



# ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

---

## **DS300** **HỆ KHUYẾN NGHỊ**

### **Lọc cộng tác**



**Giảng viên:** ThS. Nguyễn Văn Kiệt

CN. Huỳnh Văn Tín

**Bộ môn Khoa học Dữ liệu**

**Khoa Khoa học và Kỹ thuật Thông tin**

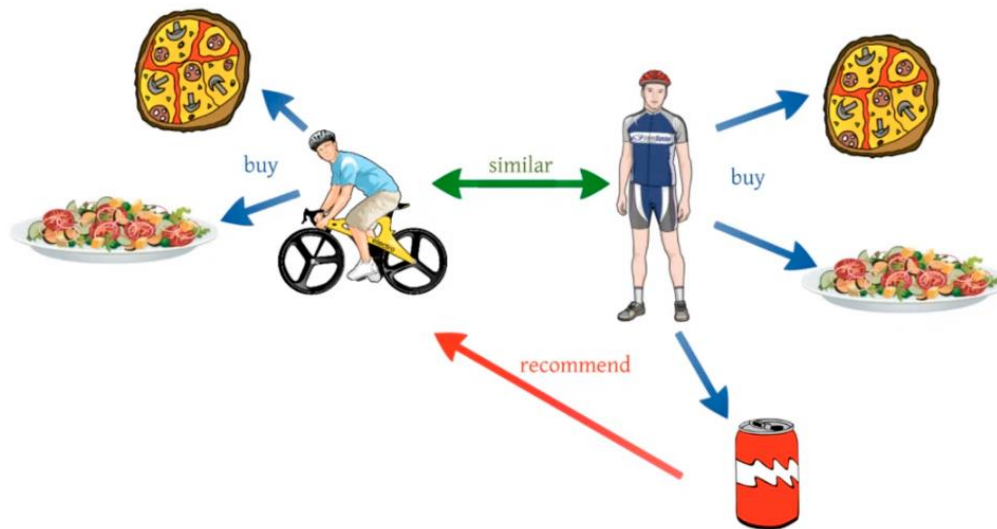
# Nội dung

---

- ❖ Giới thiệu phương pháp lọc cộng tác
- ❖ Ma trận đánh giá
- ❖ Độ đo tương đồng
- ❖ Lọc dựa trên người dùng
- ❖ Tiếp cận dựa trên bộ nhớ & mô hình
- ❖ Ưu điểm, hạn chế

# Giới thiệu CF

- Dùng phổ biến (sites thương mại điện tử, nghe nhạc, xem phim, ...)
- Sử dụng tri thức đám đông (wisdom of crowd) để khuyến nghị
- User đưa ra các đánh giá (rating) rõ ràng hoặc tiềm ẩn cho sản phẩm, dịch vụ họ quan tâm.
- **Ý tưởng và giải thuyết**
  - Những người có sở thích tương tự trong quá khứ, thì cũng sẽ có sở thích tương tự trong tương lai.



## Ma trận đánh giá

- Cho không gian người dùng  $U = \{u_1, u_2, \dots, u_n\}$  và không gian các đối tượng  $I = \{i_1, i_2, \dots, i_m\}$ . Ma trận  $R$  kích thước  $n \times m$ , chứa các giá trị đánh giá  $r_{k,j}$ , với  $k \in 1 \dots n$ ,  $j \in 1 \dots m$ . Những giá trị đánh giá  $r_{k,j}$  thể hiện mức độ hữu ích của đối tượng  $I_j$  với một người dùng  $u_k$ . Giá trị  $r_{k,j}$  có thể là nguyên hay thực trong một khoảng cho trước tùy vào bài toán cụ thể. Thông thường, giá trị đánh giá  $r_{k,j}$  trong một hệ thống ứng dụng phổ biến nhận các giá trị từ 1 (ít hữu ích) đến 5 (rất hữu ích). Nếu một người dùng  $u_k$  chưa thể hiện đánh giá với một đối tượng  $I_j$  thì  $r_{k,j} = 0$  (rỗng) và cần được tính toán, xác định.

# Ma trận đánh giá

**Movies You've Rated**

Based on your 745 movie ratings, this is the list of movies you've seen. As you discover movies on the website that you've seen, rate them and they will show up on this list. On this page, you may change the rating for any movie you've seen, and you may remove a movie from this list by clicking the 'Clear Rating' button.

Sort by > **Star Rating**

Jump to > **5 Stars**

TITLE	MPAA	GENRE	STAR RATING
<a href="#">Add</a> <a href="#">12 Angry Men</a> (1957)	UR	Classics	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">The 39 Steps</a> (1935)	UR	Classics	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">An American in Paris</a> (1951)	UR	Classics	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">The Andromeda Strain</a> (1971)	G	Sci-Fi & Fantasy	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">Apollo 13</a> (1995)	PG	Drama	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">The Battle of Algiers</a> (1965) La Battaglia di Algeri	UR	Foreign	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">Being There</a> (1979)	PG	Drama	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">Big Deal on Madonna Street</a> (1958) I soliti ignoti	UR	Foreign	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">The Birds</a> (1963)	PG-13	Thrillers	★★★★★ Clear Rating
<a href="#">Add</a> <a href="#">Blade Runner</a> (1982)	R	Sci-Fi & Fantasy	★★★★★ Clear Rating


























Value	Graphic representation	Textual representation
5	☆☆☆☆☆	Excellent
4	☆☆☆☆	Very good
3	☆☆☆	Good
2	☆☆	Fair
1	☆	Poor

Table 9.1: User-Item Matrix

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

# Người đồng sở thích

- **Người đồng sở thích:** Cho  $U$  là không gian người dùng, gọi  $S_u$  là tập hợp những người đồng sở thích với  $u \in U$ ,  $S_u \subseteq U$ . Những người đồng sở thích với  $u$  **là những người có hành vi quá khứ hay các đánh giá tương tự với trên cùng những đối tượng khuyến nghị** từ ma trận đánh giá  $R$ .

# Người đồng sở thích

---

	$I_1$	$I_2$	$I_3$	$I_4$	$I_5$
$x$	5	3	4	4	?
$U_1$	3	1	2	3	3
$U_2$	4	3	4	3	5
$U_3$	3	3	1	5	4
$U_4$	1	5	5	2	1

**Ai là người đồng sở thích với X?**

# Độ đo tương đồng

---

## Cosine Similarity

$$\text{sim}(U_u, U_v) = \cos(U_u, U_v) = \frac{U_u \cdot U_v}{\|U_u\| \|U_v\|} = \frac{\sum_i r_{u,i} r_{v,i}}{\sqrt{\sum_i r_{u,i}^2} \sqrt{\sum_i r_{v,i}^2}}.$$

## Pearson Correlation Coefficient

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

Trong đó:

- $r_{u,i}$ : Giá trị đánh giá của người dùng  $u$  đối với sản phẩm  $i$
- $\bar{r}_u$ : Giá trị đánh giá trung bình của người dùng  $u$



# Thuật toán lọc cộng tác

---

Các loại thuật toán lọc cộng tác:

□ **Memory-based**: Đề xuất trực tiếp dựa trên xếp hạng trước đó trong ma trận được lưu trữ mô tả quan hệ người dùng – đối tượng khuyến nghị

□ **Model-based**: Giả định rằng một mô hình cơ bản (hypothesis) chi phối cách người dùng xếp hạng các đối tượng khuyến nghị.

- This model can be approximated and learned.
- The model is then used to recommend ratings.
- **Example**: users rate low budget movies poorly

# Thuật toán lọc cộng tác dựa trên người dùng

---

**Input:** Ma trận đánh giá  $R$  (rating matrix)

**Output:**

- Top- $N$  những đối tượng được khuyến nghị
- Dự đoán giá trị của  $f(u,i)$ , mức độ quan tâm của người dùng  $u$  với đối tượng  $i$ .

# Cập nhật giá trị ratings

---

- Trung bình đánh giá:

$$r_{u,i} = \frac{1}{|N(u)|} \sum_{v \in N(u)} r_{v,i}$$

$N(u)$ : tập các user lân cận với  $u$

# Cập nhật giá trị ratings

---

- Tổng hợp đánh giá có trọng số:

$$r_{u,i} = \frac{\sum_{v \in N(u)} \text{sim}(u, v)(r_{v,i})}{\sum_{v \in N(u)} \text{sim}(u, v)}$$

$N(u)$ : tập các user lân cận với  $u$

# Cập nhật giá trị ratings

---

- Tổng hợp đánh giá dựa trên khoảng cách đánh giá:

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in N(u)} \text{sim}(u, v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N(u)} \text{sim}(u, v)}$$

$N(u)$ : tập các user lân cận với  $u$


$\bar{r}_u$  : rating trung bình của người dùng  $u$

$\bar{r}_v$  : rating trung bình của người dùng  $v$

## Ví dụ

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

Dự đoán rating của  
Jane đối với Aladdin



## Ví dụ

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

Dự đoán rating của Jane đối với Aladdin

1- tính ratings trung bình

$$\bar{r}_{John} = \frac{3 + 3 + 0 + 3}{4} = 2.25$$

$$\bar{r}_{Joe} = \frac{5 + 4 + 0 + 2}{4} = 2.75$$

$$\bar{r}_{Jill} = \frac{1 + 2 + 4 + 2}{4} = 2.25$$

$$\bar{r}_{Jane} = \frac{3 + 1 + 0}{3} = 1.33$$

$$\bar{r}_{Jorge} = \frac{2 + 2 + 0 + 1}{4} = 1.25$$

## Ví dụ

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

Dự đoán rating của Jane đối với Aladdin

2- Tính độ tương đồng giữa hai người dùng bằng cosine

$$\text{sim}(\text{Jane}, \text{John}) = \frac{3 \times 3 + 1 \times 3 + 0 \times 3}{\sqrt{10} \sqrt{27}} = 0.73$$

$$\text{sim}(\text{Jane}, \text{Joe}) = \frac{3 \times 5 + 1 \times 0 + 0 \times 2}{\sqrt{10} \sqrt{29}} = 0.88$$

$$\text{sim}(\text{Jane}, \text{Jill}) = \frac{3 \times 1 + 1 \times 4 + 0 \times 2}{\sqrt{10} \sqrt{21}} = 0.48$$

$$\text{sim}(\text{Jane}, \text{Jorge}) = \frac{3 \times 2 + 1 \times 0 + 0 \times 1}{\sqrt{10} \sqrt{5}} = 0.84$$



## Ví dụ

3- Tính rating của Jane đối với Aladdin bằng Tổng hợp đánh giá dựa trên khoảng cách đánh giá , giả sử neighborhood size = 2

$$\begin{aligned}r_{Jane, Aladdin} &= \bar{r}_{Jane} + \frac{sim(Jane, Joe)(r_{Joe, Aladdin} - \bar{r}_{Joe})}{sim(Jane, Joe) + sim(Jane, Jorge)} \\&\quad + \frac{sim(Jane, Jorge)(r_{Jorge, Aladdin} - \bar{r}_{Jorge})}{sim(Jane, Joe) + sim(Jane, Jorge)} \\&= 1.33 + \frac{0.88(4 - 2.75) + 0.84(2 - 1.25)}{0.88 + 0.84} = 2.33\end{aligned}$$

# Ưu, nhược điểm

## ☐ Ưu điểm

- ☐ Không cần phải có thêm thông tin về người dùng hoặc nội dung của các mặt hàng
  - Xếp hạng hoặc lịch sử mua hàng của người dùng là thông tin duy nhất cần thiết để hoạt động

## ☐ Nhược điểm

- ☐ Dữ liệu thưa (Data sparsity)
- ☐ Khởi động lạnh (Cold-start)
- ☐ Khả năng mở rộng (Scalability)
- ☐ Thiếu sự đa dạng và vấn đề long tail

# Bài tập 1

	$I_1$	$I_2$	$I_3$	$I_4$	$I_5$
$U_1$	3	0	2	3	3
$U_2$	4	3	4	3	5
$U_3$	3	3	0	5	4
$U_4$	1	5	5	0	1
$U_5$	5	3	4	4	?

**Tính giá trị rating của  $U_5$  đối với  $I_5$ ?**

- Tương quan pearson
- Tổng hợp đánh giá có trọng số
- neighborhood size = 3

## Bài tập 2

- a. Tìm những người có sở thích tương đồng với người dùng  $u3$  dùng hệ số tương quan Pearson.
- b. Nếu chọn số lân cận là 2 (tức chọn 2 người có sở thích gần nhất) với  $u3$ . Tính giá trị đánh giá của  $u3$  với  $p2$ , là  $f(u3, p2)$ , dùng phương pháp tổng hợp giá trị đánh giá dựa trên khoảng đánh giá

	p1	p2	p3	p4	p5
u1	1	2	?	2	5
u2	2	2	5	5	5
u3	1	?	5	2	4
u4	1	5	5	2	1

Trong đó: Khi tính độ tương đồng giữa hai người dùng, chỉ tính trên tập các sản phẩm đã được cả hai người dùng đánh giá.

# Bài tập 2

•  $u_3, u_4$  :  $\Rightarrow \text{Sim}(u_3, u_4) = 0,996$

\*  $u_3, u_4$  : chung  $p_1, p_2, p_4, p_5$ .

$\Rightarrow \bar{r}_{u_3} = (2 + 5 + 5 + 5)/4 = 4,25$

$\bar{r}_{u_4} = (1 + 5 + 2 + 1)/4 = 2,25$

-  $r_{u_3} - \bar{r}_{u_3}$

	$p_1$	$p_2$	$p_4$	$p_5$
$u_3$	-2,25	0,75	0,75	0,75
$u_4$	-2	2	-1	1

$\Rightarrow \text{Sim}(u_3, u_4) = \frac{2,25 \cdot 2 + 0,75 \cdot 2 - 0,75 + 0,75}{\sqrt{(2,25^2 + 0,75^2 + 0,75^2 + 0,75^2) \cdot (2^2 + 2^2 + 1^2 + 1^2)}} = 0,996$

\*  $u_3, u_4$  : chung  $p_1, p_2, p_4, p_5$ .

$\bar{r}_3 = 3$

$\bar{r}_4 = (1 + 5 + 2 + 1)/4 = 2,25$

-  $r_{u_3} - \bar{r}_{u_3}$

	$p_1$	$p_2$	$p_4$	$p_5$
$u_3$	-2	2	-1	1
$u_4$	-1,25	2,75	-0,25	-1,25

$\Rightarrow \text{Sim}(u_3, u_4) = \frac{2 \cdot 1,25 + 2 \cdot 2,75 + 0,25 - 1,25}{\sqrt{(2^2 + 2^2 + 1^2 + 1^2) \cdot (1,25^2 + 2,75^2 + 0,25^2 + 1,25^2)}} = 0,68$

$\Rightarrow$   $p_5$  đồng sở thích với  $u_3$  theo thứ tự:  $u_1, u_2, u_4$ .

	p1	p2	p3	p4	p5
u1	1	2	?	2	5
u2	2	2	5	5	5
u3	1	?	5	2	4
u4	1	5	5	2	1

# Bài tập 2

\*  $f(u_3, p_2)$  (5)

$K=2 \Rightarrow$  Người gần với  $u_3$  là  $u_1$  và  $u_2$ .

• Trung bình đánh giá.

$\Rightarrow f(u_3, p_2) = \frac{r_{u_1, p_2} + r_{u_2, p_2}}{2} = \frac{2 + 2}{2} = 2.$

• Tổng hợp giá trị đánh giá có trọng số.

$$r_{u,p} = \frac{\sum_{u_i \in S_u} \text{Sim}(u, u_i) \times r_{u_i, p}}{\sum_{u_i \in S_u} |\text{Sim}(u, u_i)|}$$

$\Rightarrow f(u_3, p_2) = \frac{0,996 \cdot 2 + 0,73 \cdot 2}{0,996 + 0,73} = 2$

\* Tổng hợp dựa trên khoảng đánh giá.

$$r_{u,p} = \bar{r}_u + \frac{\sum_{u_i \in S_u} \text{Sim}(u, u_i) \times (r_{u_i, p} - \bar{r}_{u_i})}{\sum_{u_i \in S_u} |\text{Sim}(u, u_i)|}$$

$\Rightarrow f(u_3, p_2) = 3 + \frac{0,996 \cdot (2 - 2,33) + 0,73 \cdot (2 - 4,25)}{0,996 + 0,73}$

$= 1,86.$

	p1	p2	p3	p4	p5
u1	1	2	?	2	5
u2	2	2	5	5	5
u3	1	?	5	2	4
u4	1	5	5	2	1