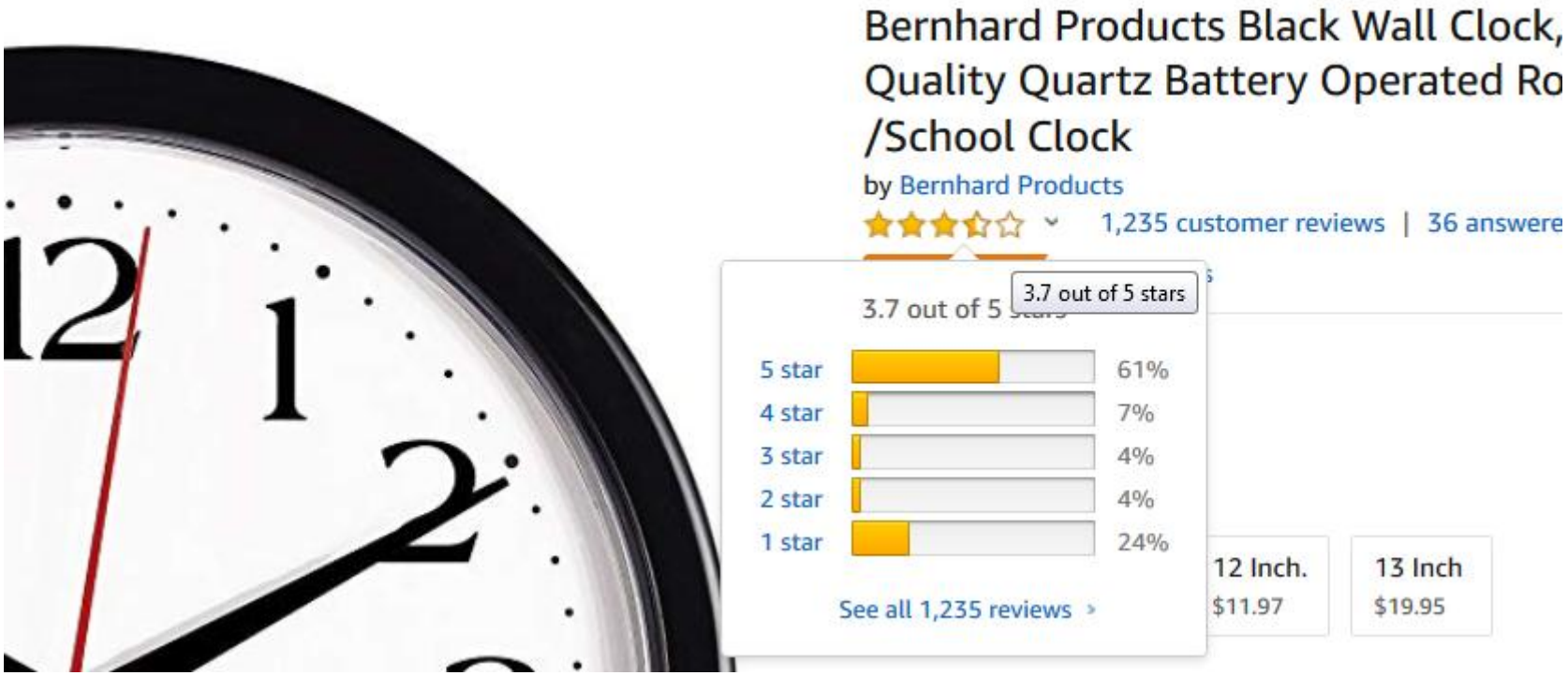
Рейтинг не репрезентативен



Marlin et al. «Collaborative Filtering and the Missing at Random Assumption» (UAI 2007)

Немного о реакции пользователей

Рейтинг не репрезентативен



Пользователи оставляют отзыв в специальных случаях

Knowledge-based Recommendations

девиз: «что удовлетворяет моим нуждам»

дорогие редко покупаемые нерейтингуемые товары

машины, квартиры, технологические продукты

требования / ограничения пользователя

«не очень дорого», «у метро», «безопасная»

CF – мало данных

CB – шумная похожесть

тут м.б. нечёткие множества

constraint-based

в явном виде определяем условия

case-based

сходство по условиям

«conversational» recommendations

уточнение в диалоге

Важность объяснений (explanations)

transparency, trustworthiness, validity, satisfaction

пользователям приятно работать с системой

убедительность (persuasiveness)

пользователи будут доверять

эффективность (effectiveness, efficiency)

пользователи быстрее принимают решение

обучение (education)

пользователи лучше понимают работу системы и учатся с ней взаимодействовать

A/B-тесты

Ron Kohavi «Seven rules of thumb for web site experimenters»

Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years// https://www.youtube.com/watch?v=qtboCGd_hTA

некоторые правила

- **Маленькие изменения могут иметь большие последствия**
- **Изменения редко бывают большими в положительную сторону**
- **Тестирование может затянуться**
- **Скорость важна**
- **Снижение числа отказов – сложно, смещение кликов – легко**
- **Избегайте сложностей – маленькие шаги**
- **Пользователей должно быть много**

Использование дополнительной информации

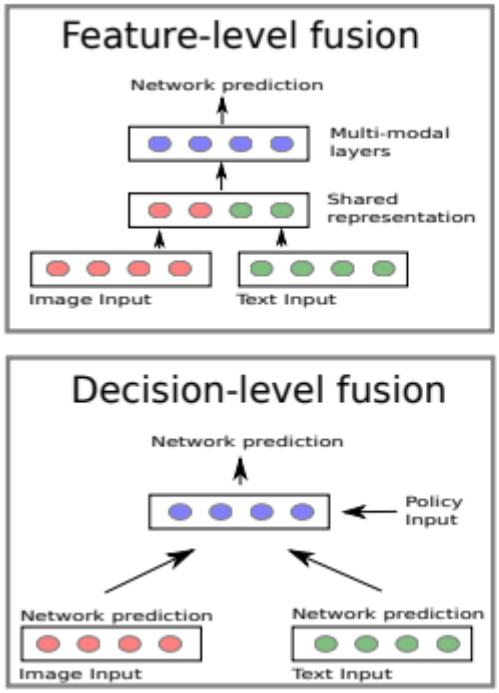
Informative title and image	Image is more informative	Title is more informative
 Title: Apple iPhone 6s Smartphone (Unlocked) Shelf: Iphone 6s	 Title: Hades VENTAIL-BRWN-11-VENTAIL - BRWN - Size - 11 Shelf: Women boots	 Title: California Umbrella 7.5' Market Patio Umbrella with Push Tilt in Straw Shelf: Patio Umbrellas
 Title: High Pressure Folding Table in Yellow (30 in. x 72 in./Yellow) Shelf: Folding Tables & Chairs	 Title: WEAR BOOTS Shelf: Baby & Toddler Underwear & Undershirts	 Title: Autumn Lip Pencil Colorganics .22 g Pencil Shelf: Moisturizers

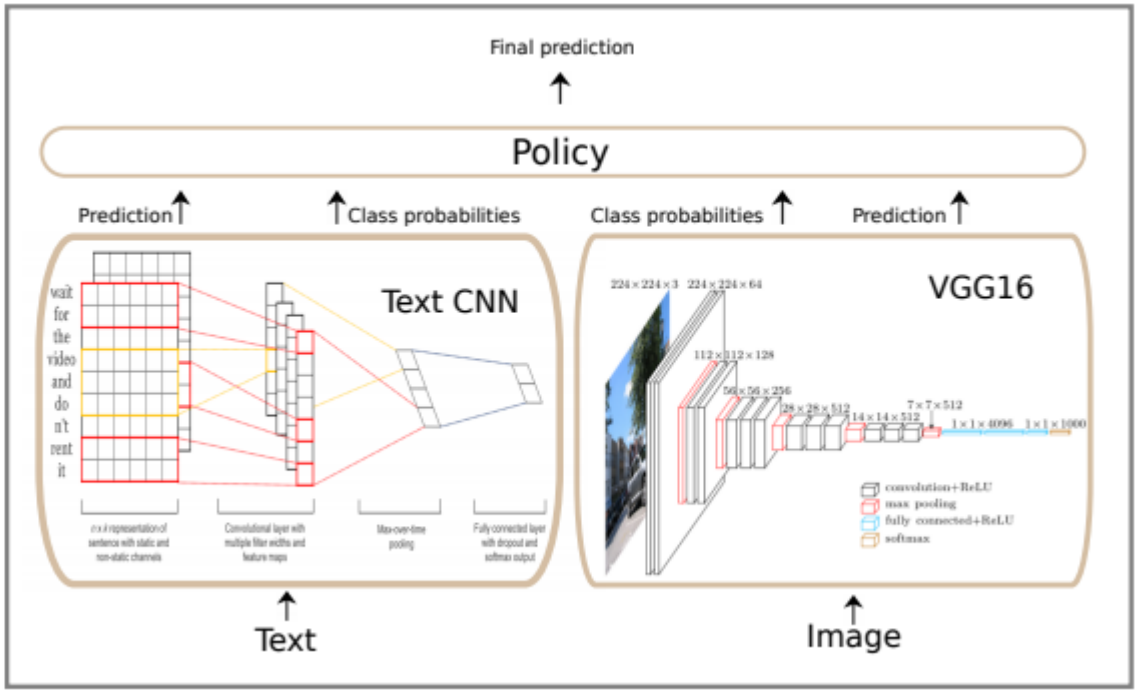
Tom Zahavy, Shie Mannor, Alessandro Magnani, Abhinandan Krishnan Is a picture worth a thousand words? A deep multi-modal fusion architecture for product classification in e-commerce

Резюме: текст информативнее

Использование дополнительной информации

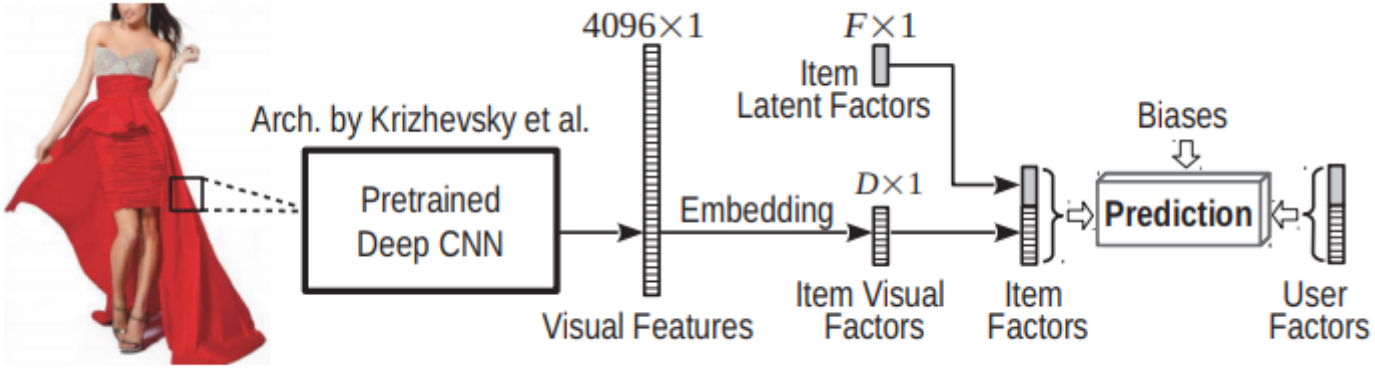
Как делают...



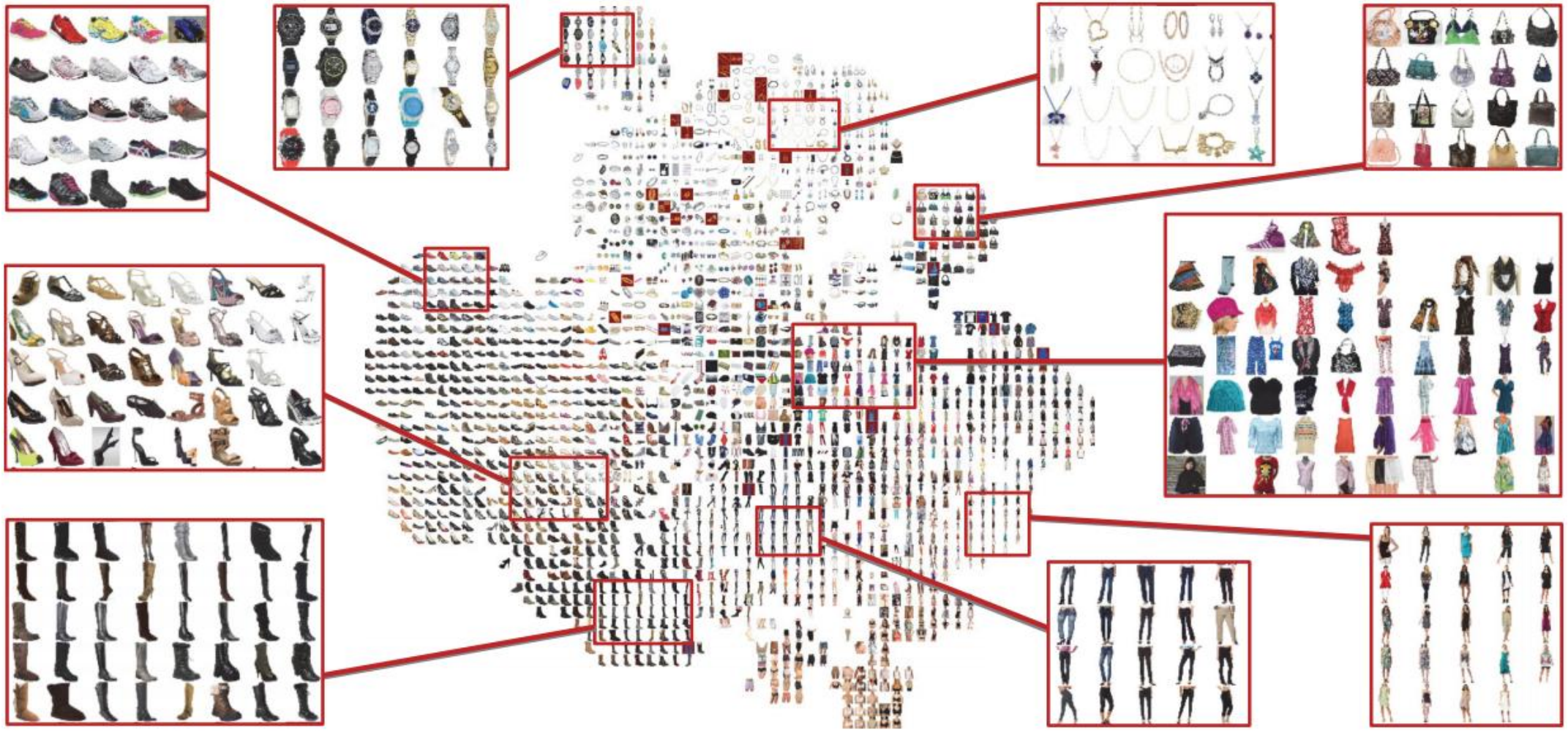


Использование дополнительной информации

Ruining He, Julian McAuley «VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback» //
https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/11914/11576



Использование дополнительной информации



Тренды

Вложения (embedding)

- A Deep Multimodal Approach for Cold-start Music Recommendation(Oramas et al.)
- Comparing Neural and Attractiveness-based Visual Features for Artwork Recommendation (Dominguez et al.)
- Translation based recommendation (He et al.)

используем контекст и отзывы

- Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction (Seo et al.)
- Recommendation of High Quality Representative Reviews in e-commerce (Paul et al.)

последовательные рекомендации

- Sequential User-based Recurrent Neural Network Recommendations (Donkers et al.)

CASE: LenKor

Пример решения (рекомендательных) задач методом ближайшего соседа

Задача «Predict Grant Applications»

Прогноз результата выполнения гранта

'Grant.Application.ID'
'Grant.Status'
'Sponsor.Code'
'Grant.Category.Code'
'Contract.Value.Band...see.note.A'
'Start.date'
'RFCD.Code.1'
'RFCD.Percentage.1'
'RFCD.Code.2'
'RFCD.Percentage.2'
'RFCD.Code.3'
'RFCD.Percentage.3'
'RFCD.Code.4'

'RFCD.Percentage.4'
'RFCD.Code.5'
'RFCD.Percentage.5'
'SEO.Code.1'
'SEO.Percentage.1'
'SEO.Code.2'
'SEO.Percentage.2'
'SEO.Code.3'
'SEO.Percentage.3'
'SEO.Code.4'
'SEO.Percentage.4'
'SEO.Code.5'
'SEO.Percentage.5'

'Person.ID.1'
'Role.1'
'Year.of.Birth.1'
'Country.of.Birth.1'
'Home.Language.1'
'Dept.No..1'
'Faculty.No..1' 'With.PHD.1'
'No..of.Years.in.Uni.at.Time.of.Grant.1'

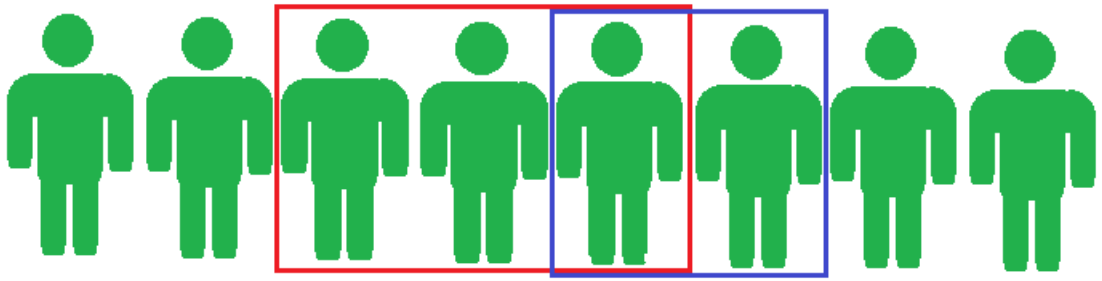
'Number.of.Successful.Grant.1'
'Number.of.Unsuccessful.Grant.1'
'A..1'
'A.1'
'B.1'
'C.1'

По описанию гранта ~ будет ли его выполнение успешным.

	Название	Область	Фин.	Коллектив	Статьи A
11	Топологические инварианты	021 – 100%	300.000	Иванов Пеший	10 3
12	Написание рекомендательной системы	217 – 60% 218 – 49%	550.000	Печенкин Белых Абашидзе	2 1 0

Технология LENKOR - именно для этой задачи и была разработана!

1. Не ясно, как измерить похожесть проектов
2. Но ясно, как измерить похожесть коллективов, спонсоров, областей и т.д.

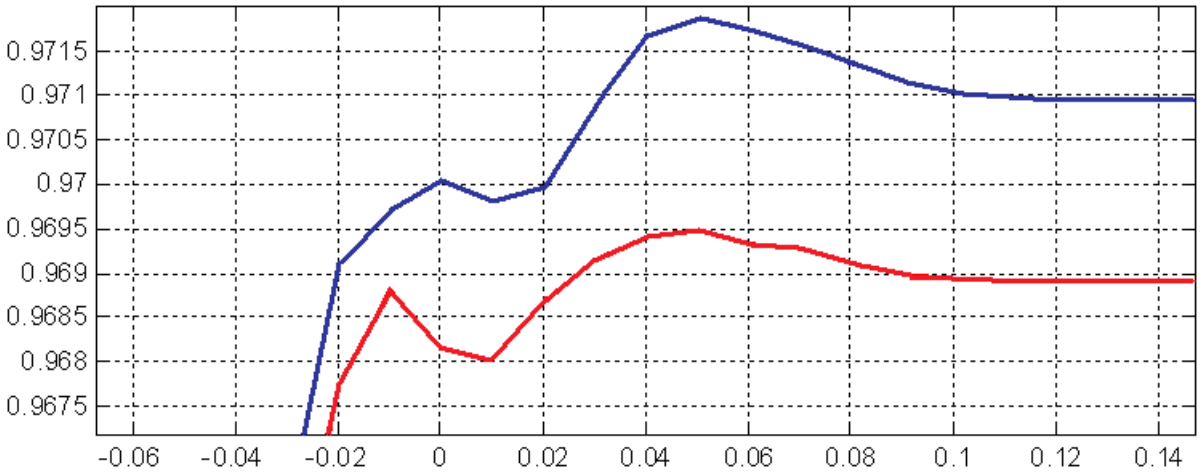


3. Близость (метрика) = сумма близостей (метрик)

$$B(x, x_i) = \sum_{\omega} c_\omega B_\omega (x, x_i)$$

4. Коэффициенты можно настраивать
5. Потом можно добавить нелинейность

Вариация коэффициента при фиксированных остальных



Добавление признака «язык» в линейную комбинацию

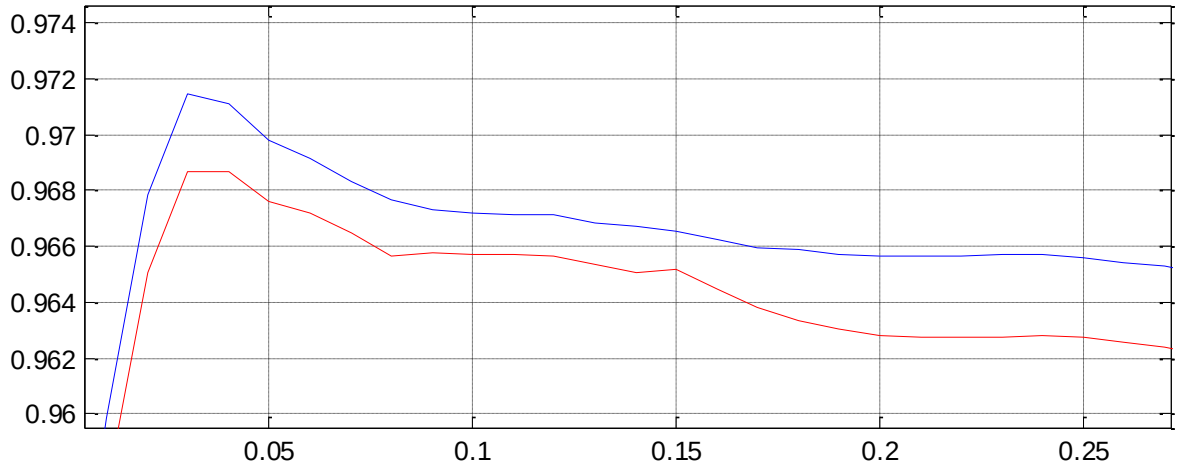
Сразу смотрим значение нужного функционала качества!

Настройка напоминает ту, что описана выше…

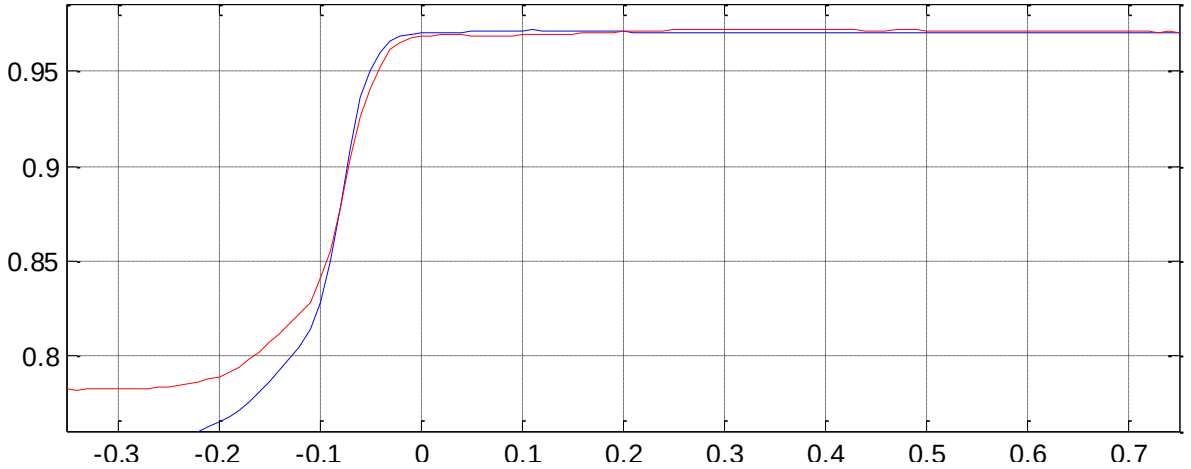
Применима во многих задачах…

Например, прогноз объёмов продаж в зависимости от рекламы.

Фиксация всех параметров, кроме одного (Grant)

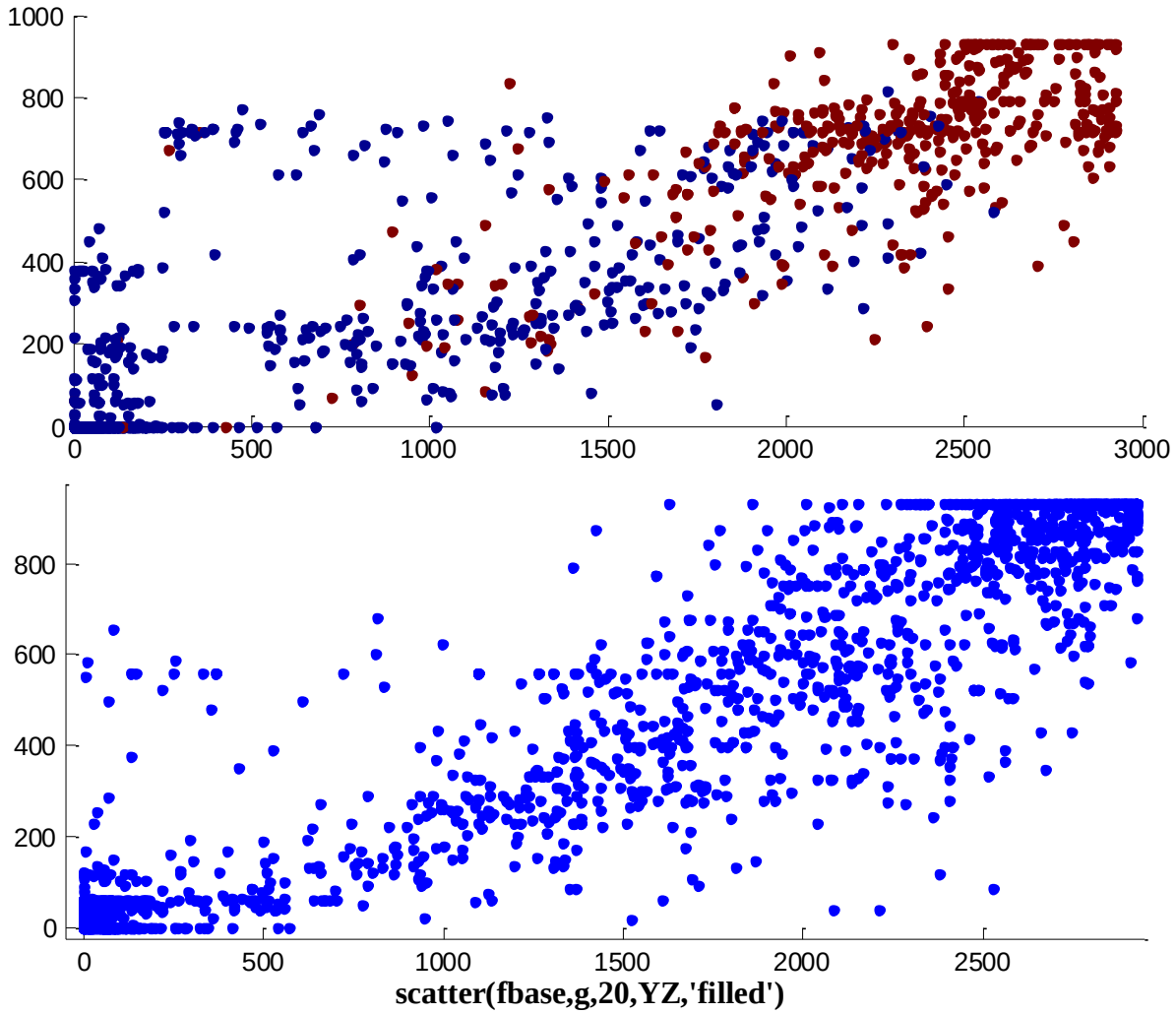


Можно улучшить (здесь: обучение и контроль)

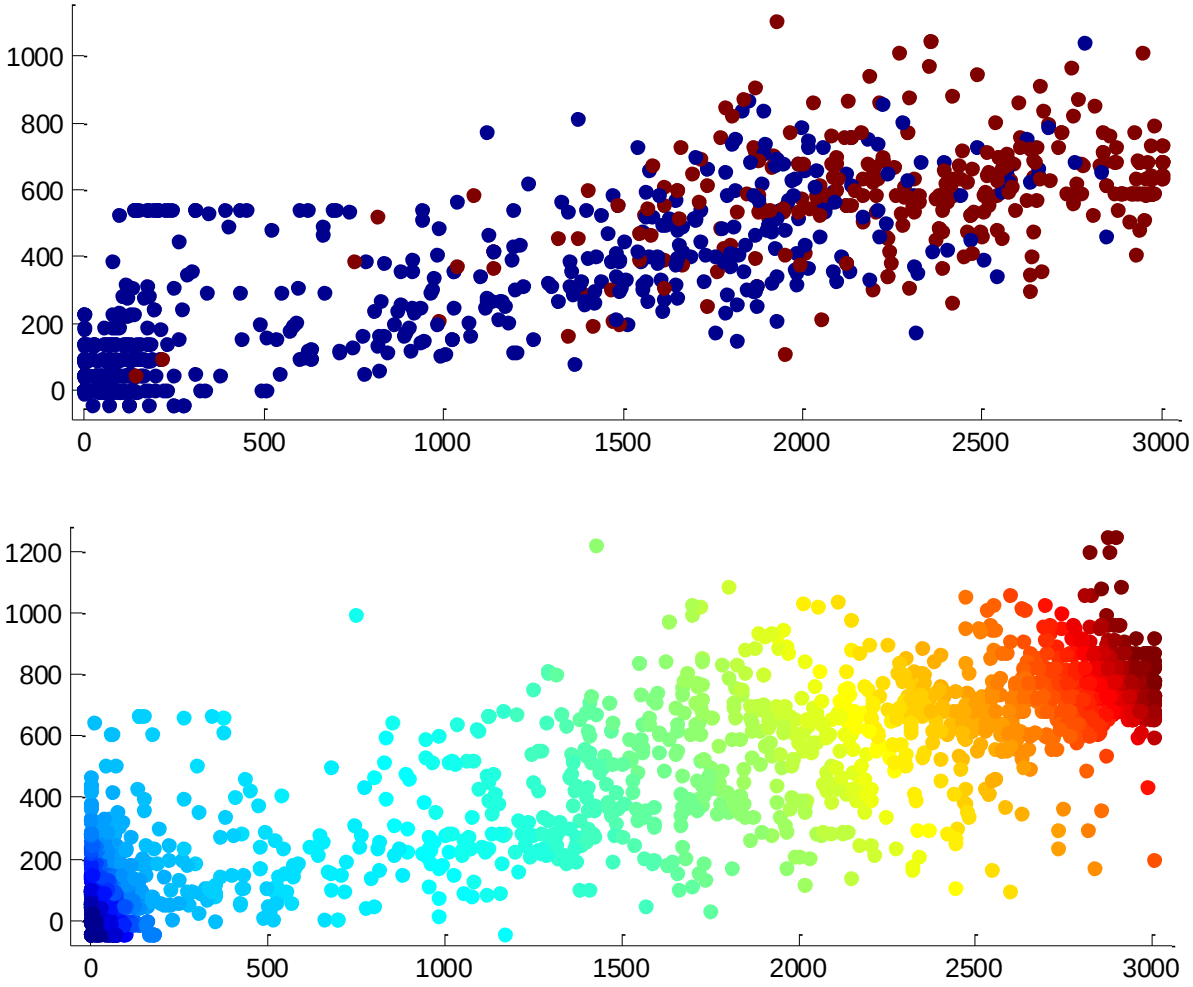


«Логичный» параметр

Кстати, ещё раз «как выглядят ответы»



Что послал (Grant)



«VideoLectures.Net Recommender System Challenge» (ECML/PKDD Discovery Challenge 2011)

Рекомендация лекций для просмотра

Задание соревнования – написать рекомендательную систему для ресурса VideoLectures.net

Первый подконкурс (cold start) – необходимо по **одной** просмотренной лекции рекомендовать лекции из множества «**новых** лекций», которые были недавно выложены на сайт, и для них ещё нет статистики популярности, только подробное описание.

N. Antulov-Fantulin, M. Bošnjak, T. Šmuc, M. Jermol, M. Žnidaršič, M. Grčar, P. Keše, N. Lavrač, ECML/PKDD 2011 - Discovery challenge: "VideoLectures.Net Recommender System Challenge", http://tunedit.org/challenge/VLNetChallenge/

Описание лекции

101, 'Lecture', 'eng', 'biology', '2008-12-04', '2009-02-12', 'Implementing a common framework on business', 'Professor Rudolf Smith', ...

Функционал качества

$$\frac{1}{|Z|}\sum_{z\in Z}\frac{|\{r_1,\ldots,r_{\min(S,R,z)}\}\cap\{s_1,\ldots,s_{\min(S,R,z)}\}|}{\min(S,R,z)}$$

$r_1,\ldots,r_R$ – **рекомендации**

$s_1,\ldots,s_S$ – **правильные ответы**

$Z=\{5,10,15,20,25,30\}$

Надо рекомендовать 30 лекций.

Нахождение метрики

$$\rho(\text{Lecture}_1,\text{Lecture}_2)=c_1\cdot\rho_1(\text{Author}_1,\text{Author}_2)+c_2\cdot\rho_2(\text{Title}_1,\text{Title}_2)+\ldots$$
$$+c_n\cdot\rho_n(\text{Subject}_1,\text{Subject}_2)$$

Не обязательно метрики по непересекающимся описаниям

Название

Аннотация

Текст

Название + Аннотация

Название + Аннотация + Текст

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

A. D'yakonov Two Recommendation Algorithms Based on Deformed Linear Combinations // Proc. of ECML-PKDD 2011 Discovery Challenge Workshop, pp 21-28 (2011).

Что использовано в решении

Близость двух лекций оценивалась используя только

1. Близость категорий.

2. Близость авторов.

3. Близость языков.

4. Близость названий.

5. Близость названий, описаний, названий и описаний событий.

+ статистика

m_{ij} **– число пользователей: смотрели и** i**-ю лекцию и** j**-ю лекцию**

Не использовано

Аналогичные данные по событиям (конференции, на которых они прочитаны, школы-семинары, циклы лекций и т.д.)

Описания слайдов

Даты съёмок

Обработка текста

Использовалось приведение к общей основной форме (стеммер Портера)

Нестандартное TF-IDF-преобразование – изменение качества на 5%.

Porter, 1980, An algorithm for suffix stripping, Program, Vol. 14, № 3, pp. 130–137.

http://tartarus.org/~martin/PorterStemmer/

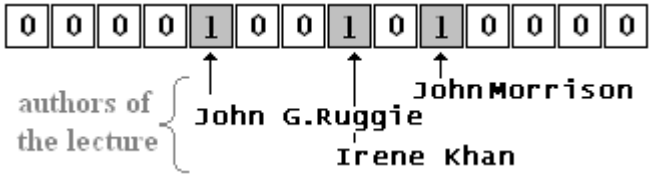
Как строилась метрика

n – число авторов всех лекций,

$f(i) = (f_1(i), \ldots, f_n(i))$ – бинарный вектор, описывает авторов i-й лекции: единицы помечают номера соответствующих авторов.

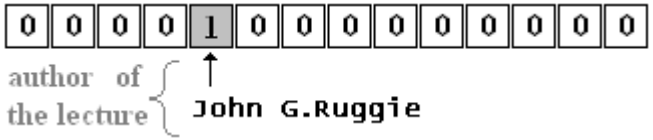
```
7445,'debate','en',7436,25,'2008-12-04','2009-02-11','Questions, Reactions from the audience',
NULL, NULL
```

$f(i) = (f_1(i), \ldots, f_n(i))$:




```
7442,'lecture','en',7436,112,'2008-12-04','2009-02-11','Implementing a common framework on
business and human rights','Professor John Ruggie, UN Special Representative on Business and
Human Right\nIrene Khan, Secretary General, Amnesty International\nModerated by: John
Morrison, Programme Director, Business Leaders Initiative on Human Rights\n(BLIHR)',NULL
```

$f(j) = (f_1(j), \ldots, f_n(j))$:



Измененная косинусная мера

$$\langle f(i), f(j)\rangle = \frac{f_1(i)f_1(j)+\cdots+f_n(i)f_n(j)}{\sqrt{f_1(i)^2+\cdots+f_n(i)^2+\varepsilon}\sqrt{f_1(j)^2+\cdots+f_n(j)^2+\varepsilon}}$$

$$\gamma(i,j) = \left\langle \sum_{t\in I}\left(\frac{m_{it}}{\sum_{s=1}^{L} m_{is}} \cdot \frac{f(t)}{\sqrt{f_1(t)^2+\ldots+f_n(t)^2+\varepsilon}} \right), f(j) \right\rangle$$

from the co-view statistics

– близость между новой j-й лекции и старой i-й лекцией (точнее, похожими на неё старыми лекциями с точки зрения пользователей)

I – множество индексов «старых лекций»,

m_{ij} – число пользователей, которые просмотрели и i-ю лекцию и j-ю лекцию при $i \neq j$, и m_{ii} – число пользователей, которые просмотрели i-ю, делённое пополам

Почему: пользователь посмотрел раздел «Биология», а новых лекций в нём нет… нечего рекомендовать?!

Окончательное решение

Вычисляем близость по формуле:

$$\gamma = 0.19 \cdot \sqrt{0.6 \cdot \gamma_{cat} + 5.6 \cdot \gamma_{auth}} + \sqrt{4.5 \cdot \gamma_{lang} + 5.8 \cdot \gamma_{dic} + 3.1 \cdot \gamma_{dic2}}\,.$$

Получено перебором различных форм ответа:

$$\gamma = C_1 \cdot \gamma_{cat} + C_2 \cdot \gamma_{auth} + C_3 \cdot \gamma_{lang} + C_4 \cdot \gamma_{dic} + C_5 \cdot \gamma_{dic2}$$

$$\gamma = C_1 \cdot \gamma_{cat} + \sqrt{C_2 \cdot \gamma_{auth} + C_3 \cdot \gamma_{lang}} + C_4 \cdot \gamma_{dic} + C_5 \cdot \gamma_{dic2}$$

$$\gamma = C_1 \cdot \gamma_{cat} + C_2 \cdot \gamma_{auth} + (C_3 \cdot \gamma_{lang} + C_4 \cdot \gamma_{dic})^2 + C_5 \cdot \gamma_{dic2}$$

$$\gamma = (C_1 \cdot \gamma_{cat} + C_2 \cdot \gamma_{auth}) + C_3 \cdot \gamma_{lang} + \sqrt{C_4 \cdot \gamma_{dic} + C_5 \cdot \gamma_{dic2}}$$

При решении задачи оптимизации использовался метод покоординатного спуска

Рекомендуем 20 лекций с наибольшими значениями γ.

Окончательное решение

Решение, выложенное на сайте

$$\left(\gamma_1 \cdot \left(1.07 - 0.07\frac{t_1 - t_{\min}}{t_{\max} - t_{\min}}\right), \ldots, \gamma_N \cdot \left(1.07 - 0.07\frac{t_N - t_{\min}}{t_{\max} - t_{\min}}\right)\right),$$

Rank	Team	Time of Submission	Preliminary Result	Final Result
1	+ D'yakonov Alexander	Jul 08, 09:27:22	0.37281	0.35857
2	+ lefman	Jul 07, 00:24:15	0.31063	0.30743
3	+ Nitram	Jul 08, 06:47:32	0.30661	0.27684
4	sofos	Jul 06, 23:22:55	0.27433	0.27151
5	+ Inner Peace	Jul 08, 10:52:35	0.27268	0.25773
6	+ DMIR	Jul 08, 11:39:37	0.26813	0.25498
7	+ ddi	Jul 08, 11:07:37	0.26298	0.24920
8	+ Haibin Liu	Jul 08, 08:45:05	0.25172	0.24559
9	+ tao	Jul 08, 09:50:24	0.22465	0.24044
10	+ Team SIG	Jul 08, 10:25:03	0.22465	0.24044

t_j – **время выкладывания на сайт j-й новой лекции,**

$t_{\min}$ – **минимальное t_j,**

$t_{\max}$ – **максимальное (вычислялось в днях).**

Увеличение качества примерно на 5%.

История одного тестирования

Бандл – множество товаров, которые покупают вместе…

Примеры

Крупная компания для интернет магазина предложила рекомендательную систему

⇒ тестирование (A/B-тест)

Итог…

Стоимость последнего бандла ~ 70000 руб.

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

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