

¿Qué recomendaciones te presenta Amazon?

amazon.com

amazon.com

Recommended for You

Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



[The Little Big Things: 163 Ways to Pursue EXCELLENCE](#)



[Fascinate: Your 7 Triggers to Persuasion and Captivation](#)



[Sherlock Holmes \[Blu-ray\]](#)



[Alice in Wonderland \[Blu-ray\]](#)

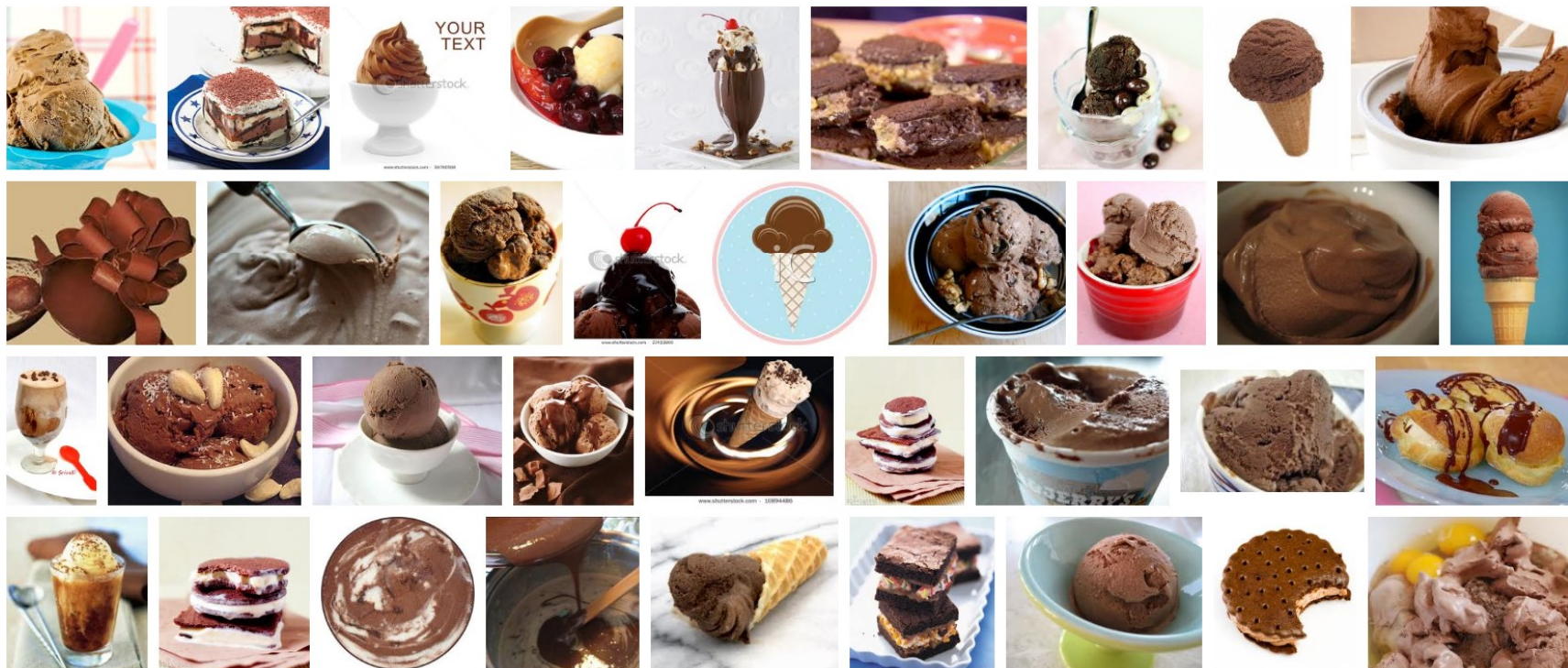
¿Qué recomendaciones te presenta TripAdvisor?



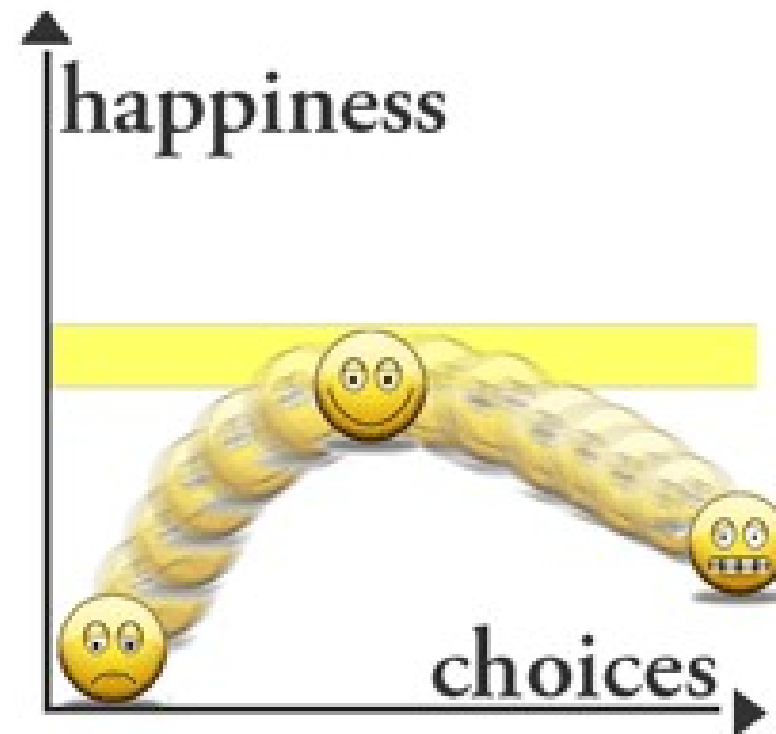
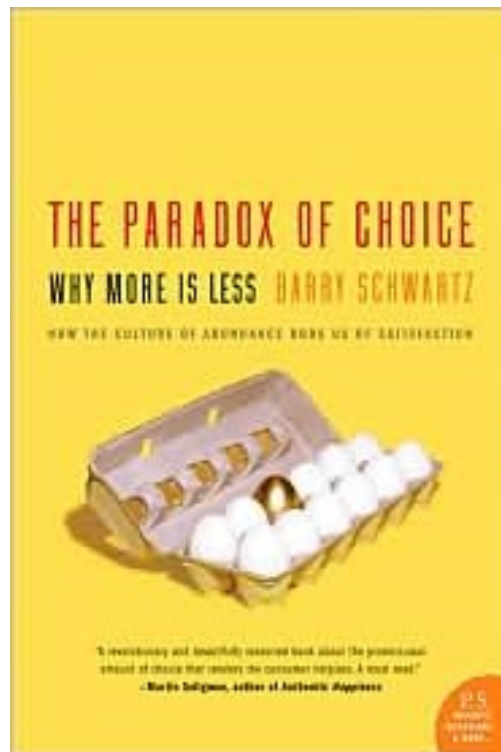
El problema de la elección



El problema de la elección



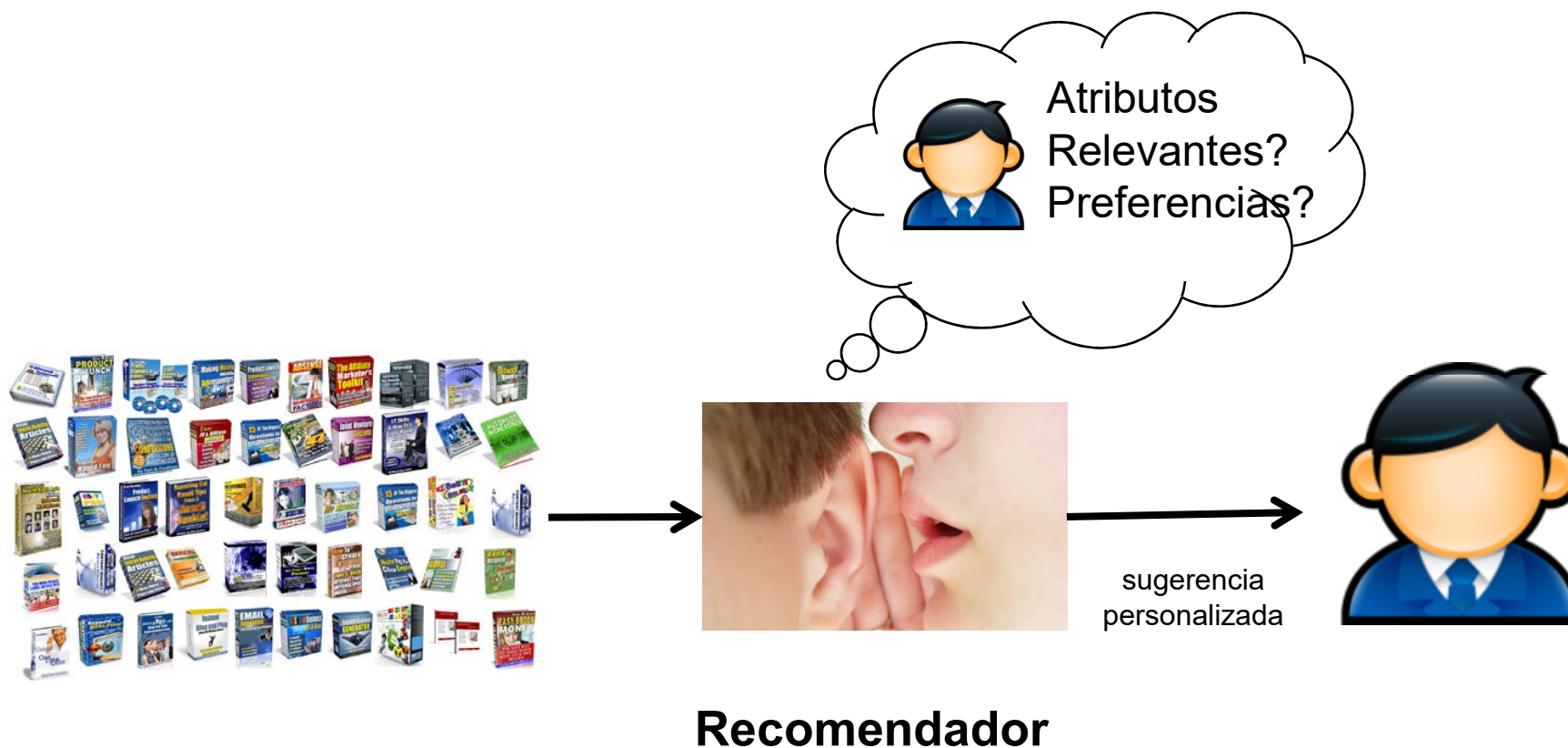
El problema de la elección



¿Qué son las recomendaciones?

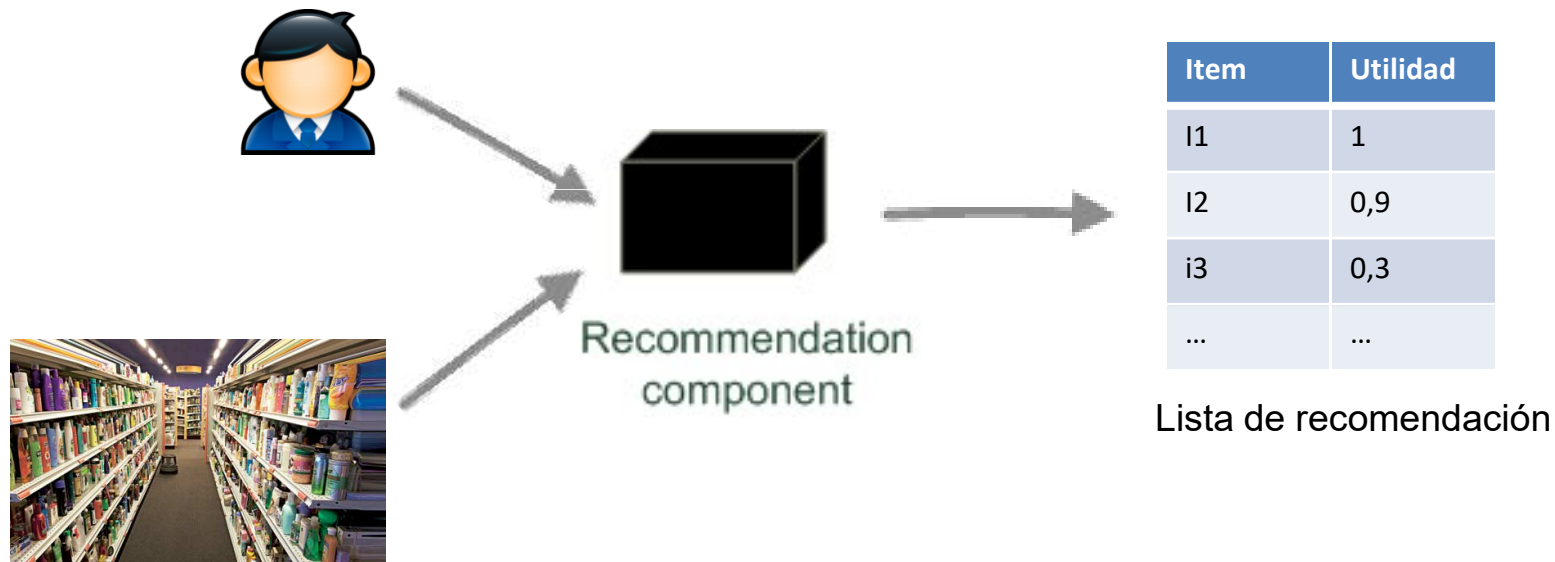


¿Qué son las recomendaciones?



¿Qué son las recomendaciones?

Problema: Predecir la utilidad de un item para un usuario!!!



Adaptado de: Recommender Systems: An introduction
(Cambridge University Press)

Recomendación Vs Búsqueda



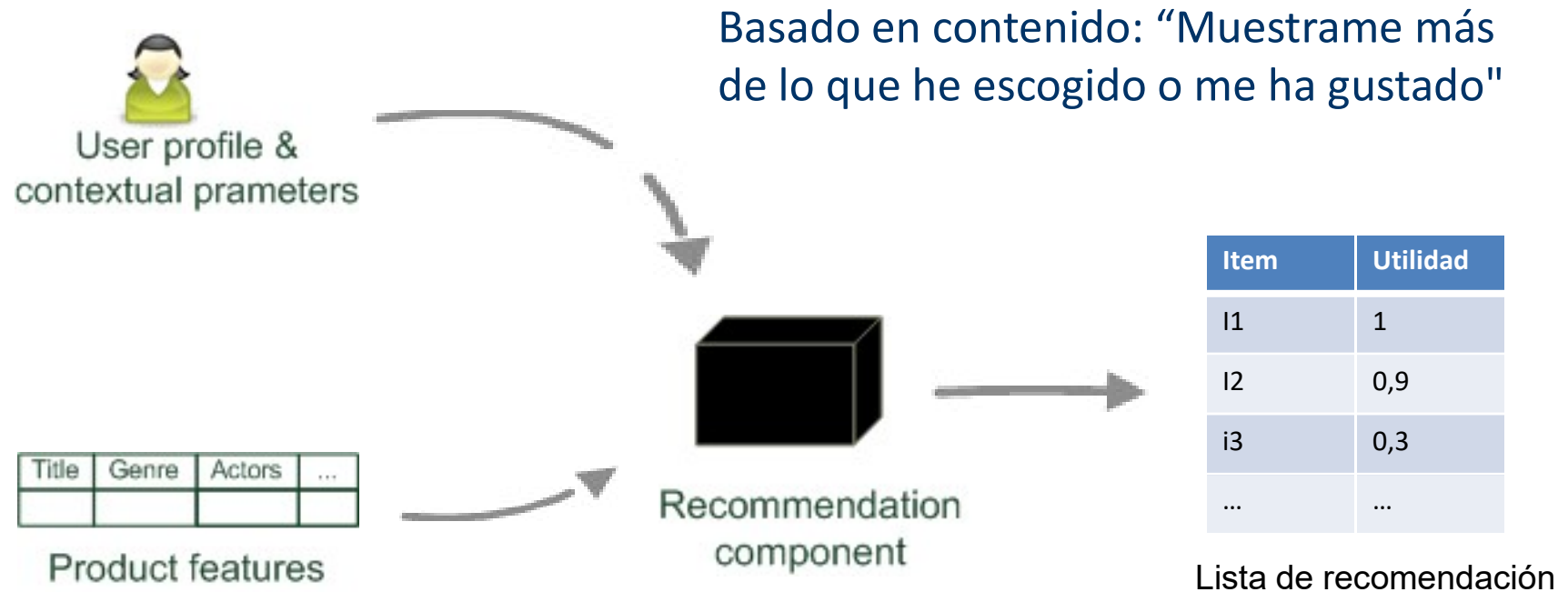
consulta



respuesta

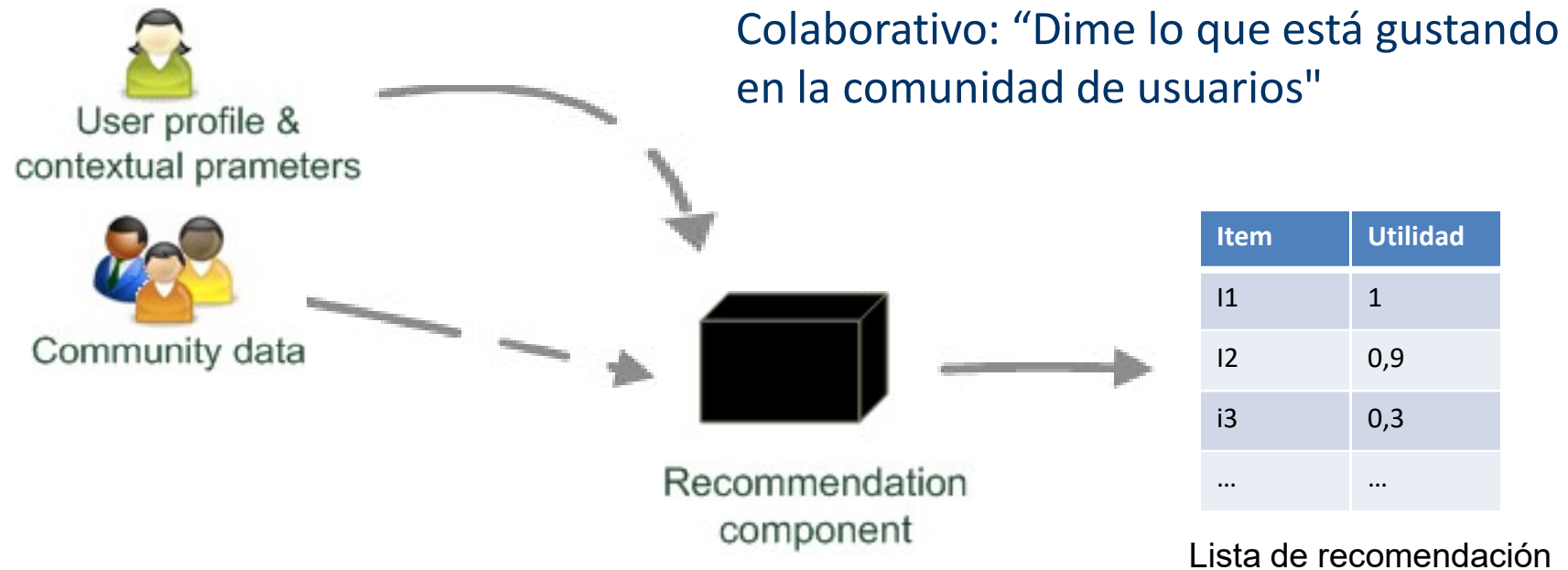


Estrategias de recomendación



Adaptado de: Recommender Systems: An introduction
(Cambridge University Press)

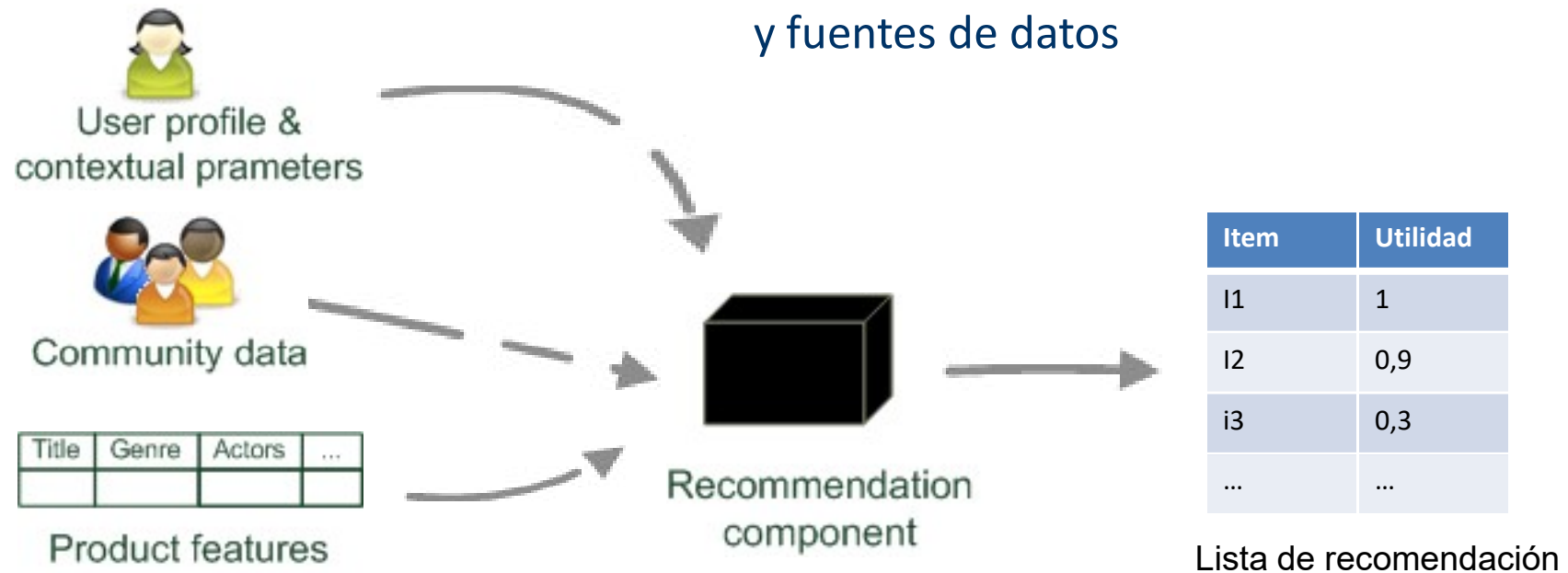
Estrategias de recomendación



Adaptado de: Recommender Systems: An introduction
(Cambridge University Press)

Estrategias de recomendación

Híbrido: combinación de varias estrategias
y fuentes de datos



Adaptado de: Recommender Systems: An introduction
(Cambridge University Press)

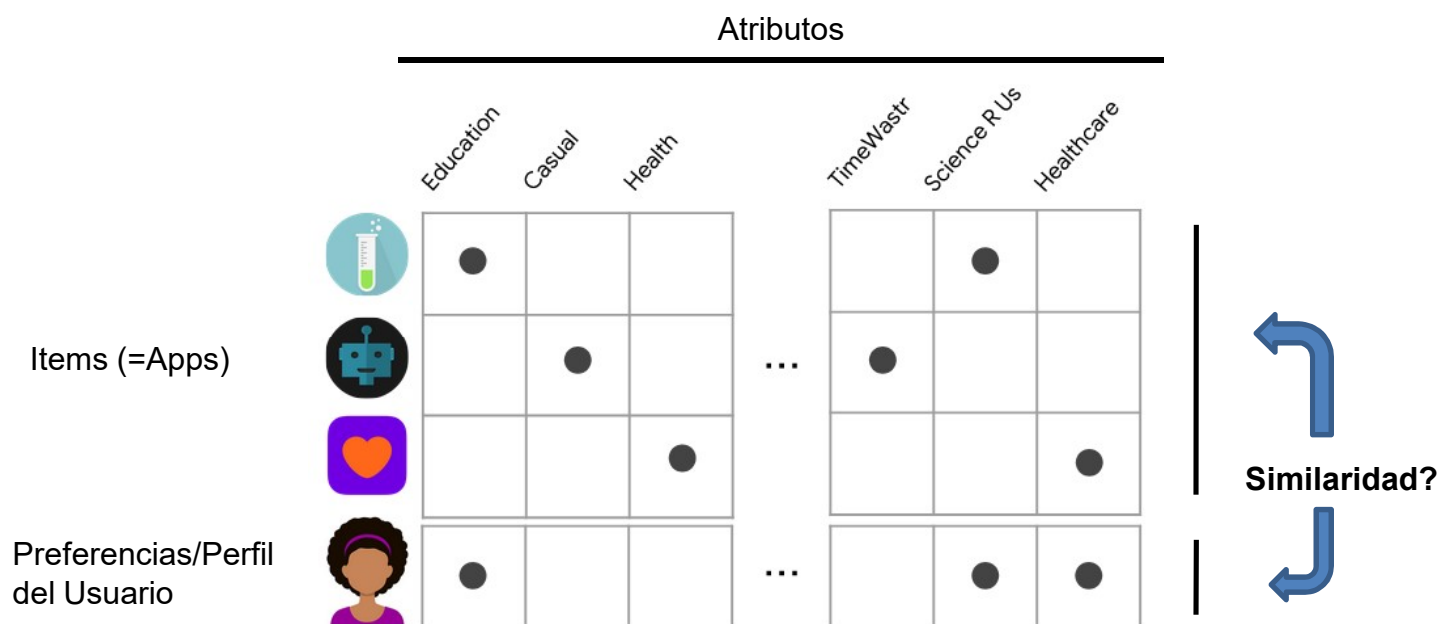
- Objetivo:

Recomendar items al usuario-objetivo similares a aquellos items que el usuario ha escogido, comprado o mostrado algún tipo de preferencia en el pasado.

- ¿Qué necesitamos?:

1. **Caracterizar el contenido** (items/alternativas): Información sobre los *atributos del contenido*. Ejemplo: el género de las películas.
2. **Aprender/estimar las preferencias del usuario.** Las preferencias representan la importancia/relevancia que tiene para el usuario cada uno de los atributos del contenido. El conjunto de preferencias constituye el *perfil del usuario*.
3. **Calcular la similaridad.** Hay que calcular la similaridad entre el perfil del usuario y cada item o alternativa que esté disponible para recomendar.
4. **Asumir utilidad = similaridad.** Vamos a aceptar el supuesto de que la utilidad de un item para un usuario viene dada por la similaridad entre el item y el perfil del usuario

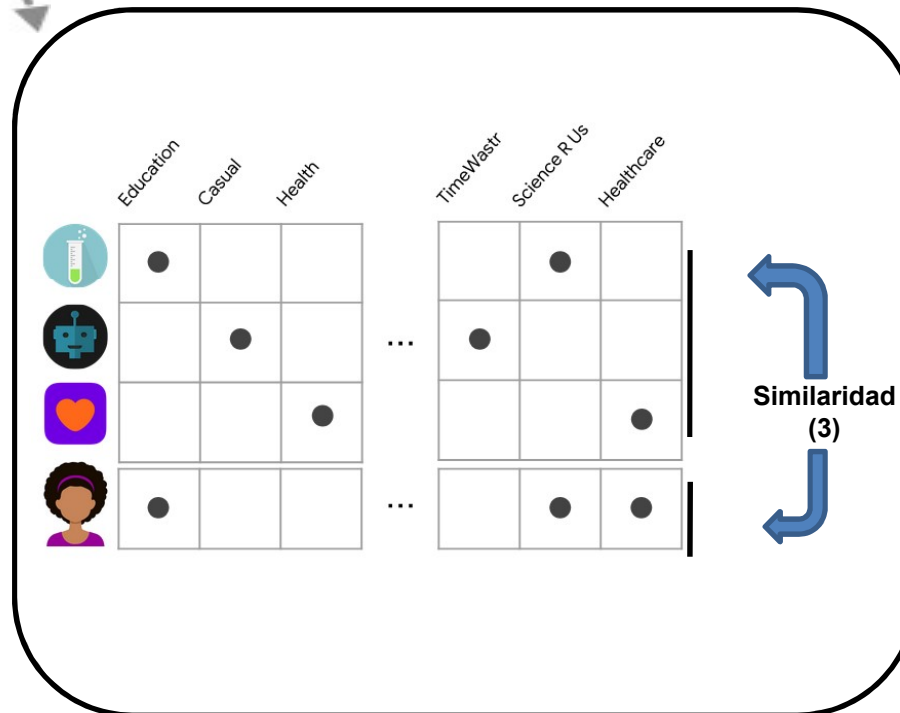
Recomendación basada en contenido



Recomendación basada en contenido



Preferencias/Perfil
del usuario
(2)



Education	Casual	Health	...

Atributos del contenido
(1)

Item	Utilidad
I1	1
I2	0,9
i3	0,3
...	...

Utilidad = Similitud
(4)

Caracterizando el contenido



- Contenidos. Ejemplo: Libros

- Atributos de un libro:
 1. Precio (variable numérica) - Valores: entre 0 y N euros
 2. Género (variable discreta) - Valores: {ficción, histórico, novela...}

- Binarización de los atributos:
 1. Precio: 0-15 euros -> "Precio Bajo" – Valores: {0,1}
 2. Precio: >15 euros -> "Precio Alto" – Valores: {0,1}
 3. Género: ficción -> "Género ficción" – Valores: {0,1}
 4. Género: histórico -> "Género histórico" – Valores: {0,1}

Caracterizando el contenido

Atributos

Items	Producto	Precio	Género
	Libro1	15	Ficción
	Libro2	50	Histórico

	LibroN	10	Ficción

Atributos

Items	Producto	Precio Bajo	Precio Alto	Género ficción	Género histórico
	Libro1	1	0	1	0
	Libro2	0	1	0	1

	LibroN	1	0	1	0

Caracterizando el contenido

Atributos

Items	Producto	Precio Bajo	Precio Alto	Género ficción	Género histórico
	Libro1	1	0	1	0
	Libro2	0	1	0	1

	LibroN	1	0	1	0

↓ Representación como vector

Vector Libro1 = (1, 0, 1, 0)

Formalizando para todo item:

Vector del
Item = ($x_{11}, x_{21}, \dots, x_{1K}, x_{2K} \dots$)

donde x es una variable binaria y k indica el número del atributo

Aprendiendo las preferencias/perfil del usuario



Conjunto de datos del individuo c (dataset)

		Atributos				
Items	Producto	Autor	Fecha	Precio	Género	Acción usuario
	Libro1	---	---	15	Ficción	Comprado
	Libro2	---	---	50	Histórico	Comprado

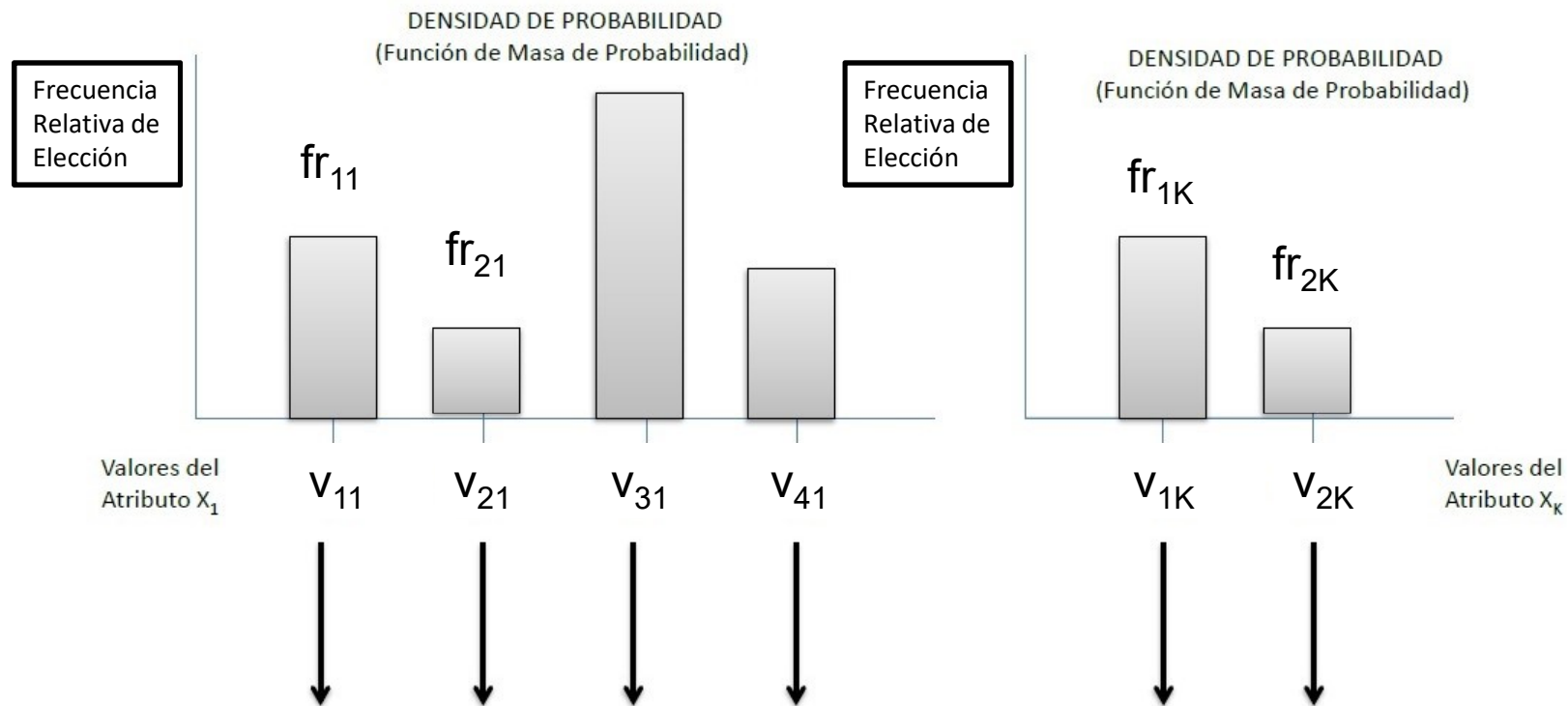
	LibroN	---	---	10	Ficción	Comprado

¿Cuáles son las preferencias del individuo c?

Frecuencia de Elección de cada valor = Veces que un valor ha sido "Comprado"

$$\text{Frecuencia Relativa de Elección de cada valor} = \frac{\text{Frecuencia de Elección de cada valor}}{\text{Número total de compras}}$$

Aprendiendo las preferencias/perfil del usuario



Vector de Preferencias = (β_{11} β_{2K})

Aprendiendo las preferencias/perfil del usuario

Conjunto de datos del usuario c (dataset)

Atributos

Items

Producto	Autor	Fecha	Precio	Género	Acción usuario
Libro1	---	---	10	Ficción	Comprado
Libro2	---	---	50	Histórico	Comprado
....
LibroN	---	---	50	Ficción	Visto



Producto	Autor	Fecha	Precio	Género	Utilidad
Libro1	---	---	10	Ficción	10
Libro2	---	---	50	Histórico	10
LibroN	---	---	50	Ficción	5

Aprendiendo las preferencias/perfil del usuario



Conjunto de datos del usuario c (dataset)

Atributos

Items	Producto	Autor	Fecha	Precio	Género	Utilidad
	Libro1	---	---	10	Ficción	10
	Libro2	---	---	50	Histórico	10
	LibroN	---	---	50	Ficción	5

¿Cuáles son las preferencias del individuo c?

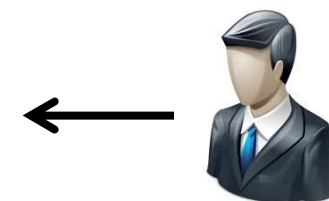
$$\text{Utilidad media de cada valor} = \frac{1}{\text{Número de compras con ese valor}} \sum \text{utilidad de la compra con ese valor}$$

Aprendiendo las preferencias/perfil del usuario

Conjunto de datos del usuario c (dataset)

Atributos

Producto	Autor	Fecha	Precio	Género	Utilidad
Libro1	---	---	10	Ficción	10
Libro2	---	---	50	Histórico	9
Libro3	---	---	---	---	
LibroN	---	---	30	Novela	3

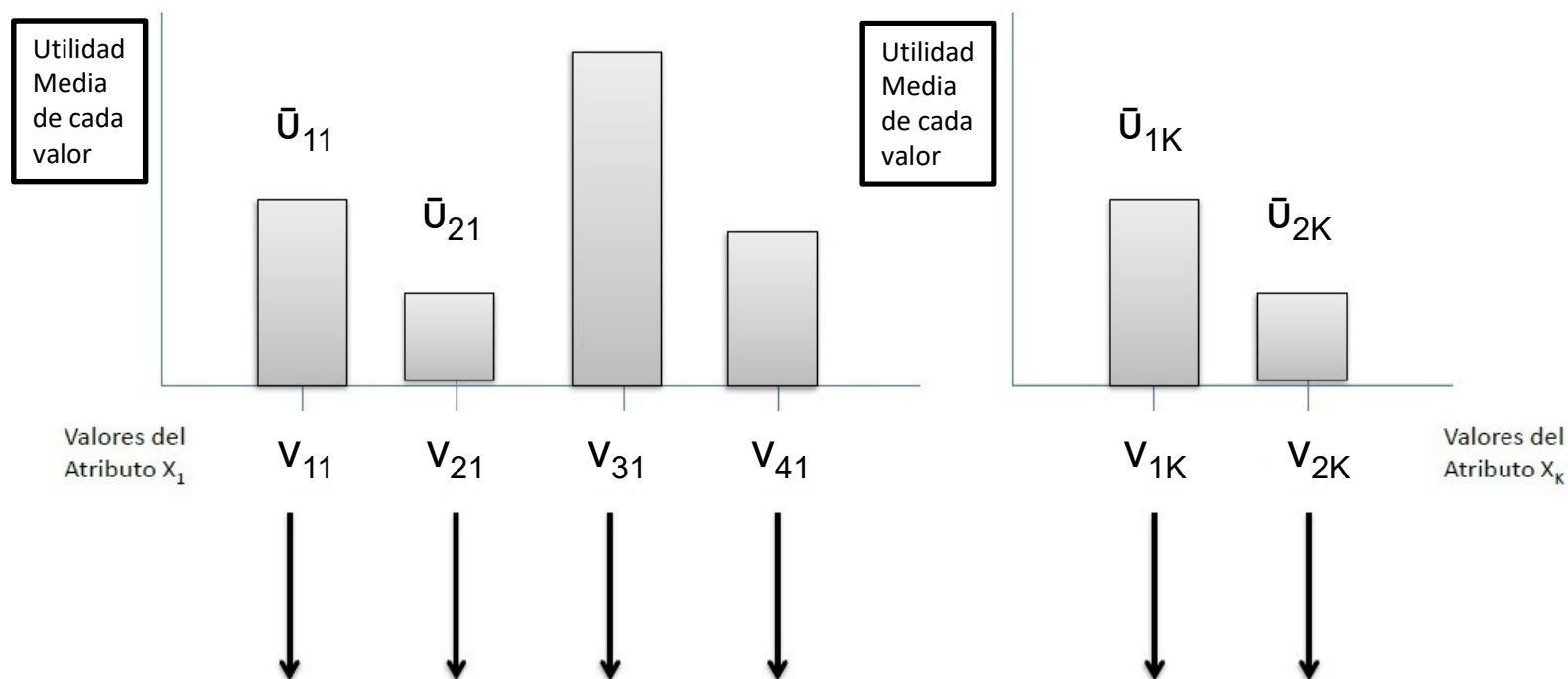


Cuestionario
para conocer
la utilidad

Técnica: Conjoint Analysis

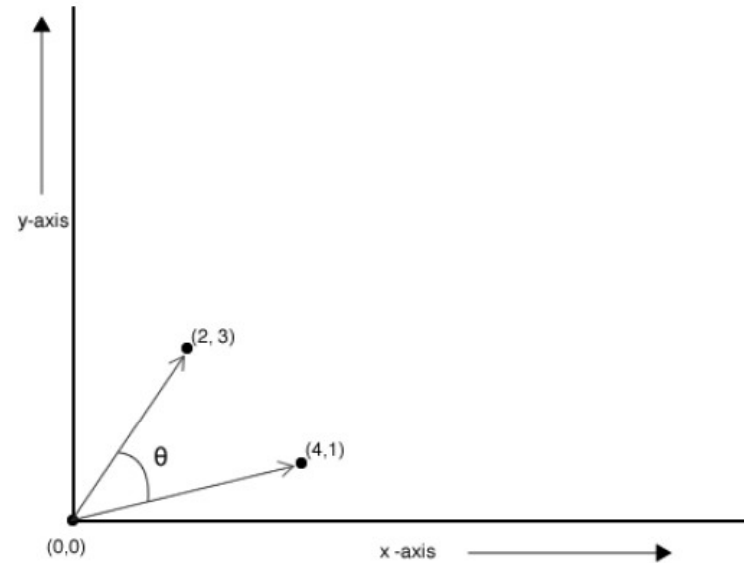
$$\text{Utilidad media de cada valor} = \frac{1}{\text{Número de compras con ese valor}} \sum \text{utilidad de la compra con ese valor}$$

Aprendiendo las preferencias/perfil del usuario



Vector de Preferencias = (β_{11} β_{2K})

Calculando la similaridad

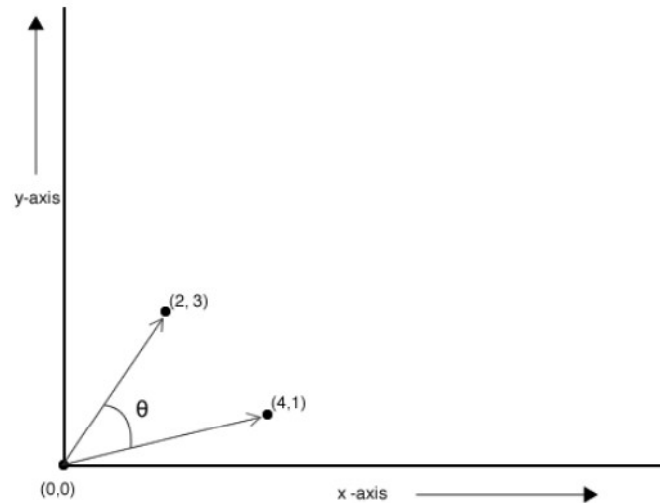


Medida del coseno:

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

Similaridad(Perfil Usuario, Item) = Coseno(Vector Preferencias, Vector Item)

Calculando la similaridad



Distancia Euclidea entre dos vectores:

$$\delta(a, b) = \|a - b\| = \sqrt{(a - b)^T (a - b)} = \sqrt{\sum_{i=1}^m (a_i - b_i)^2}$$

Similaridad(Perfil Usuario, Item) = 1 / (1 + distancia(Vector Preferencias, Vector Item))

Creando el ranking de los items



1. Asumimos: Utilidad = Similaridad
2. Creamos un ranking de los items ordenándolos de mayor a menor valor de utilidad

Item	Utilidad
I1	1
I2	0,9
i3	0,3
...	...

3. Generamos las recomendaciones: Seleccionamos los N primeros items (Top-N recommendations) .

Variaciones: contenido descrito con Keywords

- Most CB-recommendation techniques were applied to recommending text documents.
 - Like web pages or newsgroup messages for example.
- Content of items can also be represented as text documents.
 - With textual descriptions of their basic characteristics.
 - Structured: Each item is described by the same set of attributes



Title	Genre	Author	Type	Price	Keywords
The Night of the Gun	Memoir	David Carr	Paperback	29.90	Press and journalism, drug addiction, personal memoirs, New York
The Lace Reader	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
Into the Fire	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism

- Unstructured: free-text description.

Variaciones: contenido descrito con Keywords

■ Item representation

Title	Genre	Author	Type	Price	Keywords
The Night of the Gun	Memoir	David Carr	Paperback	29.90	Press and journalism, drug addiction, personal memoirs, New York
The Lace Reader	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
Into the Fire	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism

■ User profile

Title	Genre	Author	Type	Price	Keywords
...	Fiction	Brunonia, Barry, Ken Follett	Paperback	25.65	Detective, murder, New York

$keywords(b_j)$
describes Book b_j
with a set of
keywords

• Simple approach

- Compute the similarity of an unseen item with the user profile based on the keyword overlap (e.g. using the Dice coefficient)
- Or use and combine multiple metrics



$$\frac{2 \times |keywords(b_i) \cap keywords(b_j)|}{|keywords(b_i)| + |keywords(b_j)|}$$



Conexión con teorías de toma de decisiones



- ▶ En la teoría de la elección racional, la utilidad es una función lineal sobre los valores de los atributos. Dado individuo c y un ítem a :

$$u(c, a) = \sum_k \beta_{c,k} x_{a,k}$$

Con x indicando los valores del atributo k del ítem a

Con β indicando la preferencia del individuo c sobre x

Con K indicando el cto de todos los valores de los atributos de a

- ▶ En la estrategia basada en contenido, la utilidad se puede calcular a través de la similaridad, y ésta a través de la medida del coseno:

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

- ▶ Por tanto:

utilidad(estrategia_basada_contenido) = utilidad(eleccion_racional) normalizada

- ▶ ¿Qué significa esta conexión?

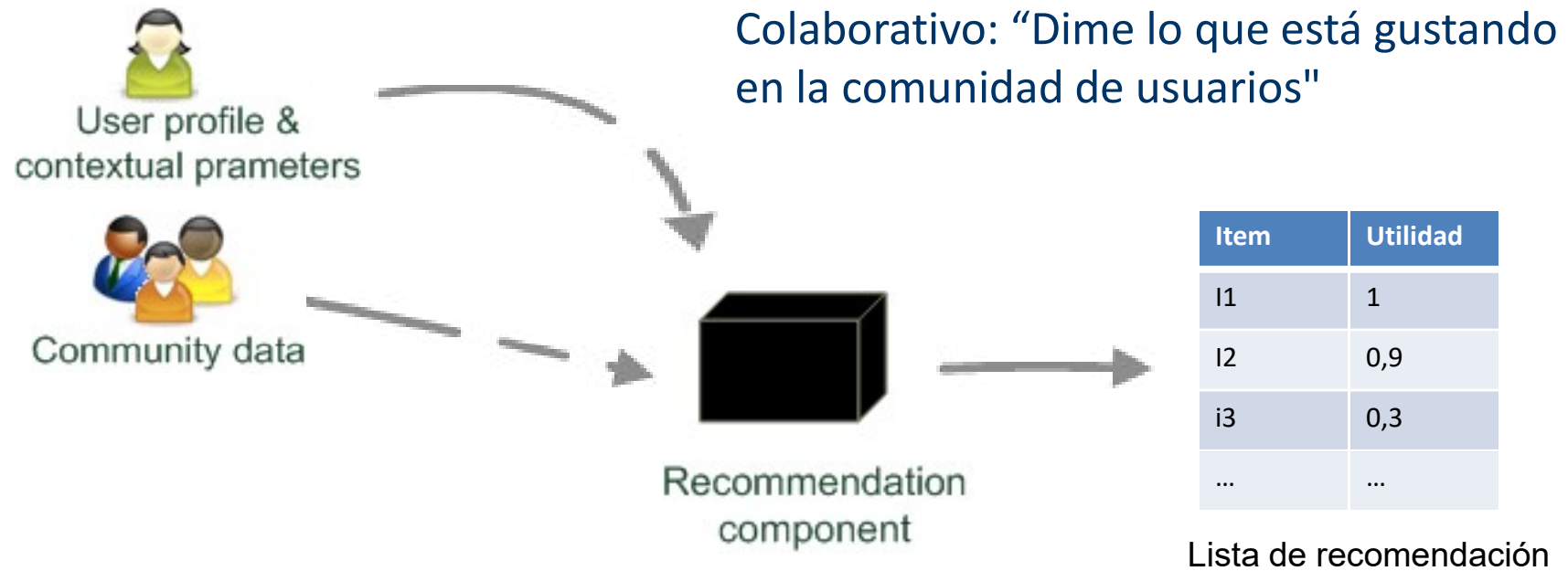
La estrategia de recomendación basada en contenido asume que la toma de decisiones de los humanos es de tipo racional

- Resumen de la estrategia:
 1. Representar los items en formato vector.
 2. Representar las preferencias del usuario en formato vector
 3. Calcular la similaridad.
 4. Asumir utilidad = similaridad.
 5. Crear un ranking de items basados en el valor de utilidad
 6. Recomendar los N primeros.

Limitaciones:

1. **Para el usuario:** Poca originalidad de las recomendaciones. Las recomendaciones suelen ser productos muy similares a los ya consumidos por el usuario.
2. **Para el ingeniero/científico:** Necesidad de conocer en detalle el dominio de la aplicación: productos, atributos y valores
3. **En la satisfacción con la recomendación:** Los valores de los atributos no aportan información acerca de la calidad del producto. Estimar la utilidad de un producto solamente a través de sus atributos no garantiza la satisfacción de la recomendación.
4. **En la hipótesis/supuestos en los que se basa la estrategia:** Asume que la toma de decisiones de los humanos se basa en una estrategia racional. ¿Es esto correcto?

Recomendación colaborativa



- Ejemplo
 - Un dataset simple con las valoraciones tanto del usuario-objetivo, Alice, como de otros usuarios.
 - Ejercicio: ¿Predicción del rating de Alice sobre el Item5?

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

- **Objetivo:**

Recomendar items al usuario-objetivo en base a las valoraciones/ratings de otros usuarios sobre los items disponibles.

- **¿Qué necesitamos?:**

- 1. Valoraciones/ratings del contenido** (items/alternativas): Información de un conjunto de usuarios sobre su satisfacción con el item/alternativa.
- 2. Predecir la valoración del item por el usuario-objetivo.** Se predice la valoración con una media ponderada de las valoraciones de los otros usuarios
- 3. Asumir utilidad = valoración predicha.**

Recomendación colaborativa

- Es la estrategia de recomendación más popular
 - Utilizada por grandes plataformas comerciales de e-commerce
 - Aplicable en cualquier ámbito ya que no depende del contenido (libros, películas, música ..)
- Aproximación
 - Utiliza la "wisdom of the crowds" para predecir las valoraciones
- Supuestos
 - Los usuarios mantienen sus gustos constantes.
 - Los usuarios que valoran de forma similar tienen gustos similares



- **Objetivo:**

Recomendar items al usuario-objetivo en base a las valoraciones aportadas por usuarios similares a él.

- **¿Qué necesitamos?:**

1. **Valoraciones/ratings del contenido** (items/alternativas): Información de un conjunto de usuarios sobre su satisfacción con el item/alternativa.
2. **Estimar la similaridad entre usuarios.** La similaridad entre el usuario-objetivo y otro usuario que ha valorado el item/alternativa se calcula comparando si las valoraciones realizadas por ambos sobre los mismos items son similares o no. *Se asume que dos usuarios tienen gustos similares si sus valoraciones también son similares.*
3. **Predecir la valoración del item por el usuario-objetivo.** Se predice la valoración con una media ponderada de las valoraciones de los otros usuarios
4. **Asumir utilidad = valoración predicha.**

Basada en usuarios

- Datos que se necesitan
 - Una matriz de valoraciones usuario-item

		<i>Items</i>					
		<i>1</i>	<i>2</i>	...	<i>i</i>	...	<i>m</i>
<i>Users</i>	<i>1</i>	5	3		1	2	
	<i>2</i>		2				4
	:			5			
	<i>u</i>	3	4		2	1	
	:					4	
	<i>n</i>			3	2		
<i>a</i>		3	5		?	1	

- Ejemplo
 - Un dataset simple con las valoraciones tanto del usuario-objetivo, Alice, como de otros usuarios.
 - Ejercicio: ¿Predicción del rating de Alice sobre el Item5?

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

- Calculando la similaridad entre usuarios

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

- **A popular similarity measure in user-based CF: Pearson correlation**

a, b : users

$r_{a,p}$: rating of user a for item p

P : set of items, rated both by a and b

- Possible similarity values between -1 and 1

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

- Ejercicio: Calcula la similaridad utilizando la correlación de Pearson y los datos anteriores

- **A popular similarity measure in user-based CF: Pearson correlation**


a, b : users

$r_{a,p}$: rating of user a for item p

P : set of items, rated both by a and b

- Possible similarity values between -1 and 1

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1



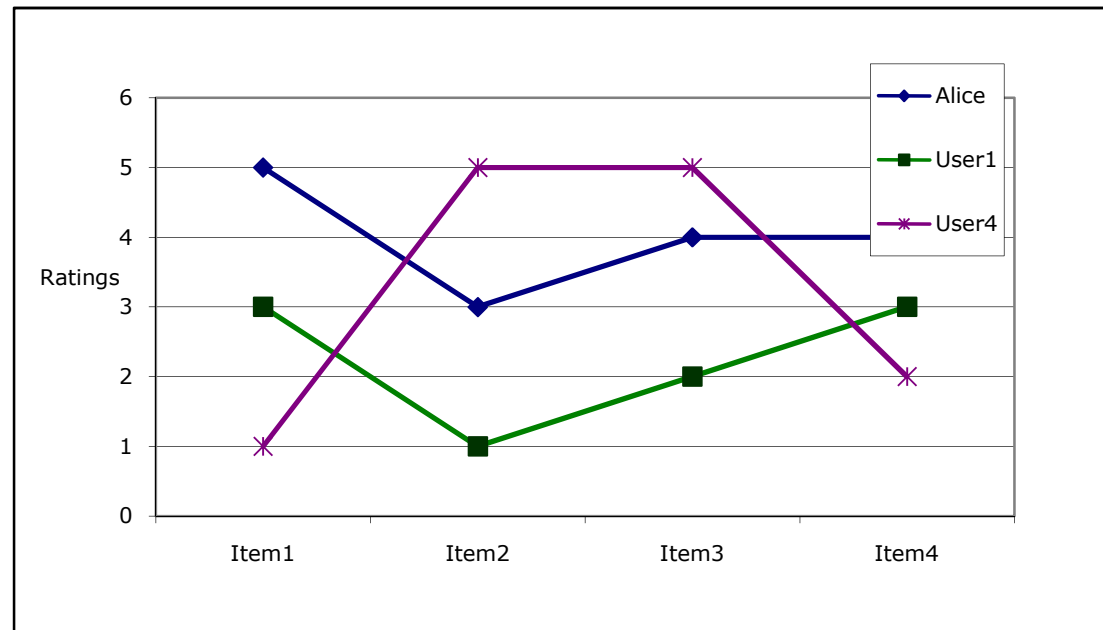
sim = 0,85

sim = 0,70

sim = 0

sim = -0,79

- Ejercicio: ¿Otras formas de calcular la similaridad?
- Compara los resultados con la similaridad obtenida por correlación de Pearson



- Predicción de la valoración con una media ponderada

$$p_{a,i} = \frac{\sum_{j \in K} r_{a,j} w_{i,j}}{\sum_{j \in K} |w_{i,j}|}$$

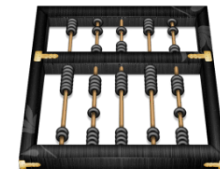
con $p_{a,i}$ la predicción del usuario-objetivo i sobre el ítem a

con $r_{a,j}$ la valoración del usuario j sobre el ítem a

y con $w_{i,j}$ la similaridad entre el usuario-objetivo i y el usuario j

- A common prediction function:

$$pred(a, p) = \bar{r}_a + \frac{\sum_{b \in N} sim(a, b) * (r_{b,p} - \bar{r}_b)}{\sum_{b \in N} sim(a, b)}$$



- Calculate, whether the neighbors' ratings for the unseen item i are higher or lower than their average
- Combine the rating differences – use the similarity with a as a weight
- Add/subtract the neighbors' bias from the active user's average and use this as a prediction

- No todas las valoraciones aportan la misma información
 - Coincidir en la valoración de items con opiniones diversas es más informativo que coincidir en la valoración de items que gustan a todos.
 - **Posible solución:** Dar más peso en la predicción a los items con mayor varianza en sus valoraciones
- Valorar el número de items valorados conjuntamente entre dos usuarios
 - El peso de un usuario en la predicción debe estar correlacionado con el número de items valorados conjuntamente.
- Selección del vecindario (número de usuarios similares)
 - Utilizar un umbral mínimo de similaridad para seleccionar el número de usuarios similares

- **Objetivo:**

Recomendar items al usuario-objetivo en base a las valoraciones realizadas sobre otros items similares a los items a recomendar.

- **¿Qué necesitamos?:**

1. **Valoraciones/ratings del contenido** (items/alternativas): Información de un conjunto de usuarios sobre su satisfacción con el item/alternativa.
2. **Estimar la similitud entre items.** La similitud entre el item-objetivo y otro item valorado por los usuarios se calcula comparando si las valoraciones realizadas por ambos usuarios son similares o no. *Se asume que dos items tienen características similares si sus valoraciones también son similares.*
3. **Predecir la valoración del item por el usuario-objetivo.** Se predice la valoración con una media ponderada de las valoraciones de los otros usuarios
4. **Asumir utilidad = valoración predicha.**

Basada en items

- Ejemplo
 - Un dataset simple con las valoraciones tanto del usuario-objetivo, Alice, como de otros usuarios.
 - Ejercicio: ¿Predicción del rating de Alice sobre el Item5?

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

- **A popular similarity measure in user-based CF: Pearson correlation**

a, b : users

$r_{a,p}$: rating of user a for item p

P : set of items, rated both by a and b

– Possible similarity values between -1 and 1

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

- Produces better results in item-to-item filtering
- Ratings are seen as vector in n-dimensional space
- Similarity is calculated based on the angle between the vectors

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



- **Adjusted cosine similarity**
 - take average user ratings into account, transform the original ratings
 - U : set of users who have rated both items a and b

- Predicción de la valoración con una media ponderada

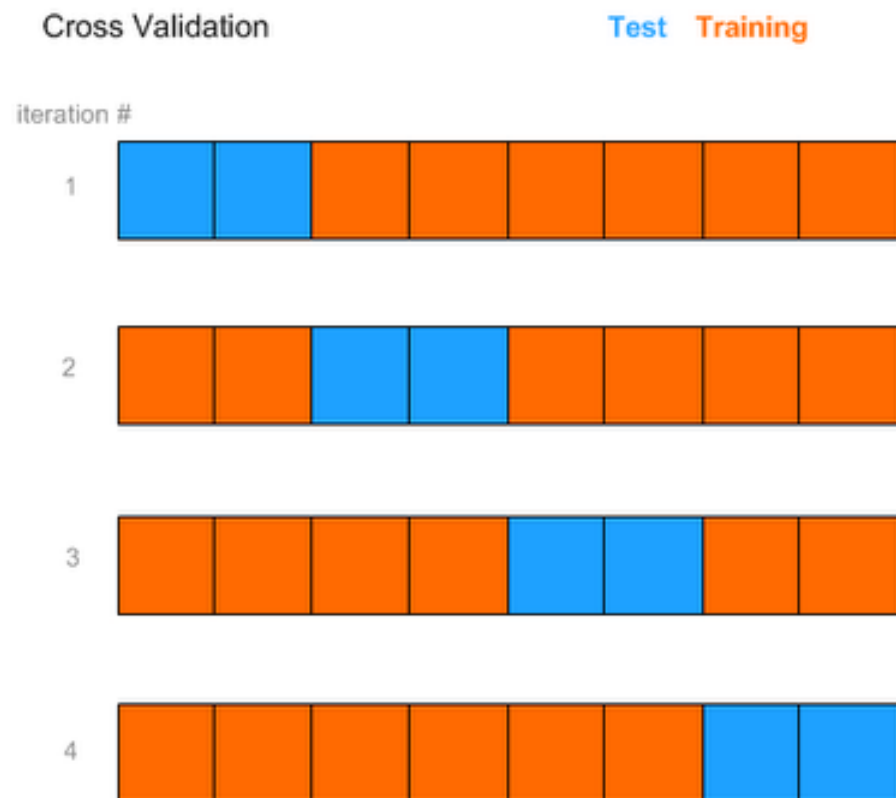
$$p_{a,i} = \frac{\sum_{j \in K} r_{a,j} w_{i,j}}{\sum_{j \in K} |w_{i,j}|}$$

con $p_{a,i}$ la predicción del usuario-objetivo i sobre el item a

con $r_{a,j}$ la valoración del usuario j sobre el item a

y con $w_{i,j}$ la similaridad entre el usuario-objetivo i y el usuario j

Validación de algoritmos



1. Validación cruzada con k subgrupos
2. Validación cruzada con 2 subgrupos
3. Validación dejando una instancia fuera (Leave one out)

MEAN SQUARED ERROR
(Error cuadrático medio)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2$$

MEAN ABSOLUTE ERROR
(Error absoluto medio)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |f_i - y_i| = \frac{1}{n} \sum_{i=1}^n |e_i|.$$

Evaluación de recomendaciones Top-N

Items Recomendados	Items que realmente han gustado (o que han sido comprados) por el usuario
i2	i2
i5	i5
i8	i8
i10	i7

Comparación entre predicción y realidad en un catálogo de 10 items

Evaluación de recomendaciones Top-N

Table 2: 2x2 confusion matrix

actual / predicted	negative	positive
negative	a	b
positive	c	d

Matriz de confusión para el ejemplo

Real/Predicción	Negativo	Positivo
Negativo	5	1
Positivo	1	3

Evaluación de recomendaciones Top-N

$$Accuracy = \frac{\text{correct recommendations}}{\text{total possible recommendations}} = \frac{a + d}{a + b + c + d}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |\epsilon_i| = \frac{b + c}{a + b + c + d},$$

$$Precision = \frac{\text{correctly recommended items}}{\text{total recommended items}} = \frac{d}{b + d}$$

$$Recall = \frac{\text{correctly recommended items}}{\text{total useful recommendations}} = \frac{d}{c + d}$$

Métricas de rendimiento para el ejemplo

Accuracy	MAE	Precision	Recall
8/10= 0,8	2/10=0,2	3/4=0,75	3/4=0,75

Problemas: la calidad de los datos

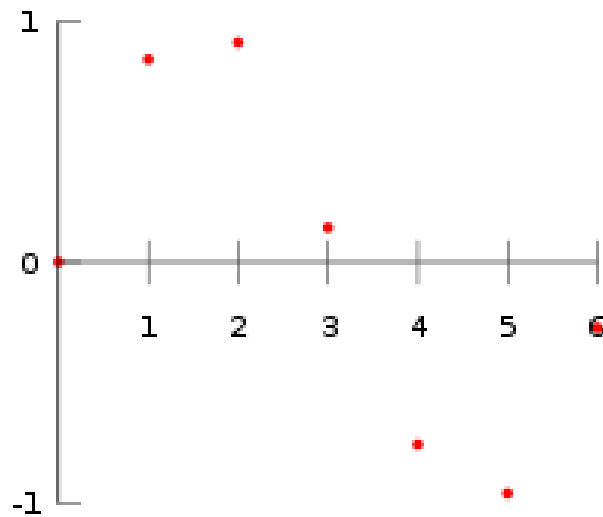
Training Set			
Id	Status	Age	Class
1	Single	20	Bad
2	Single	30	Good
3	Single	50	Bad
4	Single	60	Good
5	Married	20	Good
6	Married	30	Good
7	Married	40	Good
8	Married	50	Good
9	Divorced	40	Bad
10	Divorced	60	Good
Testing Set			
11	Single	40	(Bad)
12	Married	60	(Good)
13	Divorced	20	(Bad)
14	Divorced	30	(Bad)
15	Divorced	50	(Good)

1. Exactitud de los datos?

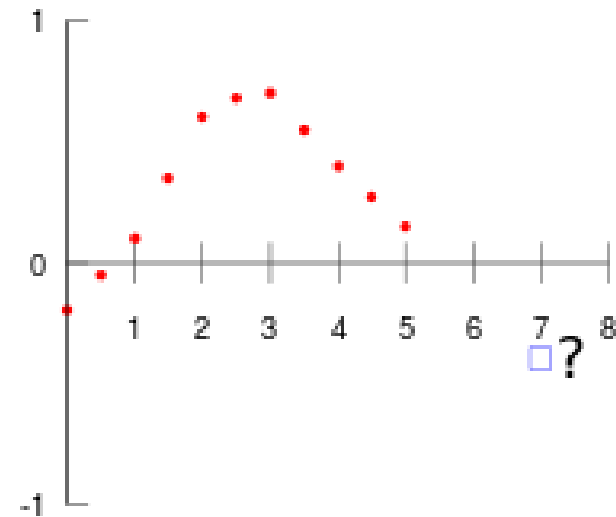
2. Imparcialidad del Testing Set Respecto al Training Set?

Problemas: La calidad de los datos

3. Completitud de los datos



Interpolación:
Situación deseable



Extrapolación:
Problema complicado

Sistemas predictivos: Sistemas de Recomendación



Eduardo M. Sánchez Vila
eduardo.sanchez.vila@usc.es

CITIUS
Grupo de Sistemas Inteligentes
Universidade de Santiago de Compostela