Sistemas de recomendación: filtro colaborativo basado en memoria

Gestión de la Información
Grado en Ingeniería Informática
Universidad de Burgos

UNIVERSIDAD DE BURGOS

José Ignacio Santos, José Manuel Galán
jisantos@ ubu.es, jmgalan@ ubu.es

Contenidos

- ➔ Filtros colaborativos basados en memoria
- ➔ Filtro colaborativo basado en usuarios
- ➔ Filtro colaborativo basado en productos
- ➔ Problemas de filtros colaborativos

Clases de sistemas de recomendación

Sistemas de recomendación
➔ Filtros basados en contenidos
➔ Filtros colaborativos

➔ Basados en modelos
➔ Basados en memoria

FC basados en modelos: utilizan los datos para ajustar modelos que después pueden ser utilizados para proponer recomendaciones (e.g. modelos de regresión)
FC basados en memoria: utilizan los datos para definir similitudes entre usuarios y productos que serán utilizados para construir las recomendaciones (e.g. Amazon)

Clases de sistemas de recomendación

- Sistemas de recomendación
  - ➔ Filtros basados en contenidos
  - ➔ Filtros colaborativos
    - ➔ Basados en modelos
    - ➔ Basados en memoria
      - ➔ Basados en usuarios
      - ➔ Basados en productos

FC basados en usuarios:
La estimación del rating de un usuario **u** al producto **i** se basa en los ratings al producto **i** de otros **usuarios similares** a **u**

FC basados en productos: (e.g. Amazon)
La estimación del rating de un usuario **u** al producto **i** se basa en los ratings del usuario **u** a **productos similares** a **i**

Filtro colaborativo basado en usuarios

La **hipótesis** de un filtro colaborativo basado en usuarios es que:

- Si dos usuarios han valorado de forma similar un conjunto de productos, lo harán también en otro producto

Los FC basados en usuarios requieren:

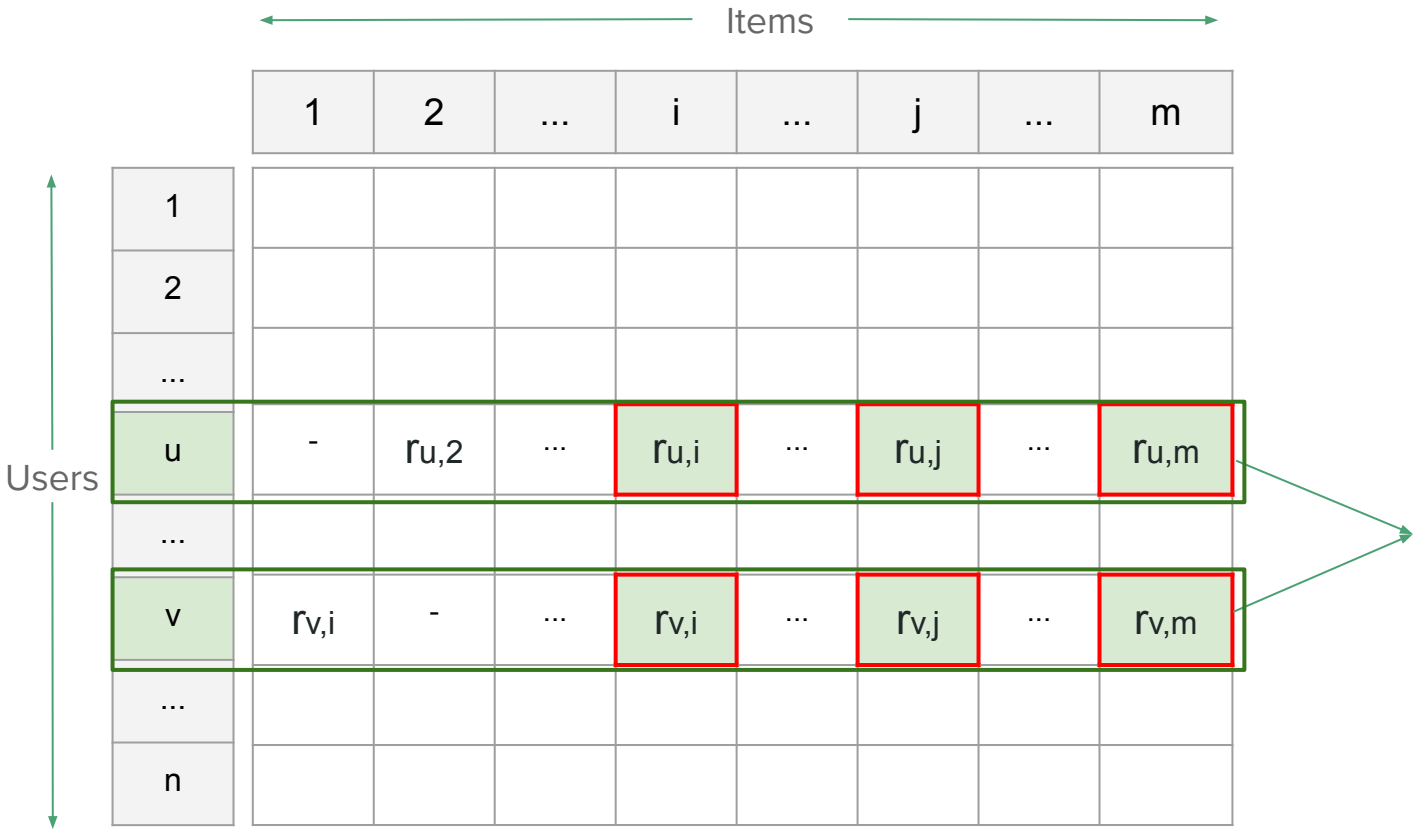
1. Una **matriz de utilidad** (e.g. ratings)
2. Una definición de **similitud** entre usuarios
3. Una definición de **vecindad** (subconjunto más próximo a un usuario)
4. Una **regla** para predecir el rating en base al conjunto de usuarios de la vecindad

Matriz de utilidad

La matriz de utilidad representa una **función de utilidad**:
U: users x items -> rating
Ejemplos rating:

- Binario (1,0)
- Escala (1-5)

Definición de similitud entre usuarios



Ratings de productos

$i \in I$

que ambos usuarios (u,v) han valorado

Definición de similitud entre usuarios

(1) Similitud del coseno (-1,1)

$$sim_{u,v} = \frac{\sum_{i \in I} r_{u,i} r_{v,i}}{\sqrt{\sum_{i \in I} r_{u,i}^2} \sqrt{\sum_{i \in I} r_{v,i}^2}}$$

Coseno:
qué tan parecidos son los gustos en dirección

(2) Coeficiente correlación de Pearson (-1,1)

$$sim_{u,v} = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}}$$

Person:
qué tan parecidos son los gustos independientemente de la escala personal de valoración

I productos evaluados por **u,v**

Intersección!!!

¡Cuidado!

En las medidas de similitud solo se consideran los **ítems que ambos usuarios han valorado** (la **intersección** de sus valoraciones no vacías)

$$I_{uv} = \{ i \in I \mid r_{u,i} \text{ y } r_{v,i} \text{ existen} \}$$

Las medidas de similitud se basan en comparar patrones de valoración comunes. Si un **ítem no fue valorado por ambos**, **no hay información** sobre cómo sus preferencias se relacionan respecto a ese ítem

Incluir valores nulos o tratarlos como ceros **introduce sesgo**. La función **pairwise_distances de SciKit no tiene en cuenta la intersección** lo que sesga las similitudes en datos dispersos

Corolario: si la intersección es **muy pequeña** (por ejemplo, 1 o 2 ítems comunes), la similitud puede ser poco fiable, generalmente se establecen algún tamaño mínimo

Regla para estimar un rating

Filtro colaborativo **basado en usuarios (user-based)**:

$$\hat{r}_{u,i} = \overline{r}_u + \frac{\sum_{v \in U} (r_{v,i} - \overline{r}_v) sim_{u,v}}{\sum_{v \in U} |sim_{u,v}|}$$

U usuarios* que han evaluado **i**

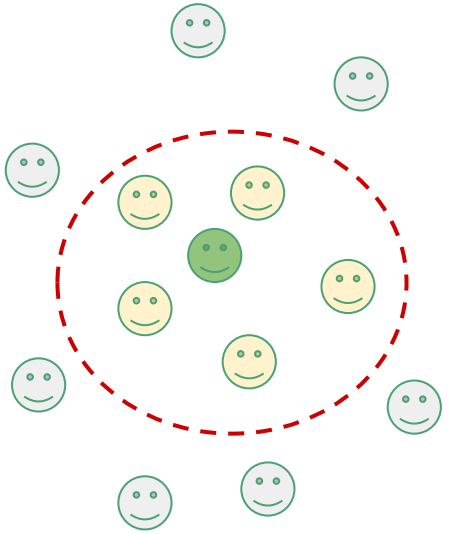
Media ponderada de los ratings (normalizados) de los usuarios que evaluaron i (siendo los pesos la similitud de los usuarios)

* Se suele hacer sobre un subconjunto de usuarios: por ejemplo los **k vecinos más próximos**

Vecindad

En un FC basado en usuarios, la predicción del rating de un usuario u para un ítem i puede calcularse utilizando solo las valoraciones de un **subconjunto (vecindad)** de usuarios más próximos a mi:

- Vecindad fija (**top-k**): por ejemplo, los 10 usuarios más parecidos
- Vecindad con "**threshold**": incluir solo usuarios con similitud > threshold



¿Por qué? Porque si usáramos todos los usuarios:

- Se introduce ruido (usuarios con gustos diferentes distorsionan la media)
- Aumenta el coste computacional
- Podría haber sesgos negativos (si muchos usuarios no han evaluado ítems similares)

Ejemplo

Matriz de utilidad:

	Star Wars VII	Batman v Superman	Ocho apellidos vascos	Marte
UserA	4	NaN	3	5
UserB	NaN	5	4	NaN
UserC	5	4	2	NaN
UserD	2	4	NaN	3
UserE	3	4	5	NaN

Queremos obtener la **predicción del usuario C a la película Marte**

Ejemplo: FC basado en usuarios

$$\hat{r}_{u,i} = \overline{r}_u + \frac{\sum_{v \in U} (r_{v,i} - \overline{r}_v) sim_{u,v}}{\sum_{v \in U} |sim_{u,v}|}$$

- Utilizaremos los ratings de aquellos usuarios U={A,D} que sí han visto Marte
- Necesitaremos los rating medios de C y de los usuarios U={A,D}
- Necesitamos la similitud entre cada usuario de U={A,D} y el usuario C (utilizaremos el coeficiente de correlación de Pearson)

Ejemplo: FC basado en usuarios

	Rating a Marte	Rating medio	Similitud con C
A	5	(4+3+5)/3=4	0.780869
D	3	(2+4+3)/3=3	-0.514496
C		(5+4+2)/3=3.67	

$$\hat{r}_{C,Marte} = 3.67 + \frac{(5-4)0.780869 - (3-3)0.514496}{0.780869+0.514496} = 4.269484$$

Filtro colaborativo basado en productos

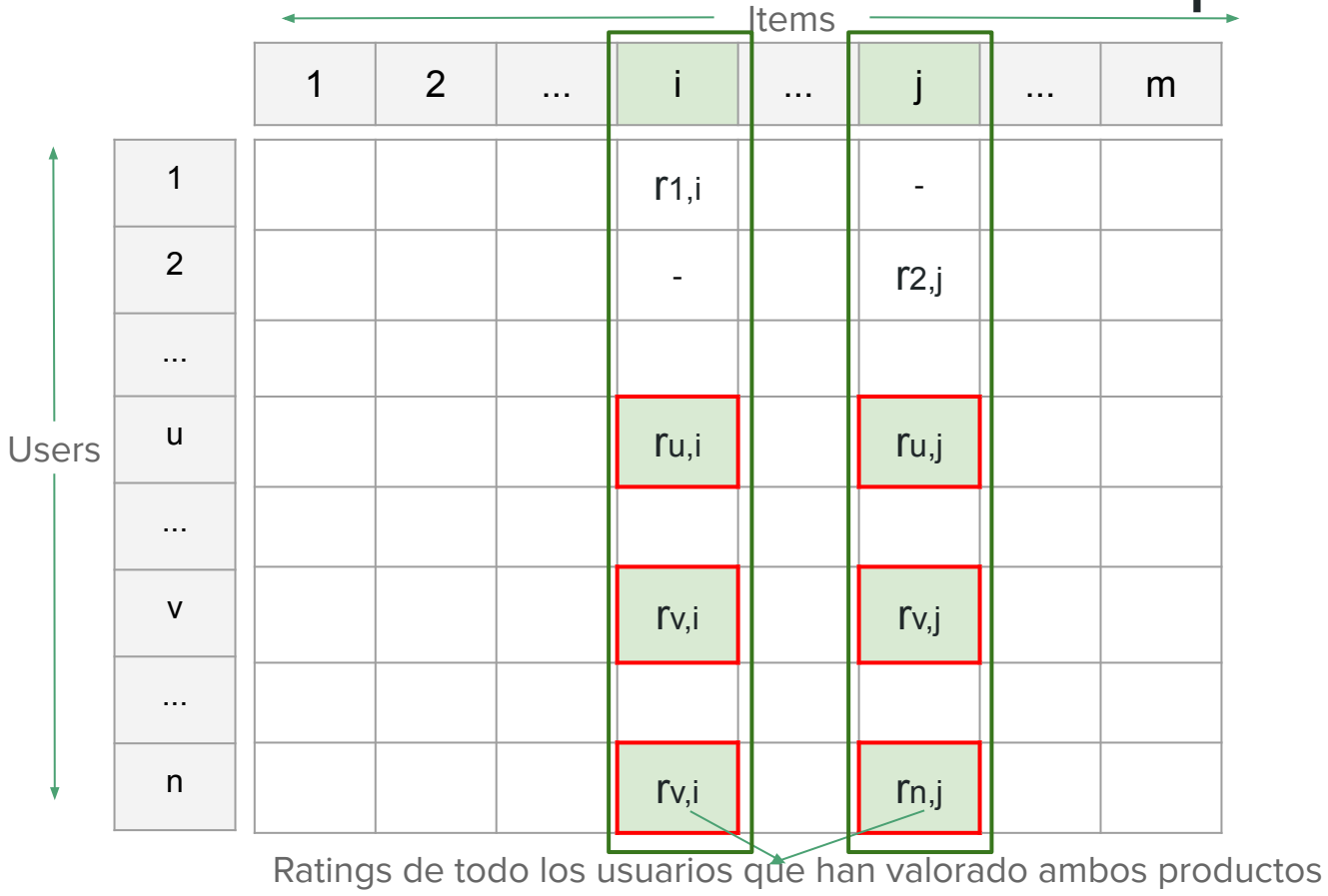
La **hipótesis** de un filtro colaborativo basado en productos es que:

- Si dos productos han sido valorados de forma similar por los usuarios, serán valorados igual por otro usuario

Los FC basados en productos requieren:

1. Una **matriz de utilidad** (e.g. ratings)
2. Una definición de **similitud** entre productos
3. Una definición de **vecindad** (subconjunto más próximo a un producto)
4. Una **regla** para predecir el rating en base al conjunto de productos de la vecindad

Definición de similitud entre productos



Definición de similitud entre productos

(1) Similitud del coseno (-1,1)

$$sim_{i,j} = \frac{\sum_{u \in U} r_{u,i} r_{u,j}}{\sqrt{\sum_{u \in U} r_{u,i}^2} \sqrt{\sum_{u \in U} r_{u,j}^2}}$$

(2) Coeficiente correlación de Pearson (-1,1)

$$sim_{i,j} = \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U} (r_{u,i} - \bar{r}_i)^2} \sqrt{\sum_{u \in U} (r_{u,j} - \bar{r}_j)^2}}$$

U usuarios que han evaluado **i, j**

Regla para estimar un rating

Filtro colaborativo **basado en productos (item-based)**:

$$\hat{r}_{u,i} = \frac{\sum_{j \in I} r_{u,j} sim_{i,j}}{\sum_{j \in I} |sim_{i,j}|}$$

I productos* evaluados por **u**

Media ponderada de los ratings del usuario u a los productos que evaluó (siendo los pesos la similitud de los productos)

* Se suele hacer sobre un subconjunto de productos: por ejemplo los **k vecinos más próximos**

Ejemplo

Matriz de utilidad:

	Star Wars VII	Batman v Superman	Ocho apellidos vascos	Marte
UserA	4	NaN	3	5
UserB	NaN	5	4	NaN
UserC	5	4	2	NaN
UserD	2	4	NaN	3
UserE	3	4	5	NaN

Queremos obtener la **predicción del usuario C a la película Marte**

Ejemplo: FC basado en productos

$$\hat{r}_{u,i} = \frac{\sum_{j \in I} r_{u,j} sim_{i,j}}{\sum_{j \in I} |sim_{i,j}|}$$

- Utilizaremos los ratings del usuario C a aquellas películas que vio I={Star Wars VII,Batman v Superman,Ocho apellidos vascos}
- Necesitamos la similitud entre cada película vista por C y Marte (utilizaremos el coeficiente de correlación de Pearson)

Ejemplo: FC basado en productos

	Star Wars VII	Batman v Superman	Ocho apellidos vascos
Rating de C	5	4	2
Similitud a Marte	0.8944	1	-1

- Podemos considerar todas las películas

$$\hat{r}_{C,Marte} = \frac{(5)0.8944+(4)1+(2)(-1)}{0.8944+1+1} = 2.2360$$

- Podemos considerar un subconjunto de ellas, por ejemplo las 2 más próximas {Star Wars VII,Batman v Superman}

$$\hat{r}_{C,Marte} = \frac{(5)0.8944+(4)1}{0.8944+1} = 4.4721$$

Problemas de filtros basados en memoria

Los Filtros Colaborativos basados en productos son muy utilizados (e.g. Amazon patent)

Pero presentan algunos inconvenientes comunes a otros sistemas de recomendación (Su & Khoshgoftaar, 2009):

- Muchos sistemas tienen datos dispersos lo que dificulta encontrar usuarios o productos similares (sobre los que basar una recomendación).
- Problema de “cold start”: dificultad de hacer recomendaciones a usuarios y/o productos nuevos (para los que no se dispone de información)
- Problemas de escalabilidad con el aumento del número de usuarios y/o productos
- Los sinónimos (dos o más referencias diferentes al mismo producto que sin embargo el filtro trata como productos distintos) introducen ruido en las recomendaciones
- “Gray sheep” y “black sheep”, cuando algunos usuarios tienen un patrón alejado del resto (o incluso único) el filtro colaborativo no hace buenas recomendaciones

Referencias

- Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. Foundations and Trends in Human-Computer Interaction, 4(2), 81-173.
- Collaborative recommendations using item-to-item similarity mappings (Amazon patent)