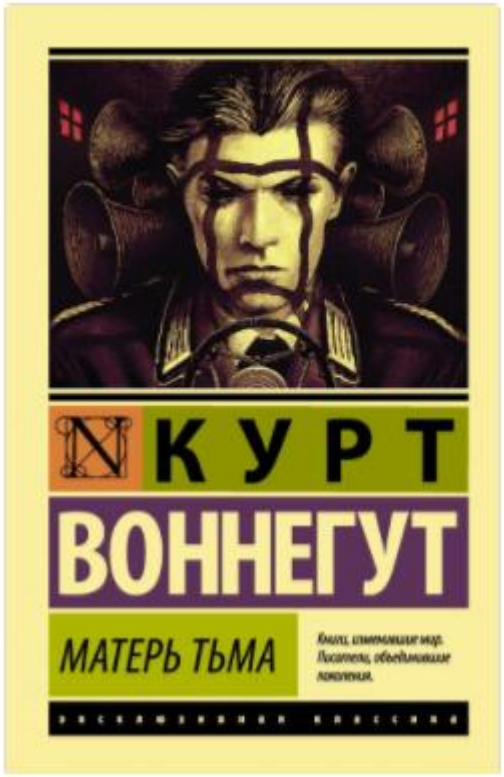
Машинное обучение и анализ данных

Введение в рекомендательные системы

Дьяконов А.Г.

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

Системы рекомендаций (Recommender Systems)



★★★★★ 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»

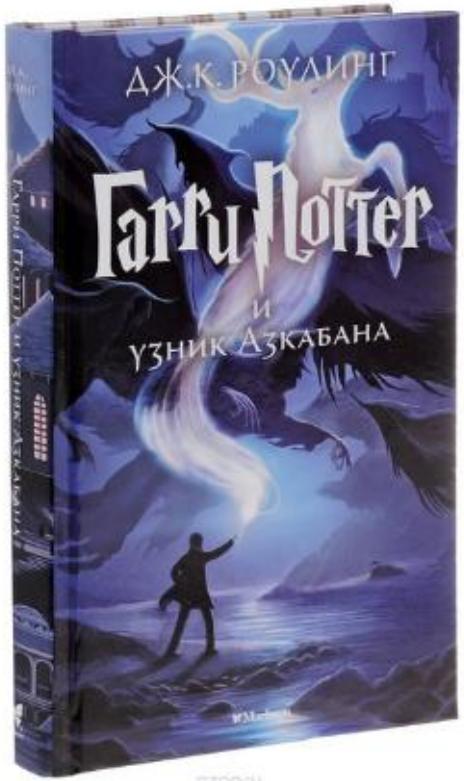
ISBN: 978-5-17-099474-8

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) от 414 ₽	Букинист. издание (2) от 2 469 ₽	Нет в продаже 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

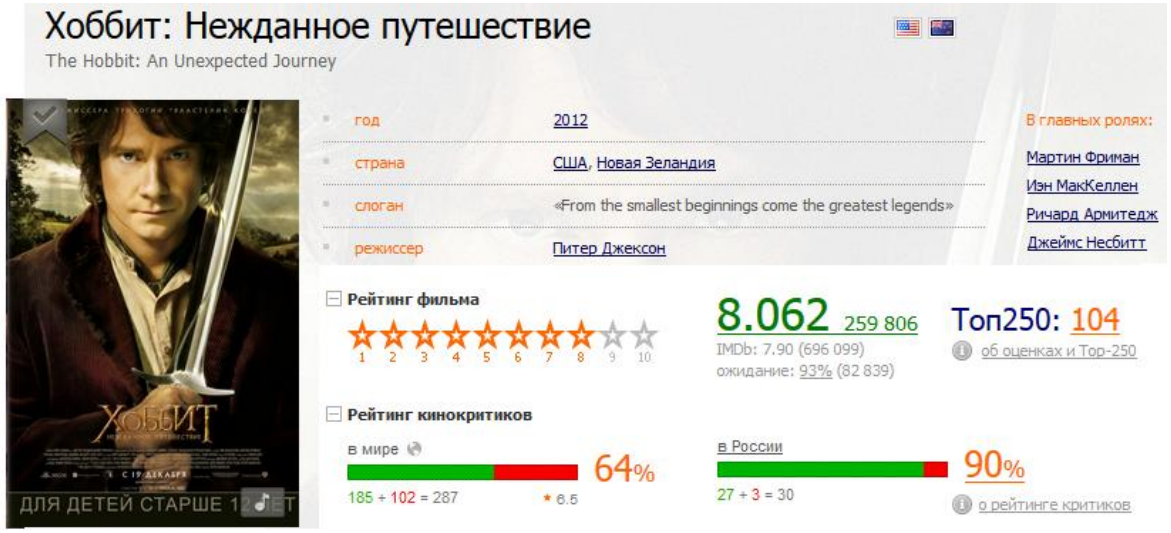
~~3 790 руб~~ 3 220 руб

Nike / Свитшот W NK TOP VERSA CREW

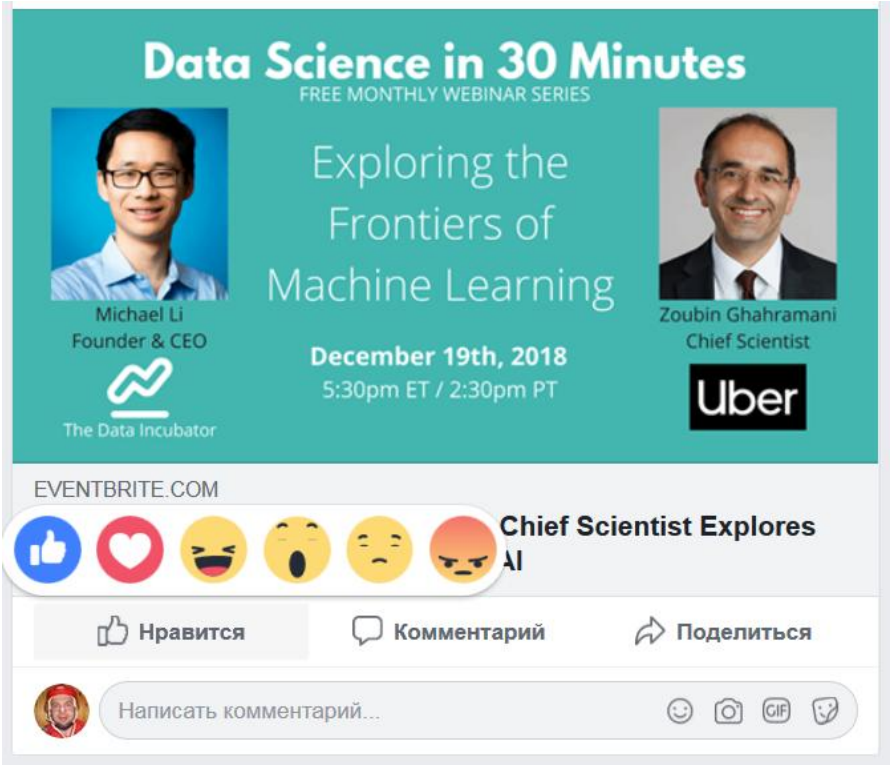
1 699 руб

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 ($)

не стоимость, а маржа + расходы на упаковку, доставку и т.п.

- **товары из категории (long-tail)**

Продать больший ассортимент / распродать

Разница между информационным поиском и рекомендательными системами

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

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)

+

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!

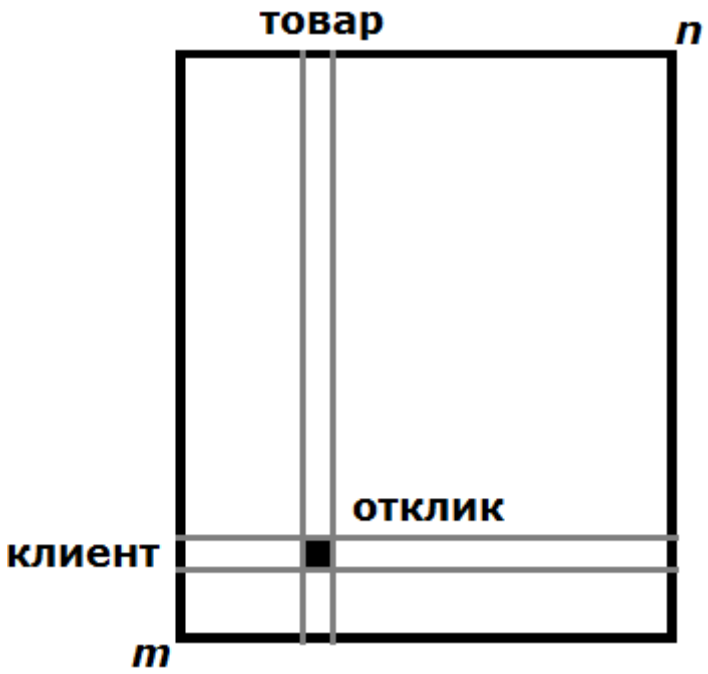
Решение на статистике поведения лучше, чем на описаниях!

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

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

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)

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

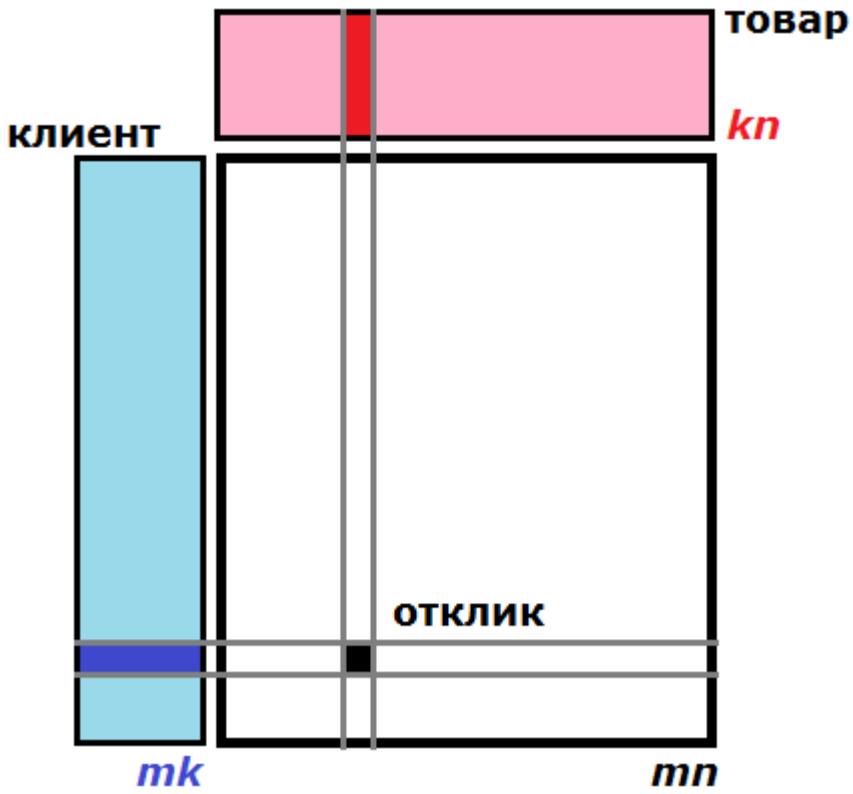
здесь – просмотры за последние 24 часа

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)**
  - **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++

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

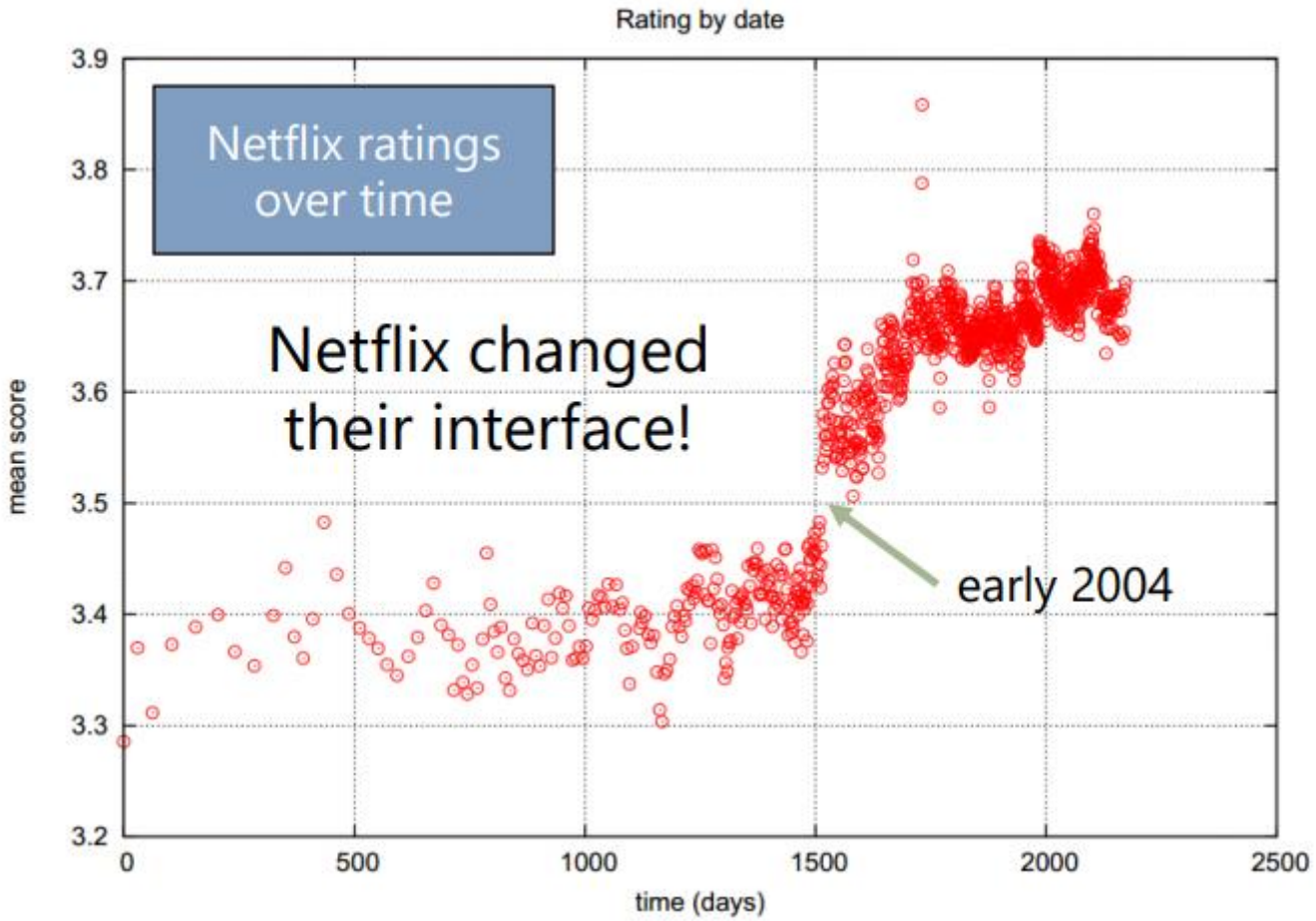
+ что просматривал, но не покупал пользователь

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

Simon Funk статья в блоге во время конкурса Netflix

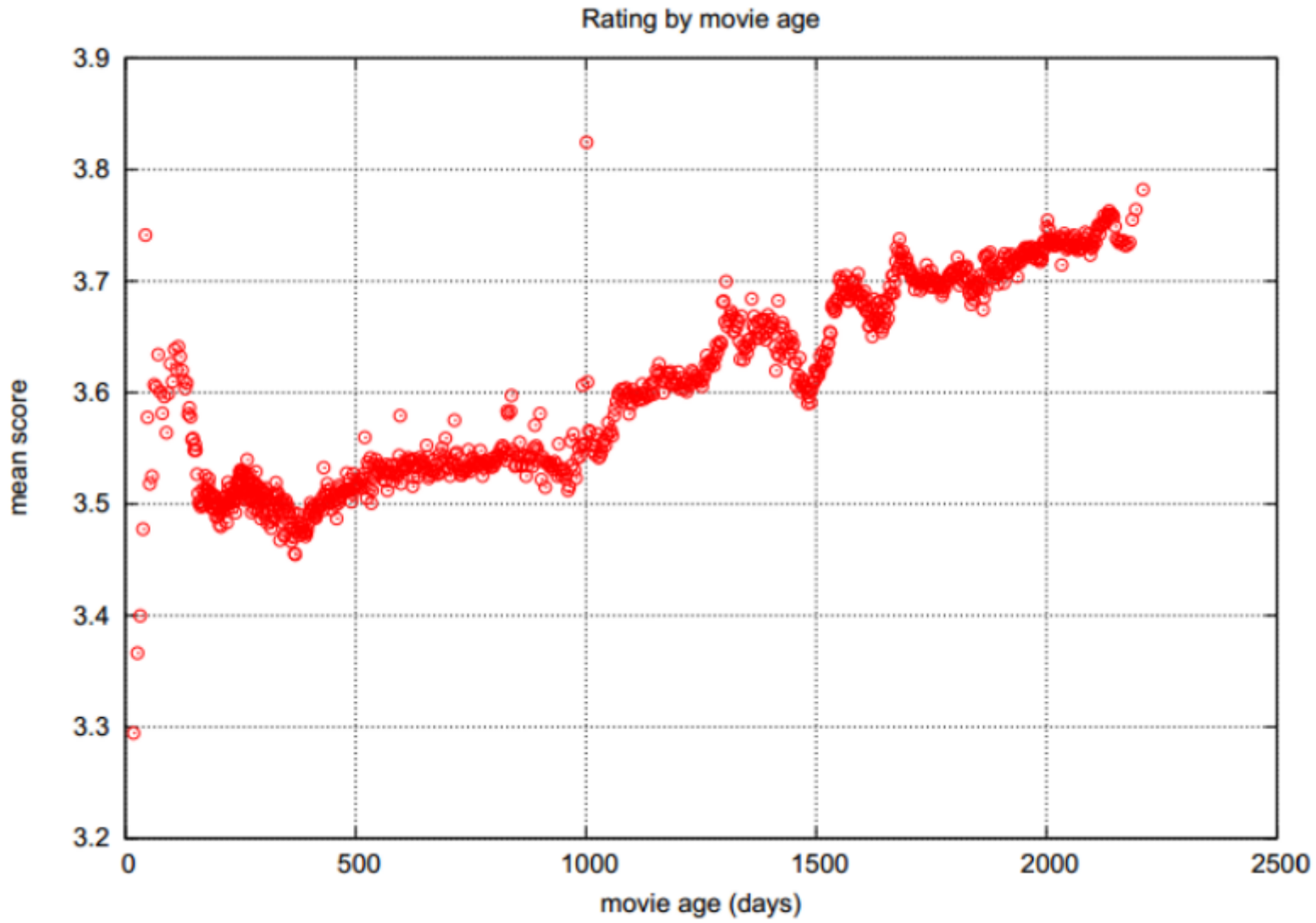
timeSVD++

Неизвестные зависят от времени…



Koren «Collaborative Filtering with Temporal Dynamics» KDD 2009

timeSVD++

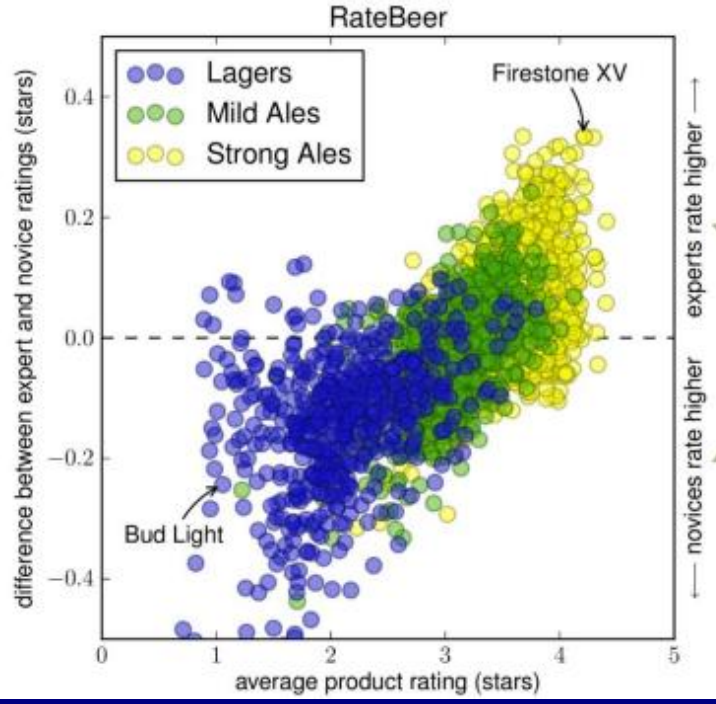


Люди склонны завышать рейтинги старых фильмов

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

timeSVD++

Регуляризация по времени

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

Адаптация SVD под социальные связи

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

$F(u)$ – **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.)

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)

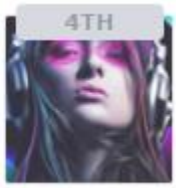
- **популярные становятся популярнее (popularity bias)**
- **условия шума (семейные аккаунты, случайные покупки и т.п.)**
- **возможны «атаки» на систему**

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…

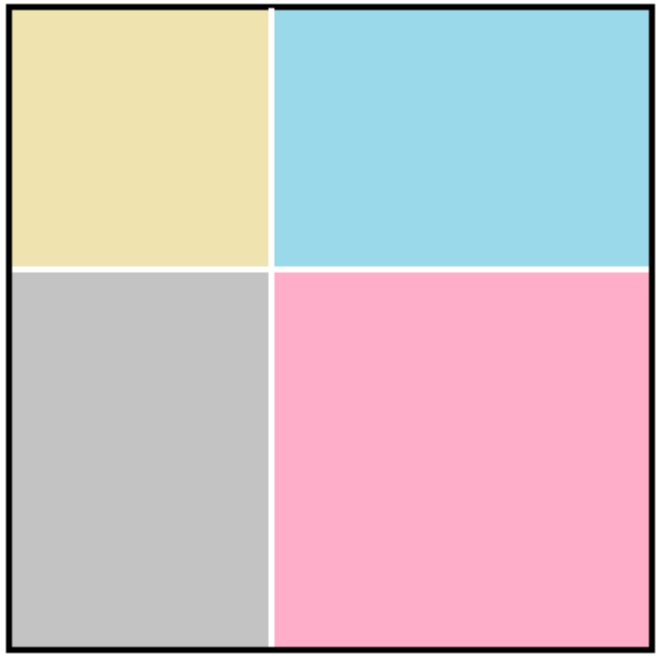
- **факторизация отдельных блоков (FFM – field-aware factorization machine)**
- **эффективное блочное хранение**

FFM – field-aware factorization machine

criteo labs Avazu

«RecSys 2015»

outbrain



Линейная модель

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

Полиномиальная модель (Poly2)

$$x^{\mathrm{T}}Wx = \sum_{1\le i<j\le n} w_{ij} x_i x_j$$

Факторизационная машина

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

Факторизационная машина с полями

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

$f(i)$ – **поле для** i

Оптимизационная задача

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

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

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)

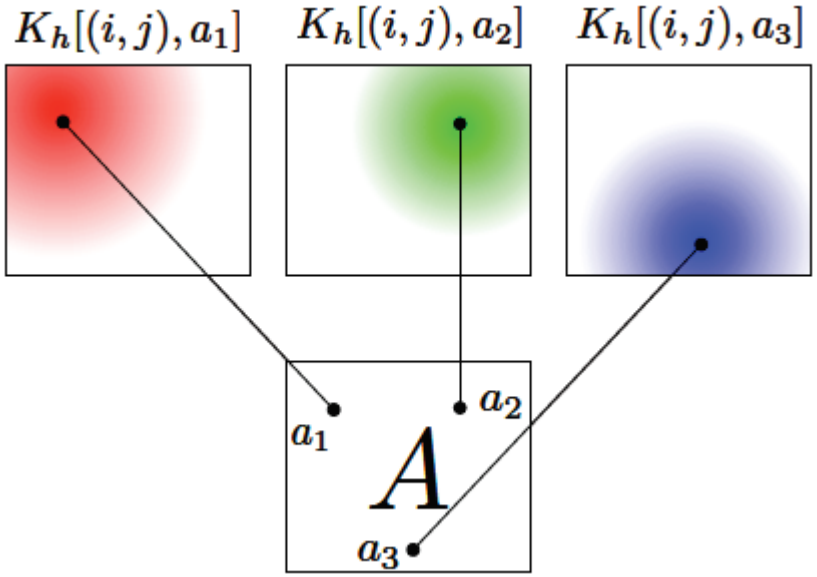
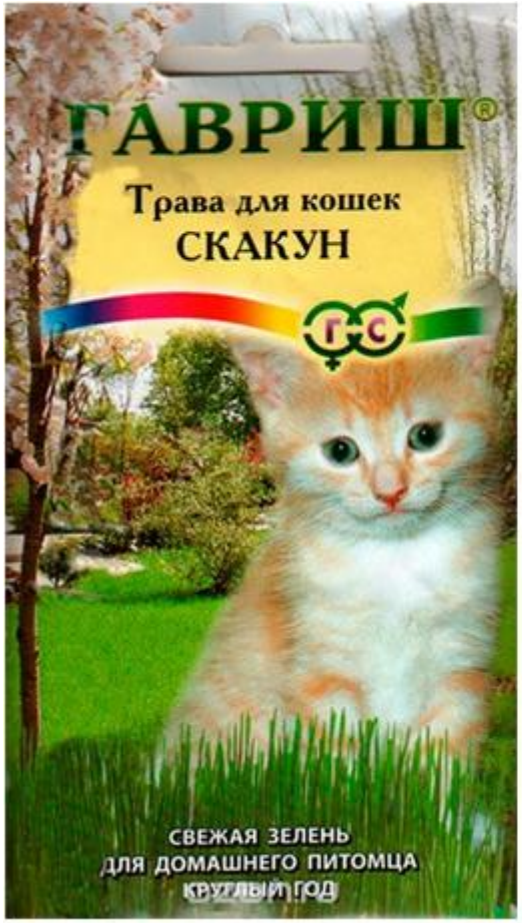


рис. из дипломной работы М.Трофимова

Простые методы



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

FPM – Frequent Pattern Mining

- **Ассоциативные правила (Association Rule Mining)**

если {A, B, C} ⇒ D (были в одной сессии)

- **Sequential Pattern Mining**

если A→...→B→...→C ⇒ D (были до)

Contiguous Sequential Pattern Mining

если A→B→C ⇒ D (были последовательно перед)

Кластеризация пользователей / товаров

(+ стандартные рекомендации)

есть и автоматические кластеры

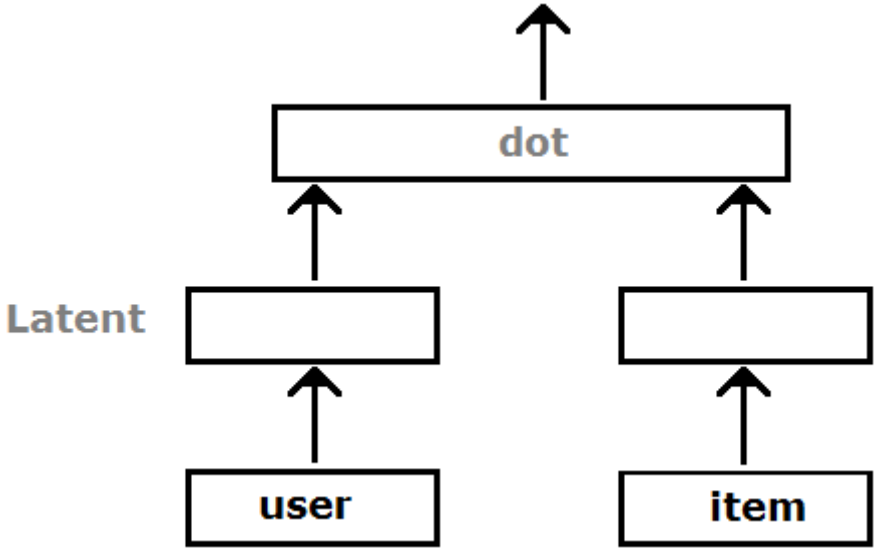
(интересы, любимые театры / жанры, актёры и т.п.)

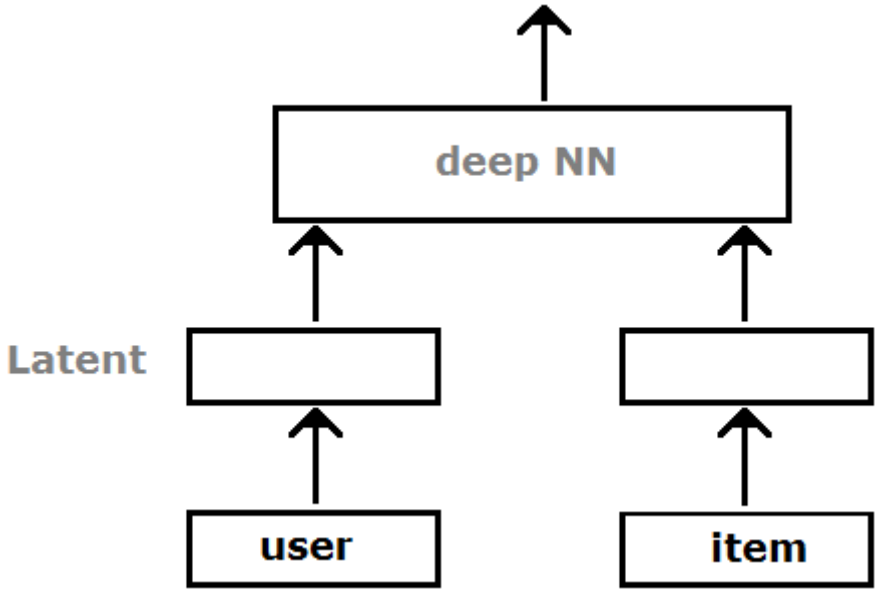
Методы на основе случайных блужданий

Laknath Semage «Recommender Systems with Random Walks: A Survey» // https://arxiv.org/pdf/1711.04101.pdf

Использование DL

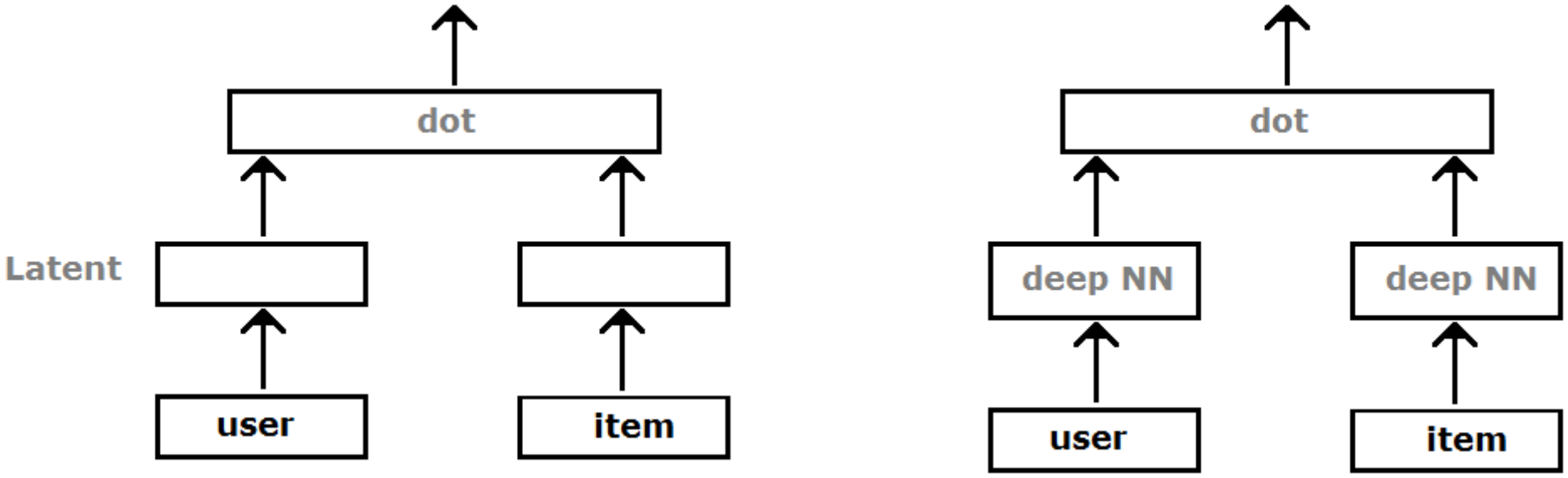
Deep CF





Использование DL

Deep Semantic Similarity Model (DSSM)



Здесь «user» – вся информация о пользователе!

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck «Learning deep structured semantic models for web search using clickthrough» // CIKM'13, P.2333–2338.

Ali Elkahky, Xiaodong He «A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems»
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/frp1159-songA.pdf

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

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

Хорошая презентация

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