Recommendation System

Introduction

RECOMMENDATION MODEL

- X = SET OF CUSTOMERS
- S = SET OF ITEMS

- Utility function $u: X \times S \rightarrow R$
  - R = set of ratings
  - R is a totally ordered set
  - e.g., **0-5** stars, real number in **[0,1]**

注：x 和 s 分别是用户和推荐项目，他们共同组成下面的 utility matrix，其中是用户对项目的评分。

	AVATAR	LOTR	Matrix	Pirates
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

TYPES OF RECOMMENDATION

关键问题：utility matrix U 是稀疏的

- 大多数人没有对大多数项目进行评级
- 冷启动 (cold start)：新项目没有评级，新用户没有历史记录

推荐系统的方法：

- Content-based
- Collaborative

CONTENT-BASED RECOMMENDATION

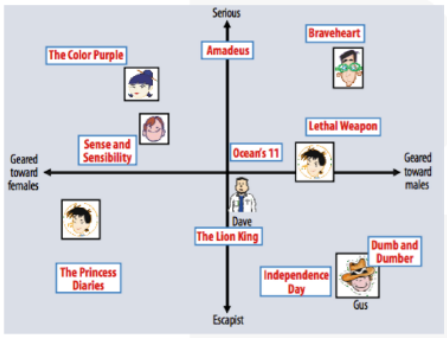
主要思想：向客户 X 推荐的项目，类似于之前被 X 高度评价的项目。就是根据历史推送相似内容。

MATRIX FACTORIZATION

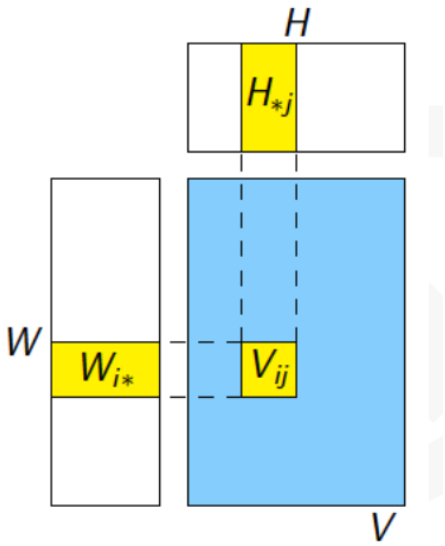
MATRIX FACTORIZATION

- User vectors:
  $(W_{u*})^T \in \mathbb{R}^r$
- Item vectors:
  $H_{*i} \in \mathbb{R}^r$
- Rating prediction:

$$V_{ui} = W_{u*} H_{*i}$$
$$= [WH]_{ui}$$



Figures from Koren et al. (2009)



- User vectors:
  $\mathbf{w}_u \in \mathbb{R}^r$
- Item vectors:
  $\mathbf{h}_i \in \mathbb{R}^r$
- Rating prediction:
  $v_{ui} = \mathbf{w}_u^T \mathbf{h}_i$

注：我们有一个用户的评分矩阵 V，可以将 V 分解成用户向量 U 和项目向量 W，再根据这两个预测。

COLLABORATIVE FILTERING

对于用户 X，找到 N 个评分与 X 的评分"相似"的其他用户。根据 N 中用户的评分估算 X 的评分。

FIND “SIMILAR” USERS

设 R_X 是用户 X 的评分矩阵。

JACCARD SIMILARITY MEASURE

- Jaccard distance = $1 - \frac{|v_1 \cap v_2|}{|v_1 \cup v_2|}$
- **Problem:** Ignores the value of the rating

$$J(A,B) = 1 - \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

注：要计算 Jaccard distance，会给定两个向量 A 和 B，它们由 0 和 1 组成 (也可能由其他数组成)。其中 M_{01} 代表 A 是 0 但 B 是 1 (或 A 没评分 B 有评分) 的位置的个数；M_{10} 代表 A 是 1 但 B 是 0 (或 A 有评分 B 没评分) 的位置的个数；M_{11} 代表 A 是 1 B 也是 1 (或 A 有评分 B 也有评分) 的位置的个数。

Cosine similarity measure

- sim(***x***, ***y***) = arccos(***r*** , ***r***) = $\frac{r_x \cdot r_y}{\|r_x\| \cdot \|r_y\|}$
- **Problem:** Treats missing ratings as “negative”

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Example

- Calculate the following distance measures between the two users with the different ratings, r1 = [0, 1, 1, 0, 0, 0,. 1], and r2 = [1, 0, 1, 0, 1, 0, 0].

- (a) What is the Jaccard distance between two users?
  (b) What is the Cosine distance between two user? (You can use arccos(x) to present the answer).

(a)，Jaccard distance = 1 - 1/5 = 4/5

(b)，Cosine distance = 1/3

Other example:

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- **Intuitively we want: sim(*A*, *B*) > sim(*A*, *C*)**
- **Jaccard similarity: 1/5 < 2/4**
- **Cosine similarity: 0.386 > 0.322**

ITEM-ITEM COLLABORATIVE FILTERING

之前我们介绍的实际上是 user-user collaborative filtering，即使用用户之间的相似性。

而现在的 item-item，是使用项目之间的相似性。

- 对于项目 i，查找其他类似项目
- 根据类似项目的评级估算项目 i 的评级
- 可以使用与 user-user 模型中相同的相似度量和预测函数

$$r_{xi} = \frac{1}{k} \sum_{y \in N} r_{yi}$$

注：上式是一个预测评分的方法，即对相似的 N 个项目的评分求均值。

Example

对于下面这些数据，我们要估计出用户 5 对电影 1 的评分，即红色区域的值，这里 N = 2，即选择两个最相似的项目进行预测。

USERS

movies	1	2	3	4	**5**	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

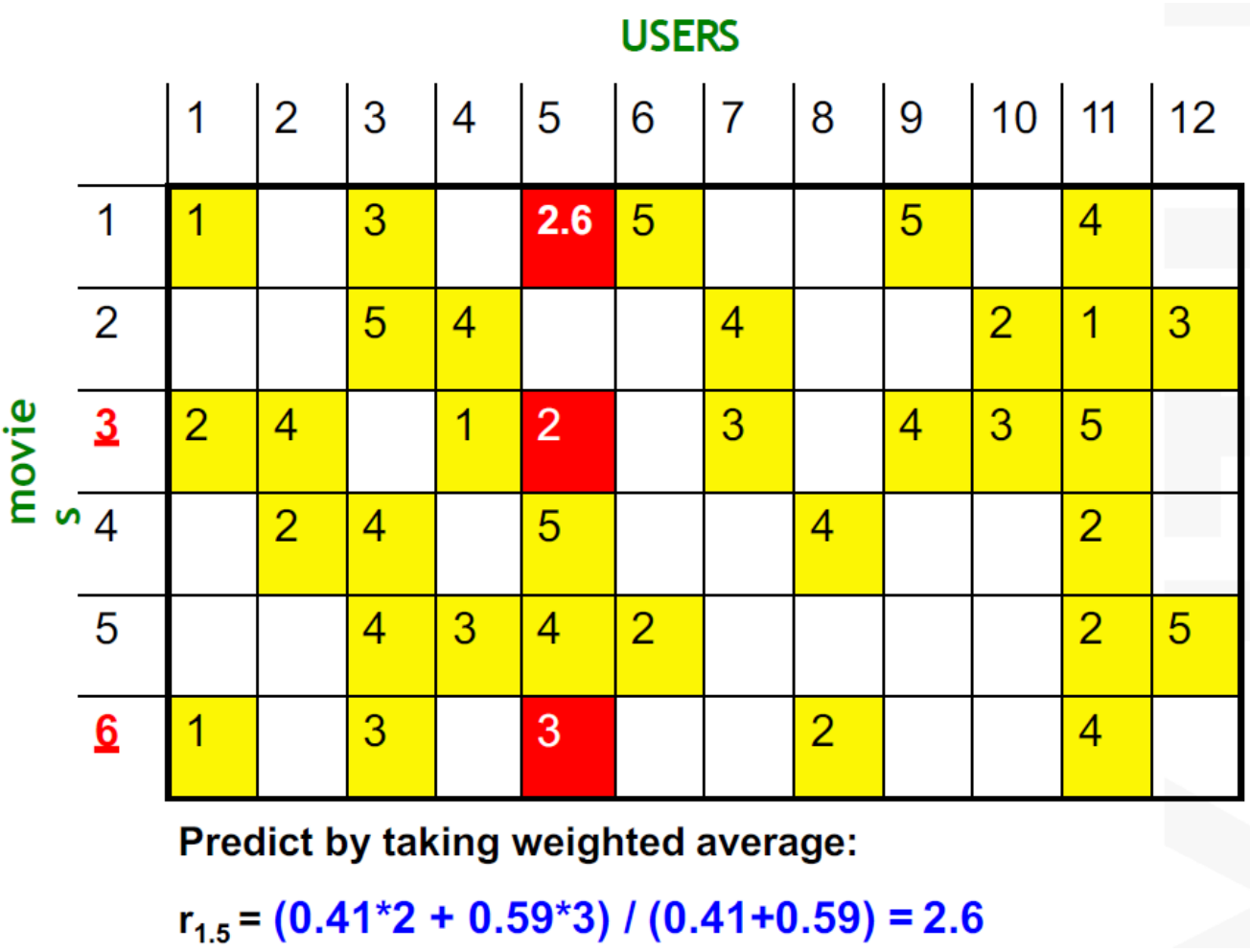
第一步，计算项目之间的相似性。比如我们现在要求的是项目 1，现在就有计算所有项目和它的相似性，即求每一行和第一行的余弦距离。

movies	1	2	3	4	**5**	6	7	8	9	10	11	12	sim(1,m)
1	1		3		?	5			5		4		1.00
2			5	4			4			2	1	3	-0.18
3	2	4		1	2		3		4	3	5		**0.41**
4		2	4		5			4			2		-0.10
5			4	3	4	2					2	5	-0.31
6	1		3		3			2			4		**0.59**

Neighbor selection:
Identify movies similar to movie **1**, **rated by user 5**

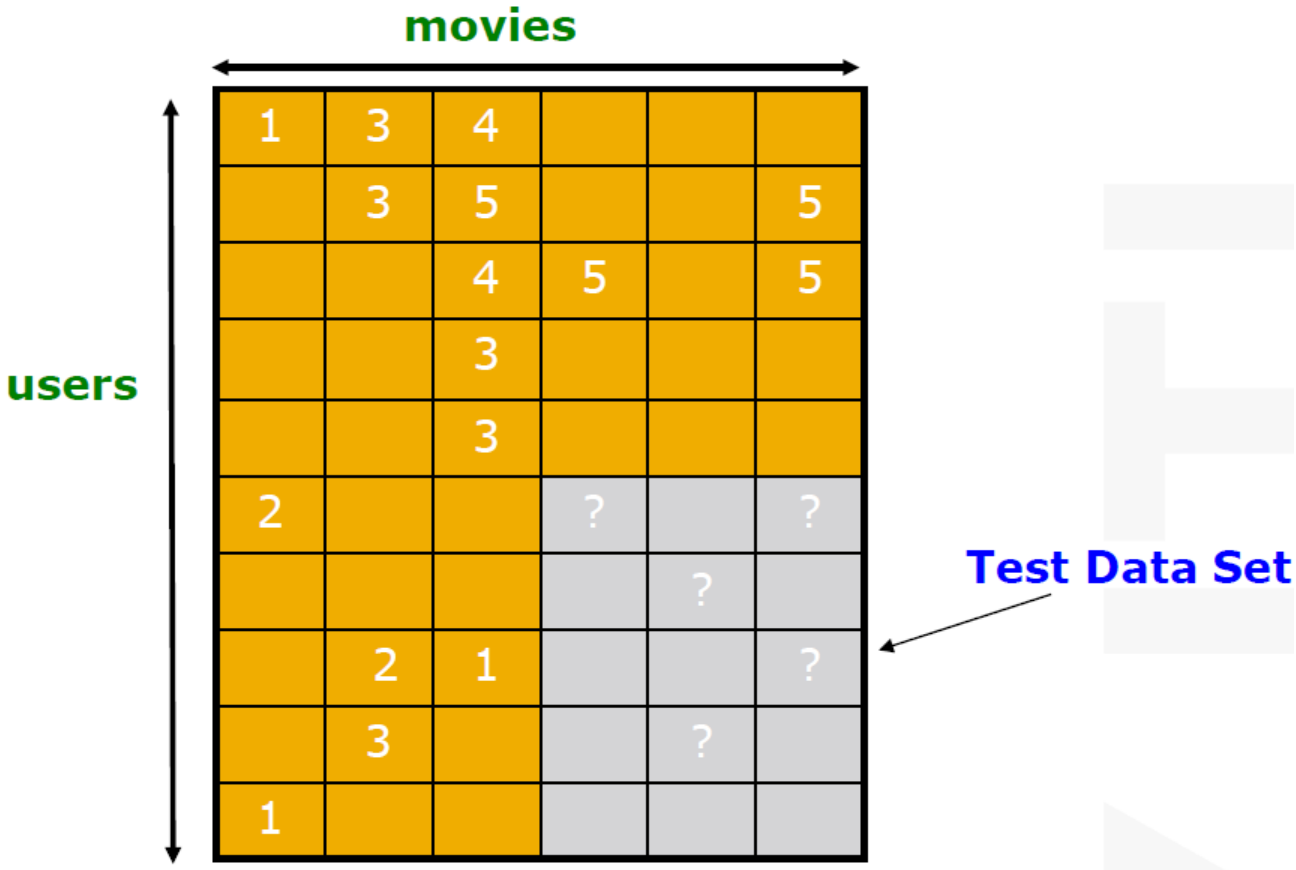
1. Compute cosine similarities between rows

现在我们知道了项目 3 和 6 与 1 比较相似。我们就可以根据用户 5 对项目 3 和 6 的评分估计出他对项目 1 的评分。



EVALUATION

可以将数据的一部分作为测试集。



Question

- Consider a dataset containing information about movies: genre, director and release decade. We also have information about which users have seen each movie. The rating for a user on a movie is either 0 or 1.

Movie	Release decade	Genre	Director	Total number of ratings
A	1970s	Humor	D_1	40
B	2010s	Humor	D_1	500
C	2000s	Action	D_2	300
D	1990s	Action	D_2	25
E	2010s	Humor	D_3	1

- Consider user U1=[2000s, D2, Humor]. We have some existing recommender system R that recommended the movie B to user U1.

(a) Given the above dataset, which one(s) do you think R could be?
  - User-user collaborative filtering.
  - Item-item collaborative filtering
  - Content-based recommender system.

(b) If some user U2 wants to watch a movie, under what conditions can our recommender system R recommend U2 a movie?

(c) If R recommends a movie, how to do it? If R cannot recommend a movie, please explain why it cannot be recommended.

(d) State any additional information R might want from U2 for predicting a movie for this user, if required.

(a)，User-user CF 和 item-item CF 都可以。