



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

DS300 **HỆ KHUYẾN NGHỊ**

Lọc cộng tác (tt)

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

- ❖ Nhắc lại lọc cộng tác dựa trên user
- ❖ Bài tập
- ❖ Lọc cộng tác dựa trên item
- ❖ Độ đo tương tự
- ❖ Vấn đề dữ liệu thừa

Lọc cộng tác dựa trên item

- Ý tưởng: Dùng sự tương tự trên đối tượng khuyến nghị i (item) để dự đoán giá trị của $f(u,i)$ thay vì tương tự dựa trên người dùng u .
- $f(X,i5) = ?$
- Chọn 2 lân cận gần nhất cho $i5 \rightarrow i1, i4$. Dựa trên $f(X,i1)$ và $f(X,i4)$ để tính $f(X,i5)$

	i1	i2	i3	i4	i5
X	<u>5</u>	?	?	<u>4</u>	?
u1	3	1	4	3	3
u2	4	3	2	3	5
u3	3	3	1	5	4
u4	1	5	5	2	1

Độ đo tương đồng cosine

- Thông tin đánh giá (rating) có thể xem như một vector trong không gian n chiều.
- Với a,b là 2 đối tượng khuyến nghị cần xem mức độ liên quan với nhau

$$\text{sim}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

	a	b
X	4	?
u1	3	3
u2	3	5
u3	5	4
u4	2	1

Độ đo tương đồng Pearson (cosine hiệu chỉnh)

- Với a, b là 2 đối tượng khuyến nghị cần xem xét mức độ liên quan với nhau
- U : tập người dùng cùng có đánh giá trên a, b .
- $r_{u,a}$: đánh giá u trên a

$$\text{sim}(\vec{a}, \vec{b}) = \frac{\sum_{u \in U} (r_{u,a} - \bar{r}_a)(r_{u,b} - \bar{r}_b)}{\sqrt{\sum_{u \in U} (r_{u,a} - \bar{r}_a)^2} \sqrt{\sum_{u \in U} (r_{u,b} - \bar{r}_b)^2}}$$

	a	b
X	4	?
u1	3	3
u2	3	5
u3	5	4
u4	2	1

Cập nhật giá trị ratings

- Trung bình đánh giá:

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

$N(i)$: Tập các item lân cận với i

Cập nhật giá trị ratings

- Tổng hợp đánh giá có trọng số:

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

$N(i)$: Tập các item lân cận với i

Cập nhật giá trị ratings

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

$$r_{u,i} = \bar{r}_i + \frac{\sum_{j \in N(i)} \text{sim}(i,j)(r_{u,j} - \bar{r}_j)}{\sum_{j \in N(i)} \text{sim}(i,j)}$$

$N(i)$: Tập các item lân cận với i


\bar{r}_i : Giá trị rating trung bình của item i

\bar{r}_j : Giá trị rating trung bình của item j

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

$$\bar{r}_{Lion\ King} = \frac{3 + 5 + 1 + 3 + 2}{5} = 2.8$$

$$\bar{r}_{Aladdin} = \frac{0 + 4 + 2 + 2}{4} = 2.$$

$$\bar{r}_{Mulan} = \frac{3 + 0 + 4 + 1 + 0}{5} = 1.6$$

$$\bar{r}_{Anastasia} = \frac{3 + 2 + 2 + 0 + 1}{5} = 1.6$$

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 bộ phim bằng cosine

$$\text{sim}(\text{Aladdin}, \text{Lion King}) = \frac{0 \times 3 + 4 \times 5 + 2 \times 1 + 2 \times 2}{\sqrt{24} \sqrt{39}} = 0.84$$

$$\text{sim}(\text{Aladdin}, \text{Mulan}) = \frac{0 \times 3 + 4 \times 0 + 2 \times 4 + 2 \times 0}{\sqrt{24} \sqrt{25}} = 0.32$$

$$\text{sim}(\text{Aladdin}, \text{Anastasia}) = \frac{0 \times 3 + 4 \times 2 + 2 \times 2 + 2 \times 1}{\sqrt{24} \sqrt{18}} = 0.67$$

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

$$\begin{aligned} r_{Jane, Aladdin} &= \bar{r}_{Aladdin} + \frac{sim(Aladdin, Lion King)(r_{Jane, Lion King} - \bar{r}_{Lion King})}{sim(Aladdin, Lion King) + sim(Aladdin, Anastasia)} \\ &\quad + \frac{sim(Aladdin, Anastasia)(r_{Jane, Anastasia} - \bar{r}_{Anastasia})}{sim(Aladdin, Lion King) + sim(Aladdin, Anastasia)} \\ &= 2 + \frac{0.84(3 - 2.8) + 0.67(0 - 1.6)}{0.84 + 0.67} = 1.40 \end{aligned}$$

Vấn đề dữ liệu thừa

❑ Cold start

New user

	p1	p2	p3	p4	p5
Ux	?	?	?	?	?
u1			2		3
u2		3			2
u3	5		1	4	
u4				2	

New item

	p1	p2	p3	p4	Px
u1			2		?
u2			1		?
u3		2			?
u4	5		2	4	?
u5			5		?

Vấn đề dữ liệu thưa

- ❑ Cold start, giải pháp?
 - Ép người dùng rating
 - Gán giá trị mặc định
 - Dùng thông tin khác (cá nhân, content,...)
 - “Chuyển tiếp/lan truyền” lân cận
 - Graph-based
-

Bài tập

- a. Tìm những sản phẩm có sự tương đồng với sản phẩm p_2, p_3 dùng hệ số tương quan Pearson.
- b. Nếu chọn số lân cận là 2 (tức chọn 2 sản phẩm tương đồng nhất) với p_2, p_3 . Tính giá trị đánh giá của u_1 với p_3 , là $f(u_1, p_3)$ và u_3 với p_2 , là $f(u_3, p_2)$, 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 sản phẩm, chỉ tính trên tập những người dùng cùng đánh giá trên 2 sản phẩm.