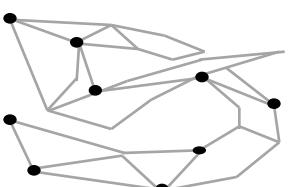


近期推荐系统概述

陈晓理 Data Scientist

数据应用学院



提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤+: Matrix Factorization (Alternative Least Squares)
 - 协同过滤++: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统 ~~CNN~~
 - 评估推荐系统结果 ~~☆~~
- 其他注意事项
 - 数据采集
 - 隐式变量与显式变量模型
 - 信息整合
 - 结果保护

推荐系统应用场景

e.g. Search Engine
Recommendation
System

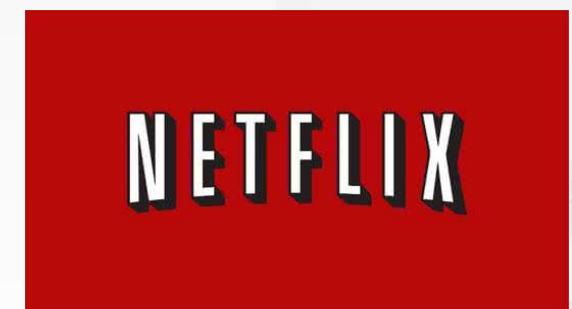
- 为什么要使用推荐系统?
1. Too much information
2. User doesn't know exactly what he/she needs to search.
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台

推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台

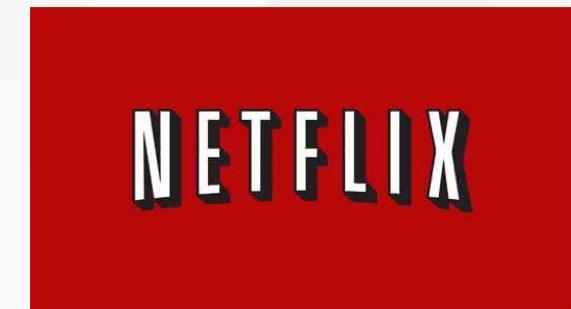
推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台



推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台
- 效果可好可坏



推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域**
 - 电商平台
 - 音乐影视
 - 社交平台
- 效果可好可坏**



Abhishek Gupta

Director of Engineering, Hired

[Follow | 597](#)

[Turn On Notifications](#) [Ask Question](#)

To put things in perspective, 50% of total job applications and job views by members are a direct result of recommendations. Interestingly, in the past year and half it has risen from 6% to 50%.

50%的岗位申请直接来源于推荐系统。很有意思，一年前，这个数字才只是6%。



推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台
- 效果可好可坏



推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台
- 效果可好可坏



推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台
- 效果可好可坏
- 怎样才是好的推荐系统？

推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台
- 效果可好可坏
- 怎样才是好的推荐系统？
 - 比用户还了解他/她需要什么产品

推荐系统应用场景

- 为什么要使用推荐系统？
- 以前的解决方案是什么？
- 应用领域
 - 电商平台
 - 音乐影视
 - 社交平台
- 效果可好可坏
- 怎样才是好的推荐系统?
 - 比用户还了解他/她需要什么产品
 - 准确，有新意

提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤⁺: Matrix Factorization (Alternative Least Squares)
 - 协同过滤⁺⁺: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统
- 评估推荐系统结果
- 其他注意事项
 - 数据采集
 - 信息整合

1. Popularity
2. Graphic

推荐系统算法概述



• 流行度

- 简单易行
- 缺乏个性化
- 不准确



推荐系统算法概述

②

- 基于图（关系）的推荐
 - 社交网络



推荐系统算法概述

- ③ • 基于内容(content based)
 - 需要显式的标签项
 - 优点：没有冷启动问题，不需要历史数据，没有流行度偏见，可以用用户内容特性来提供解释
 - 缺点：项目内容必须是机器可读的和有意义的，很难有意外，还是缺少多样性，很难发觉用户或者商品间的关联性

招聘商业分析师，要求：

1. 性别：男
2. 工作经验：2年
3. 985院校毕业
4. 理工科专业
5. 北京户口

推荐系统算法概述

- 基于内容(content based)
 - 需要显式的标签项
 - 优点：没有冷启动问题，不需要历史数据，没有流行度偏见，可以使用用户内容特性来提供解释
 - 缺点：项目内容必须是机器可读的和有意义的，很难有意外，还是缺少多样性，很难发觉用户或者商品间的关联性

招聘商业分析师，要求：

1. 性别：男
2. 工作经验：2年
3. 985院校毕业
4. 理工科专业
5. 北京户口

推荐系统算法概述

- 基于内容(content based)
 - 需要显式的标签项
 - 优点：没有冷启动问题，不需要历史数据，没有流行度偏见，可以用用户内容特性来提供解释
 - 缺点：项目内容必须是机器可读的和有意义的，很难有意外，还是缺少多样性，很难发觉用户或者商品间的关联性

招聘商业分析师，要求：

1. 性别：男
2. 工作经验：2年
3. 985院校毕业
4. 理工科专业
5. 北京户口

推荐系统算法概述

- 基于内容(content based)
 - 需要显式的标签项
 - 优点：没有冷启动问题，不需要历史数据，没有流行度偏见，可以使用用户内容特性来提供解释
 - 缺点：项目内容必须是机器可读的和有意义的，很难有意外，还是缺少多样性，很难发觉用户或者商品间的关联性
 - 真实条件下，缺乏显式标签项，需要通过关联性来揣测用户真实需求

招聘商业分析师，要求：

1. 性别：男
2. 工作经验：2年
3. 985院校毕业
4. 理工科专业
5. 北京户口

提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤⁺: Matrix Factorization (Alternative Least Squares)
 - 协同过滤⁺⁺: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统
- 评估推荐系统结果
- 其他注意事项
 - 数据采集
 - 信息整合

推荐系统算法概述

④

协同过滤推荐算法

- 基于历史用户行为模式或商品特征模式
- 优点：通过用户或者商品间的**关联性**，满足用户未挑明/隐式的要求
- 缺点：冷启动问题（没有历史数据怎么办）

推荐系统算法概述

- 协同过滤推荐算法
 - 基于历史用户行为模式或商品特征模式
 - 优点：通过用户或者商品间的关联性，满足用户未挑明/隐式的要求
 - 缺点：冷启动问题（没有历史数据怎么办）
- 协同过滤推荐算法几个步骤
 - **搜集历史数据**
 - 计算相似度
 - 进行推荐

	万科	新浪	腾讯	新希望	蒙牛
张三		Y	Y		
李四	Y			Y	Y
王五	Y	Y	Y		

推荐系统算法概述

- 协同过滤推荐算法
 - 基于历史用户行为模式或商品特征模式
 - 优点：通过用户或者商品间的关联性，满足用户未挑明/隐式的要求
 - 缺点：冷启动问题（没有历史数据怎么办）
- 协同过滤推荐算法几个步骤
 - 搜集历史数据
 - 计算相似度
 - 进行推荐

$$T(x, y) = \frac{x \cdot y}{\|x\|^2 \times \|y\|^2} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

	万科	新浪	腾讯	新希望	蒙牛
张三	0	1	1	0	0
李四	1	0	0	1	1
王五	1	1	1	0	0

张三和王五的相似度 (0, 1, 1, 0, 0) 与
(1, 1, 1, 0, 0) 的cosine 相似度是：
 $(1 + 0 + 0 + 0 + 1) / \sqrt{2 * 3} = 2 / \sqrt{6}$

The Formula of similarity is based on cos(vector 1, vector 2)

more similar -> cos value is larger. (if exactly the same, then cos (0 Degree) =1; If complete different (no overlap), then cos (90 Degree) =0)

$\text{cos} (\text{vec1}, \text{vec2}) = (\text{vec1} \cdot \text{dot} \text{vec2}) / [\text{mod}(\text{vec1}) * \text{mod}(\text{vec2})]$

推荐系统算法概述

- 协同过滤推荐算法
 - 基于历史用户行为模式或商品特征模式
 - 优点：通过用户或者商品间的**关联性**，满足用户**未挑明/隐式**的要求
 - 缺点：冷启动问题（没有历史数据怎么办）
- 协同过滤推荐算法几个步骤
 - 搜集历史数据
 - **计算相似度**
 - 进行推荐

$$T(x, y) = \frac{x \cdot y}{\|x\|^2 \times \|y\|^2} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

张三和王五的相似度 (0, 1, 1, 0, 0) 与
(1, 1, 1, 0, 0) 的cosine 相似度是：
 $(1 + 0 + 0 + 0 + 1) / \sqrt{2 * 3} = 2/\sqrt{6}$

Similarity Matrix

	张三	李四	王五
张三	1	0.0	$2/\sqrt{6}$
李四	0.0	1	??
王五	$2/\sqrt{6}$???	1

推荐系统算法概述

- 协同过滤推荐算法
 - 基于历史用户行为模式或商品特征模式
 - 优点：通过用户或者商品间的关联性，满足用户未挑明/隐式的要求
 - 缺点：冷启动问题（没有历史数据怎么办）
- 协同过滤推荐算法几个步骤
 - 搜集历史数据
 - 计算相似度
 - 进行推荐

Because 张三 和 李四 相似度
< 张三和王五 相似度,
therefore, if 张三 选择了新公
司, then most likely, 王五 也会
选择新公司, 而不是李四

$$T(x, y) = \frac{x \cdot y}{\|x\|^2 \times \|y\|^2} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

张三和王五的相似度 (0, 1, 1, 0, 0) 与
(1, 1, 1, 0, 0) 的cosine 相似度是
 $(1 + 0 + 0 + 0 + 1) / \sqrt{2 * 3} = 2 / \sqrt{6}$

	新公司
张三	?
李四	1
王五	0

推荐系统算法概述

- 协同过滤推荐算法
 - 基于历史用户行为模式或商品特征模式
 - 优点：通过用户或者商品间的关联性，满足用户未挑明/隐式的要求
 - 缺点：冷启动问题（没有历史数据怎么办）
- 协同过滤推荐算法几个步骤
 - 搜集历史数据
 - 计算相似度
 - 进行推荐
 - **潜在问题：**当用户或商品过多时，数据阵列非常稀疏

$$T(x, y) = \frac{x \bullet y}{\|x\|^2 \times \|y\|^2} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

张三和王五的相似度 (0, 1, 1, 0, 0) 与
(1, 1, 1, 0, 0) 的cosine 相似度是
 $(1 + 0 + 0 + 0 + 1) / \sqrt{2*3} = 2/\sqrt{6}$

推荐系统算法概述

- 协同过滤推荐算法
 - 基于历史用户行为模式或商品特征模式
 - 优点：通过用户或者商品间的关联性，满足用户未挑明/隐式的要求
 - 缺点：冷启动问题（没有历史数据怎么办）
- 协同过滤推荐算法几个步骤
 - 搜集历史数据
 - 计算相似度
 - 进行推荐
 - User based v.s Item based

$$T(x, y) = \frac{x \cdot y}{\|x\|^2 \times \|y\|^2} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

张三和王五的相似度 (0, 1, 1, 0, 0) 与
(1, 1, 1, 0, 0) 的cosine 相似度是
 $(1 + 0 + 0 + 0 + 1) / \sqrt{2*3} = 2/\sqrt{6}$

推荐系统算法概述

- 协同过滤算法
 - User based v.s. Item based
 - 从数据量与计算量上考虑
 - 从关注点上考虑
 - User based: 圈子
 - Item based: 物品功能

Size of Similarity Matrix

提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤+: 矩阵分解Matrix Factorization (Alternative Least Squares)
 - 协同过滤++: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统
- 评估推荐系统结果
- 其他注意事项
 - 数据采集
 - 信息整合

推荐系统算法概述：CF+矩阵分解

4

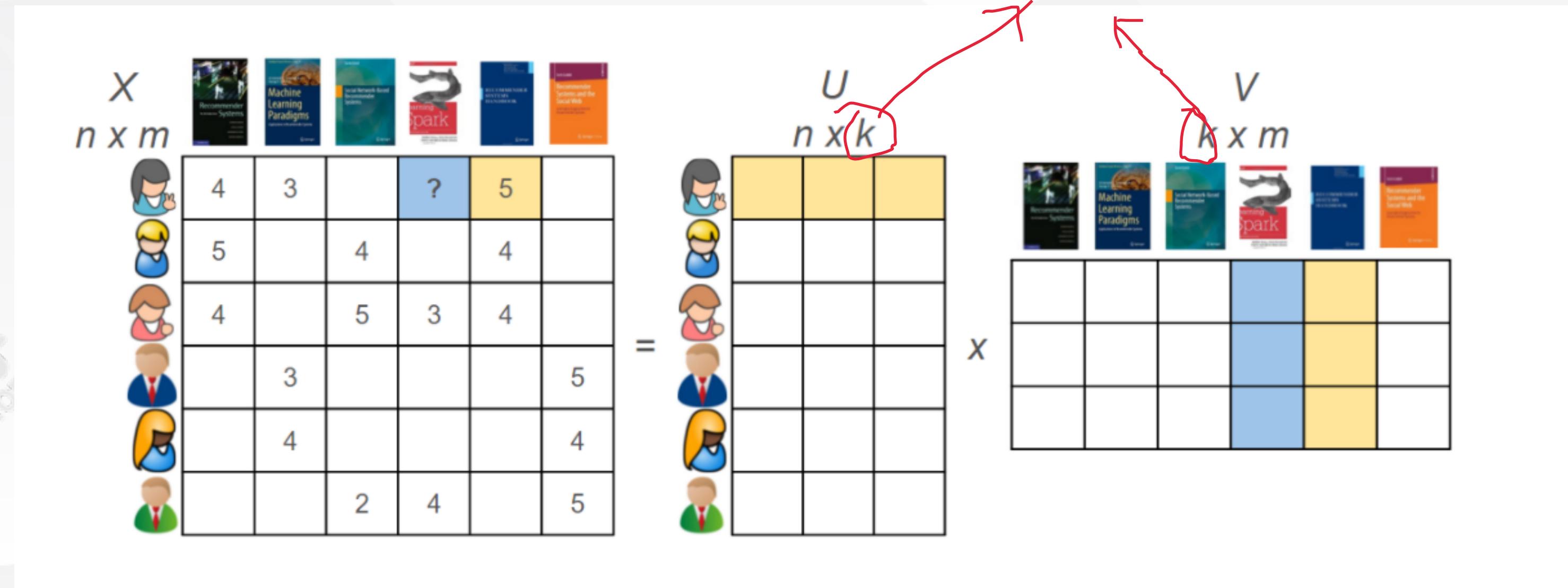
- 解决稀疏性问题

Similarity Matrix Factorization generates a set of hidden Variable (隐变量), with size of k.

意义: 产生的隐变量可以帮助发现一些新的特征 for relationship between e.g. User and Book shown below. (隐变量个数 \rightarrow 特征个数)

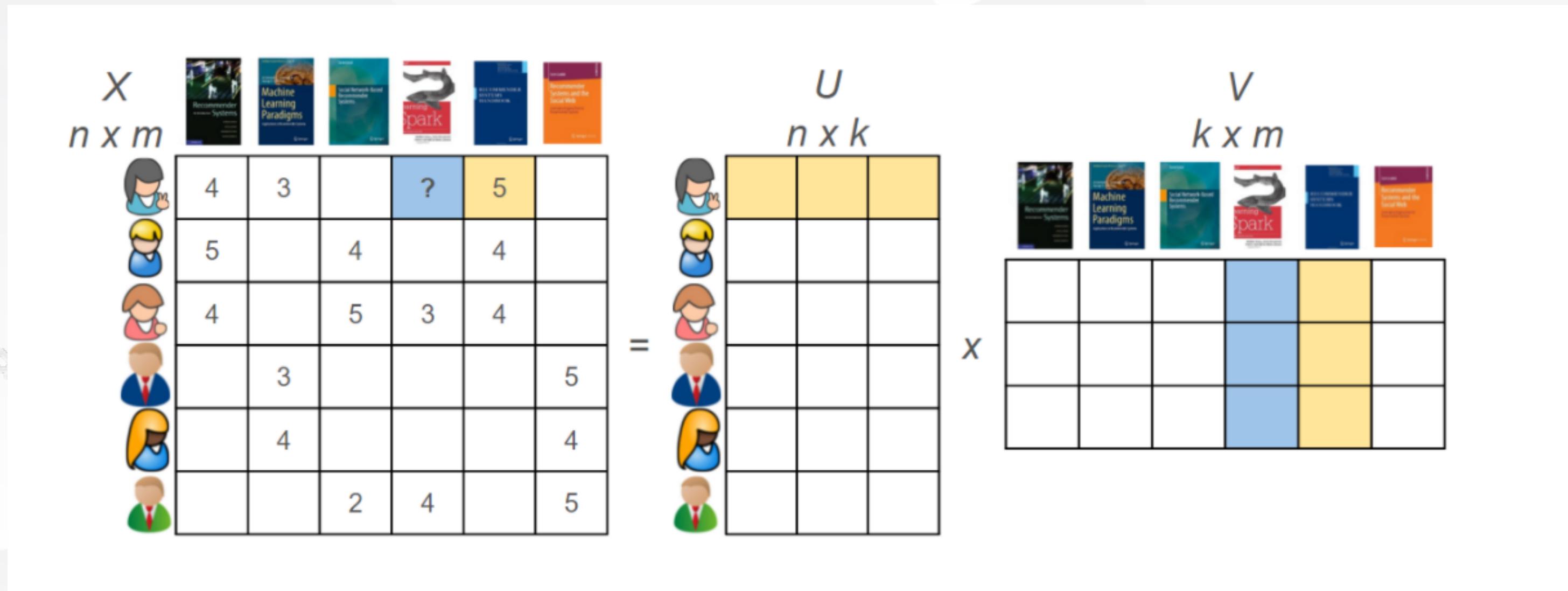
| For example:

k 个 隐变量 below can be:
爱好数据科学的人;
爱好计算机的人;
爱好数学的人



推荐系统算法概述：CF + 矩阵分解

- 解决稀疏性问题
- 分解方法：数值逼近Alternating Least Squares(ALS)



推荐系统算法概述：CF+矩阵分解

- 解决稀疏性问题
- 分解方法：数值逼近Alternating Least Squares(ALS)

$$Cost\ Function = \|X - U \times V^T\|_2 + \lambda(\|U\|_2 + \|V\|_2)$$

1. X已知，固定U，以V为未知量，使用最小二乘法求V，使Cost Function最小
2. X已知，固定V，以U为未知量，使用最小二乘法求U，使Cost Function最小
3. 重复步骤1, 2, 直至收敛

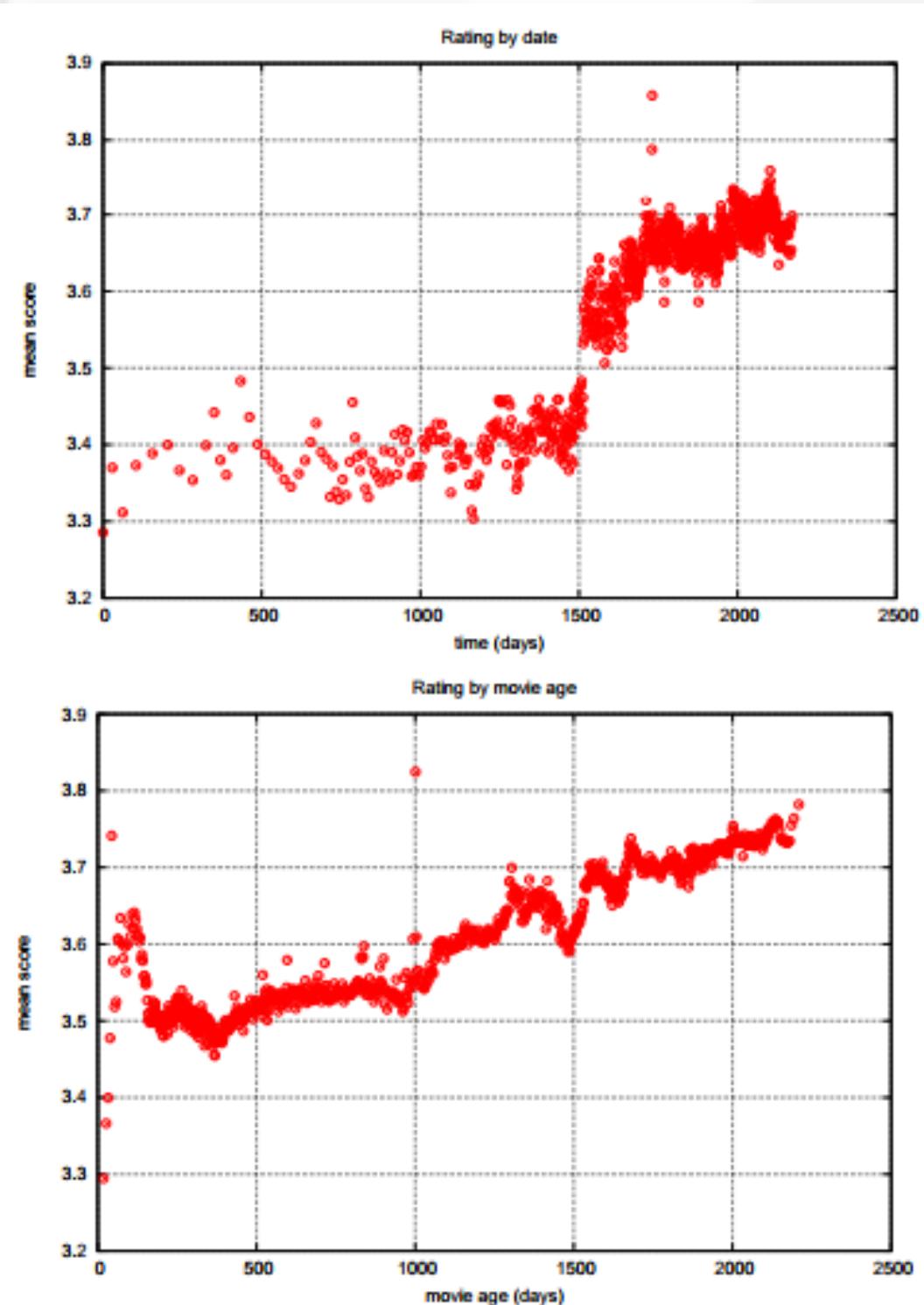
提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤⁺: 矩阵分解Matrix Factorization (Alternative Least Squares)
 - 协同过滤⁺⁺: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统
- 评估推荐系统结果
- 其他注意事项
 - 数据采集
 - 信息整合

推荐系统算法概述 : CF + 时间维度变化

4.2.

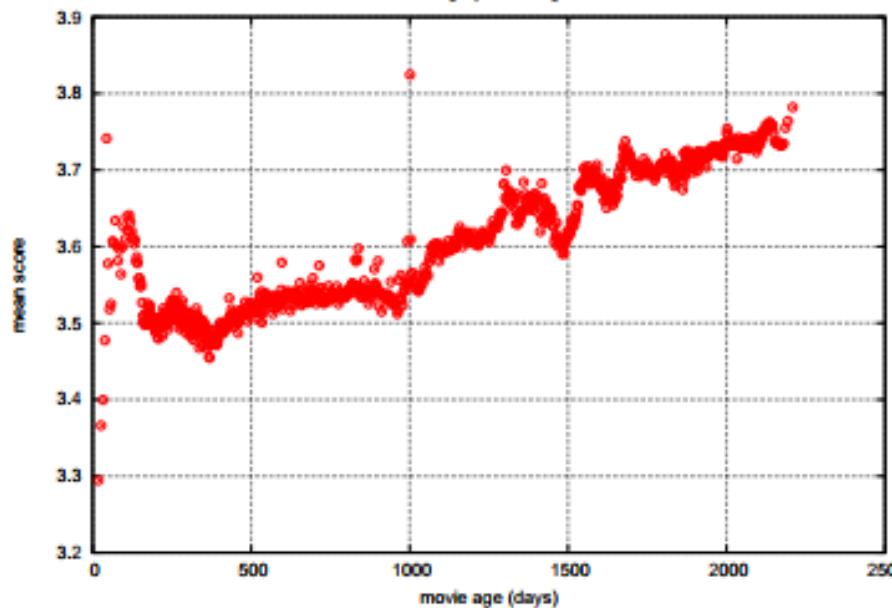
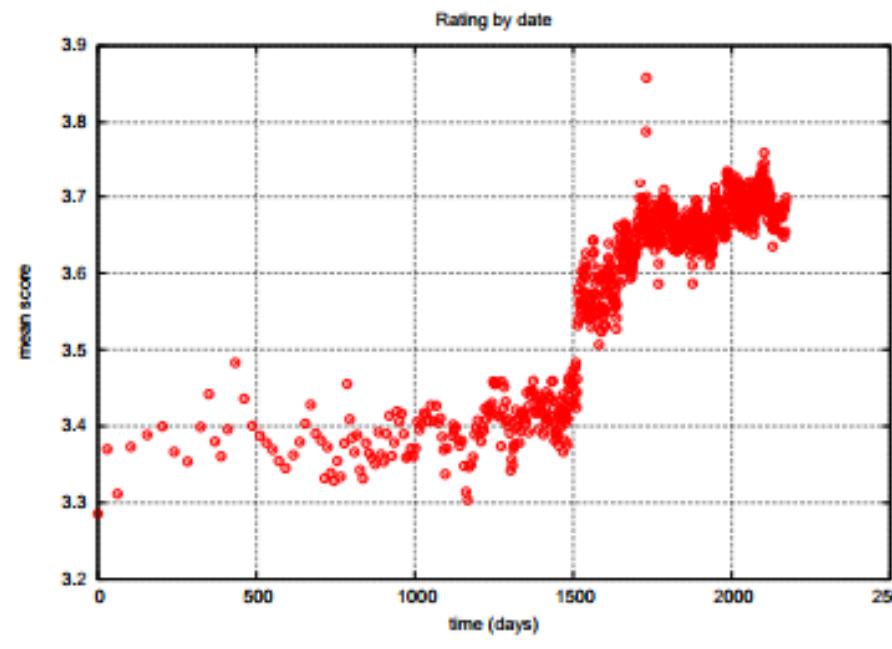
- 解决用户偏好随时间变化的问题(time drifting data)



Koren, Y. (2009). Collaborative filtering with temporal dynamics. *Knowledge Discovery and Data Mining {KDD}*, 447–456.
<https://doi.org/10.1145/1557019.1557072>

推荐系统算法概述：CF+矩阵分解

- 解决用户偏好随时间变化的问题(time drifting data)



分离基准模型与偏移量

$$b_{ui} = \mu + b_u + b_i$$

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

Algorithm: Too complicated, did not cover in lecture, can consult paper reference if interested below

引入时间变化后的分离基准模型与偏移量

$$b_i(t) = b_i + b_{i,\text{Bin}(t)} + b_{i,\text{period}(t)}$$

$$b_u(t) = b_u + \alpha_u \cdot \text{dev}_u(t) + b_{u,t} + b_{u,\text{period}(t)}$$

$$\hat{r}_{ui}(t) = \mu + b_i(t) + b_u(t) + q_i^T \left(p_u(t) + |R(u)|^{-\frac{1}{2}} \sum_{j \in R(u)} y_j \right)$$

Koren, Y. (2009). Collaborative filtering with temporal dynamics. *Knowledge Discovery and Data Mining {KDD}*, 447–456.
<https://doi.org/10.1145/1557019.1557072>

提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤+: 矩阵分解Matrix Factorization (Alternative Least Squares)
 - 协同过滤++: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统
- 评估推荐系统结果
- 其他注意事项
 - 数据采集
 - 信息整合

推荐系统算法概述：LSTM

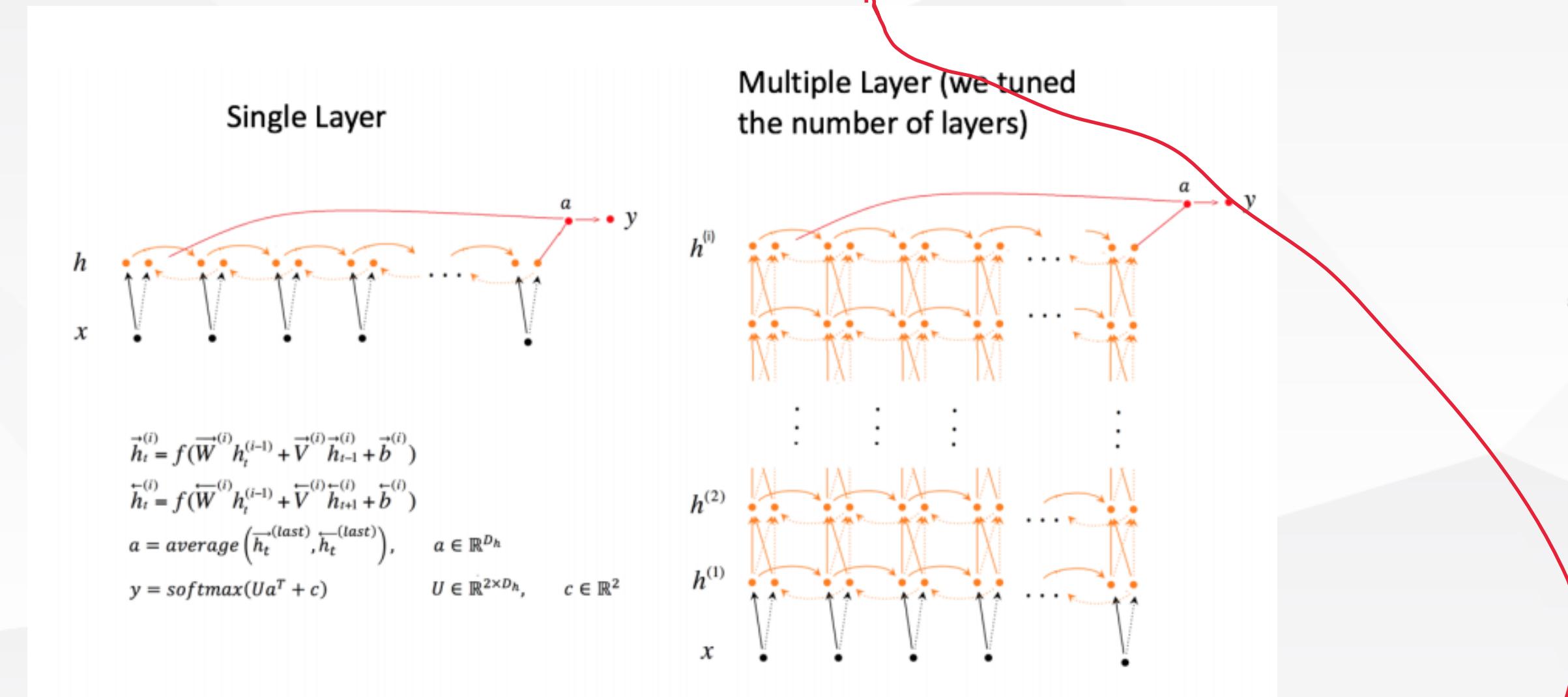
4.3.



推荐系统算法概述：LSTM

- 语义处理
- 考虑用户评价：Yelp评论

Algorithm Too complicated, not covered in lecture, consult paper reference below if interested.



Liu, D. Z., & Singh, G. (2013). A recurrent neural network based recommendation system. *Stanford Lecture*, 53(9), 1–30. Retrieved from <https://cs224d.stanford.edu/reports/LiuSingh.pdf> <http://ebooks.cambridge.org/ref/id/CBO9781107415324A009>

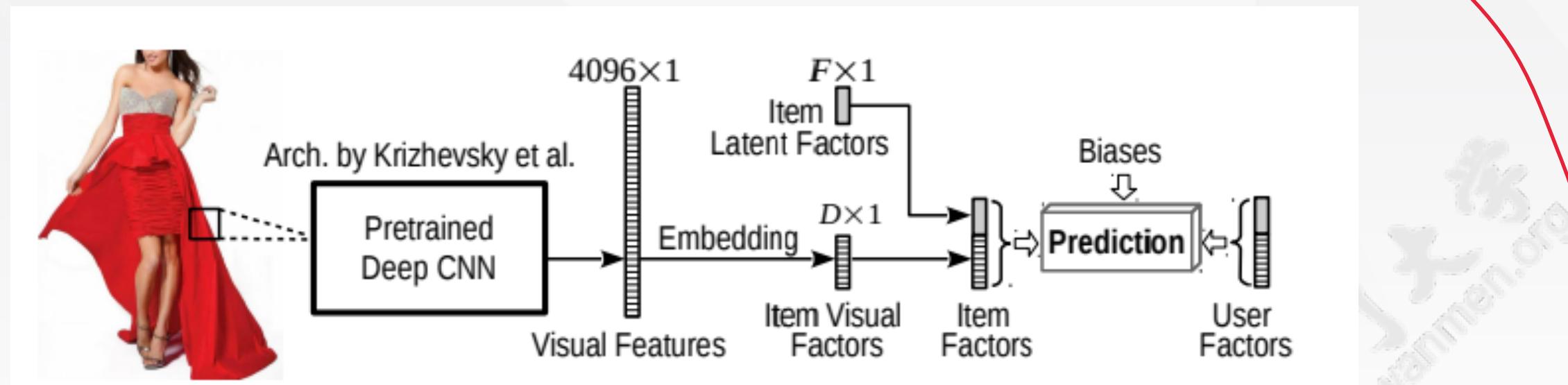
提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤⁺: 矩阵分解Matrix Factorization (Alternative Least Squares)
 - 协同过滤⁺⁺: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统
- 评估推荐系统结果
- 其他注意事项
 - 数据采集
 - 信息整合

推荐系统算法概述：基于图像的推荐

4.4

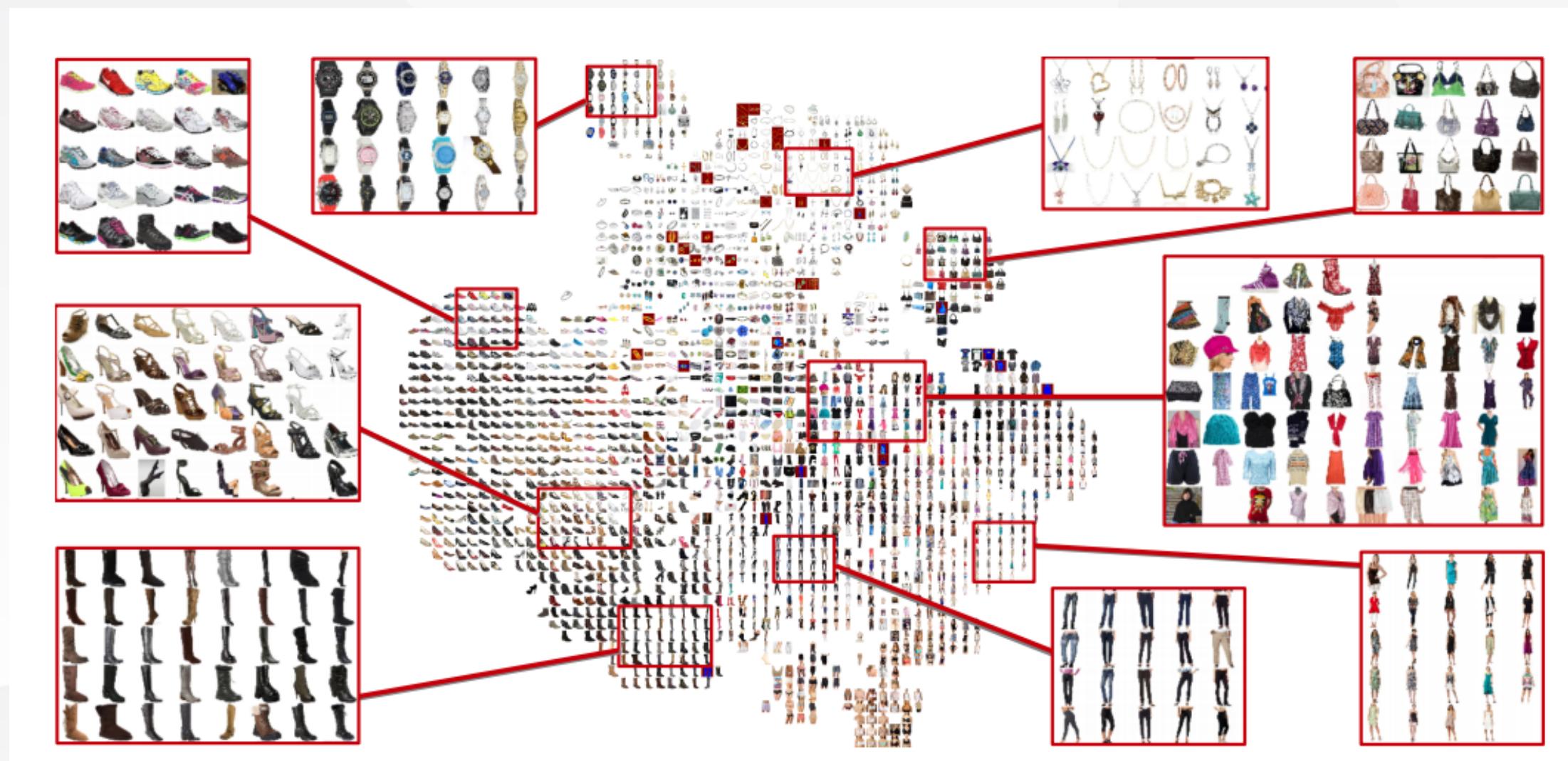
- 使用CNN提取特征
- 结合MF矩阵分解



He, R., & McAuley, J. (2015). VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. Retrieved from <http://arxiv.org/abs/1510.01784>

推荐系统算法概述：基于图像的推荐

- 使用CNN提取特征
- 结合MF矩阵分解



提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤⁺: 矩阵分解Matrix Factorization (Alternative Least Squares)
 - 协同过滤⁺⁺: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统
- 评估推荐系统结果
- 其他注意事项
 - 数据采集
 - 信息整合

评估

A=What user liked
B=What was predicted by Recommendation System

- 指标 $\frac{A \cap B}{A}$. $\frac{A \cap B}{B}$
 - 准确度, 召回率 : Precision Rate, Recall Rate
 - 覆盖度: 供应商关注的, 也与新颖程度相关
 - 新颖程度: 你知道的但顾客之前并没有意识到自己喜欢的
- On-line vs off-line
 - online: A/B testing 拿你的推荐系统apply 到网站上看效果, e.g. 使用后商品销量的变化
 - CTR vs Root-Mean-Square Error (RMSE)
 - off-line 用历史数据自己develop model



提纲

- 推荐系统应用场景
- 推荐系统算法概述
 - 基于内容的推荐(content based)
 - 协同过滤Collaborative Filtering
 - 协同过滤⁺: 矩阵分解Matrix Factorization (Alternative Least Squares)
 - 协同过滤⁺⁺: 引入延时间维度的变化
 - 基于Long Short Term Memory (LSTM)的推荐系统
 - 引入图像识别的推荐系统
- 评估推荐系统结果
- 其他注意事项
 - 数据采集
 - 信息整合

其他注意事项

- 数据采集：不断更新推荐系统
- 信息整合：行为数据

Activity Data from User.
E.g. User A bought item K
at Time t.