

Tema 9

Evaluación de Sistemas de Recomendación

SCAR

**Sistemas Complejos Adaptativos y
Recomendación**

Introducción



Evaluación

En general, una
buena
recomendación
debe ser

- Medida de la **calidad** de la recomendación
- Selección del SR más apropiado
- **Precisa:** cercana a los gustos del usuario
- **Novedosa:** debe introducir elementos diferentes a los que le gustan al usuario, pero que se adapten a su perfil

Pasos

Diseño

- Decidir como queremos que funcione el recomendador

Propiedades

- Decidir que elementos nos interesa evaluar para que funcione de esa forma

Método de evaluación

- Decidir el método a emplear en la evaluación
- Será diferente según el punto del desarrollo del SR en el que estemos

Propiedades a tener en cuenta en la evaluación de SR

Propiedades de un SR

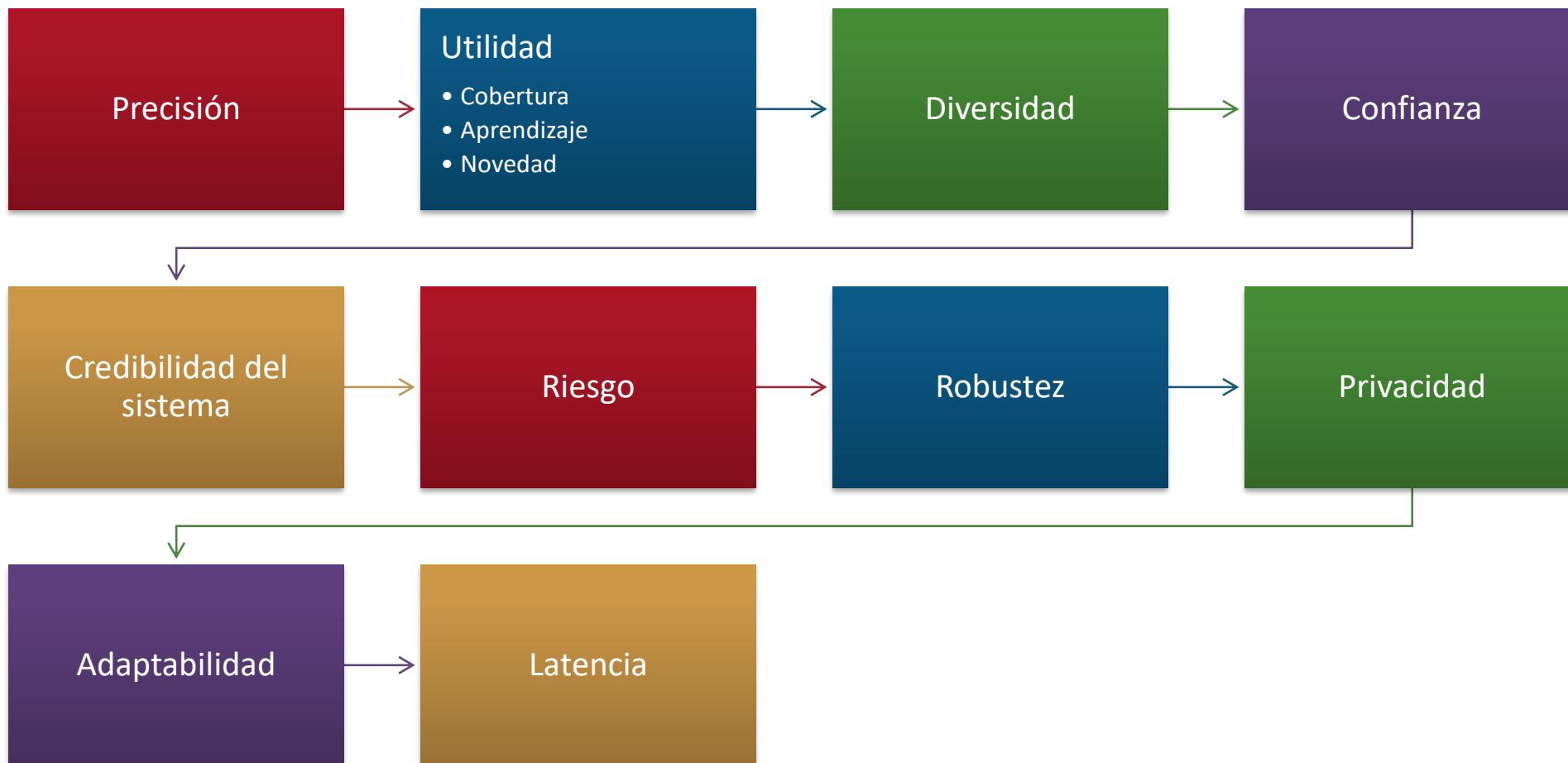
Cada SR tiene propiedades diferentes

Se debe seleccionar que propiedades son deseables, y evaluarlas

Se debe evaluar como el cambio en dichas propiedades afecta a la recomendación ofrecida al usuario

Se debe seleccionar el SR (BRT, HRS) que mejore las características seleccionadas

Posibles propiedades a medir en un SR



Propiedades de un SR: Precisión

Precisión

- Mide lo acertada que es la recomendación
- Mide lo acertada que es la predicción de los ratios de los ítems (aunque hay distintos tipos)
- Es decir, lo parecido que es el ratio calculado al ratio que habría dado el usuario

Propiedades de un SR: **Utilidad**

Utilidad

Capacidad del sistema de obtener una recomendación o una buena recomendación

- Cobertura
- Aprendizaje
- Novedad

Propiedades de un SR: Utilidad

Cobertura

- Proporción de los **ítems** que el sistema puede recomendar
 - Ítems más comunes vs ítems de características peculiares
- Proporción de los **usuarios** a los que el sistema puede recomendar ítems
 - Usuarios con gustos comunes vs ovejas negras
- Cold start (para ítems y para usuarios)

Propiedades de un SR: Utilidad

Novedad y sorpresa (serendipity)

Recomendar ítems que el usuario no conoce, o ítems de características diferentes a las que él prefiere pero que piensa que le pueden interesar

Mide lo que sorprende la recomendación a un usuario

Una recomendación aleatoria será muy sorprendente, pero puede que no sea muy precisa

Propiedades de un SR: **Utilidad**

Novedad y sorpresa

Recomendación de un ítem que el usuario no conoce

Serendipity

- Recomendación de un ítem interesante y sorprendente que el usuario no hubiera encontrado de otro modo

Propiedades de un SR: **Utilidad**

Aprendizaje

- Mide si el sistema es capaz de aprender sobre los gustos del usuario durante sucesivas recomendaciones

Propiedades de un SR: Diversidad

Diversidad

El conjunto de ítems recomendados al usuario no deberían ser muy parecidos entre sí

- Aunque, no siempre interesa diversidad

Normalmente, un incremento en la precisión hace que decrezca la diversidad

- No siempre, puede interesarle sólo eso

Propiedades de un SR: Confianza

Confianza del SR en la recomendación ofrecida

- Puede depender de la cantidad de ítems, de la información sobre el usuario,...
- Simula cuando una persona que te conoce te dice: “Seguro que te va a gustar...” o “No sé si te va a gustar”

Ejemplo

- El SR muestra dos ítems I1 e I2 con el mismo ratio obtenidos mediante recomendación colaborativa
- I1 lo obtiene basándose en 40 vecinos del usuario
- I2 lo obtiene basándose en un solo vecino
- Si se muestra al usuario el nivel de confianza que el sistema tiene en la recomendación que ofrece junto con la recomendación, el usuario tiene más información y debería elegir I1

Propiedades de un SR: **Credibilidad**

Credibilidad o reputación del sistema

- Confianza del usuario en el sistema en general
- Confianza del usuario en la recomendación ofrecida por el sistema
- Puede ser aconsejable que el sistema recomiende algunos ítems que sabe que el usuario conoce y le gustan, esto hace aumentar la confianza en el sistema

Propiedades de un SR: **Riesgo**

Riesgo

En algunos sistemas los ítems pueden llevar asociado un valor de riesgo

Ejemplo: si se recomienda una inversión de capital, si se recomienda un tratamiento médico o si se recomiendan determinados destinos turísticos

Propiedades de un SR: Robustez

Robustez Estabilidad de la recomendación en presencia de datos falsos

Puede introducirse información falsa con el objeto de orientar la recomendación a determinados ítems

Ejemplo: el dueño de un restaurante podría entrar numerosas veces en un SR de restaurantes y puntuar el suyo muy favorablemente

Propiedades de un SR: **Privacidad**

Privacidad

- Muchos usuarios desean que la información sobre sus preferencias sea privada
- Los SR colaborativos utilizan las preferencias de otros usuarios para obtener la recomendación

Propiedades de un SR: **Adaptabilidad**

Adaptabilidad

- Muchos SR trabajan con datos que cambian constantemente o con cambios bruscos en las tendencias
- Los cambios bruscos pueden afectar no solo a ítems nuevos, sino a otros antiguos
- **Ejemplo:** si el nuevo libro de un autor se convierte en best-seller, automáticamente se ponen de moda sus libros anteriores

Propiedades de un SR: Latencia

Latencia

- Tiempo que tarda el sistema en obtener la recomendación
- Los SR trabajan con gran cantidad de datos, por lo que deben ser lo suficientemente veloces
 - Pueden tener algoritmos de recomendación muy precisos, pero si la latencia es alta, no sirven
- Se pueden pre-procesar datos para aumentar la velocidad
 - La obtención de vecinos en un SR colaborativo es un proceso lento, pero pueden calcularse los vecinos de todos los usuarios del sistema y almacenarlos en el perfil de usuario

Métodos de evaluación

Métodos de evaluación

Probar el sistema cuando ya
está funcionando



Estudio **online** con los clientes
potenciales

Probar el sistema antes de
ponerlo en funcionamiento



Estudio con **usuarios reales**
(buscados a propósito para el
estudio)



Estudio **offline** (datos sintéticos, no
hay usuarios)

Métodos de evaluación

Cuestiones a tener en cuenta

¿En qué punto estoy del diseño?

- Decisión sobre técnicas o ajustes
- Pruebas
- Versión final

Evaluación online vs offline

- ¿Se puede evaluar offline o se requieren usuarios online?

Datos reales vs datos simulados

- ¿Se pueden utilizar datos simulados?
- No todos los data set son adecuados para evaluar un sistema
- ¿Qué características debe reunir el data set para evaluar el recomendador teniendo en cuenta la tarea para la que ha sido diseñado?

Métodos de evaluación

Estudio con el sistema en funcionamiento (online)

- El sistema lo usan usuarios interaccionando con el sistema ya funcionando (clientes potenciales)
- **Ventajas**
 - Información **fiable**
 - Mejor método, más **preciso**
- **Inconvenientes**
 - El sistema debe estar ya en **funcionamiento**
 - Los usuarios deben ser **variados** para poder obtener información precisa
 - Más **difícil** de obtener. Depende de que el usuario quiera dar feedback
 - Cuanto más datos se le pidan al usuario, más preciso será, pero más costoso para el usuario y más probable que **no esté dispuesto**

Métodos de evaluación

Estudio con usuarios de test reales

- Se recopila la información **real** de usuarios interaccionando con el sistema (testers)
- Se selecciona **un grupo de usuarios de test**
 - Variado
 - Público al que va dirigido
- Se **observa** a los usuarios mientras interactúan con el sistema y se recopila información
- **Ventajas**
 - Permite recopilar **gran cantidad** de información
- **Inconvenientes**
 - Son **caros**, debe realizarse con un número elevado de usuarios y en diferentes escenarios

Métodos de evaluación

Estudio offline (usando data sets)

- Data set: conjunto de ítems previamente puntuados por un grupo de usuarios
- Se intenta simular el comportamiento de los usuarios cuando acceden al sistema usando los datos disponibles
- Se utilizan parte de los datos como entrenamiento del sistema y otra parte para testear el resultado
- **Ventajas**
 - Son los más sencillos
 - No requieren interacción con el usuario
 - Permite obtener resultados usando diversas técnicas
- **Inconvenientes**
 - No son aptos para medir algunas de las propiedades, en general, útiles para métricas de precisión
 - Menor precisión de los resultados obtenidos

Método de evaluación offline

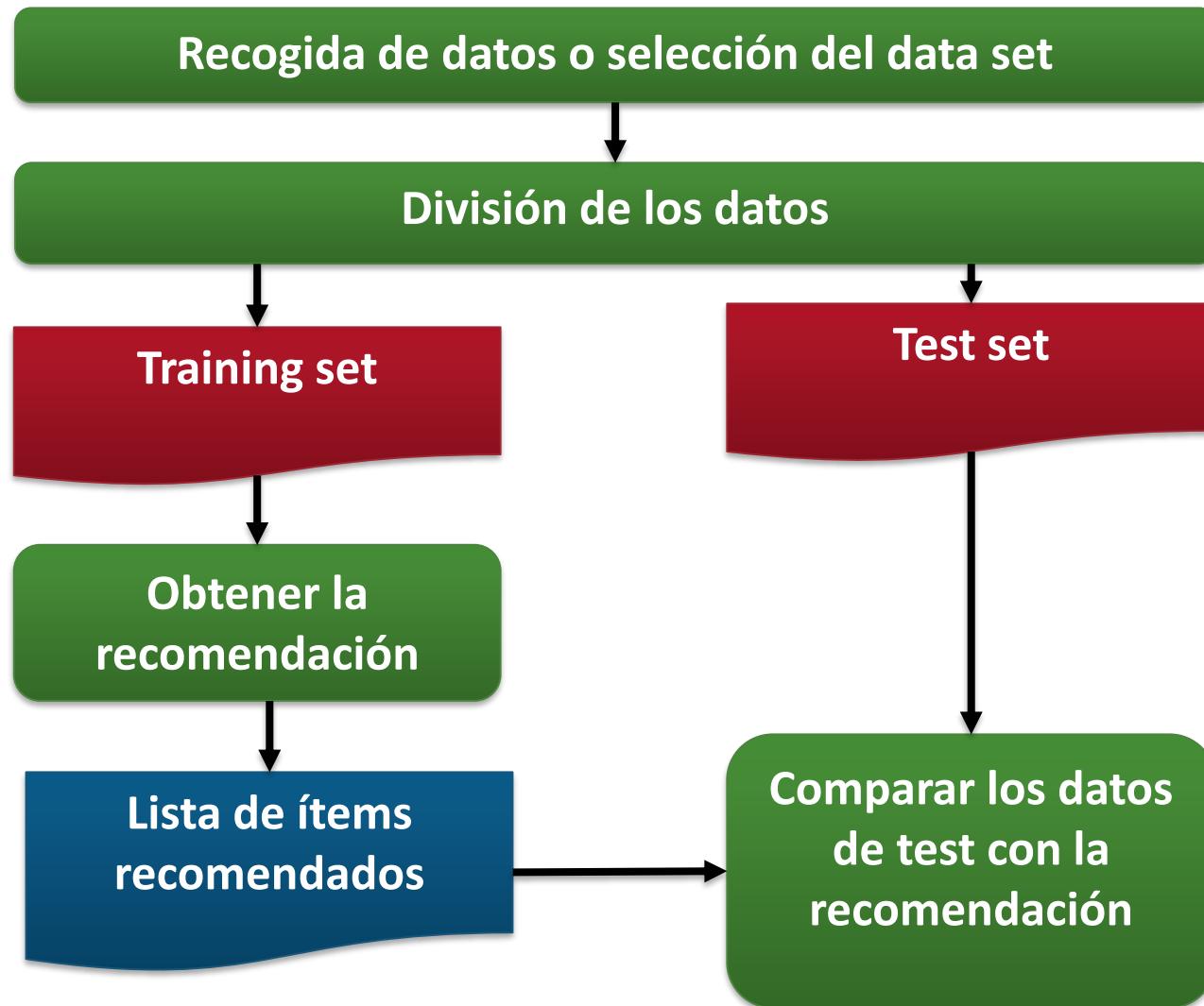
Método de evaluación offline



Data
set

Conjunto de ítems
previamente
puntuados por un
grupo de usuarios

Método de evaluación offline



Tipos de data set

Real

Resultado de la propia aplicación que ya está en funcionamiento o de una versión anterior

Generado

Cuestionarios a usuarios reales

Obtenidos de otra aplicación (Movielens)

Simulado

Datos sintéticos generados aleatoriamente o con un patrón

Ejemplos de data set

- Movielens • Películas
- Jester • Chistes
- Book-Crossings • Libros
- Last.fm • Música
- Wikipedia • Datos de interés en general
- OpenStreetMap • Puntos de interés geográficos
- Python Git Repositories • Funciones y datos de programación

Dataset	Users	Items	Ratings	Density	Rating Scale
Movielens 1M	6040	3883	1,000,209	4.26%	[1-5]
Movielens 10M	69,878	10,681	10,000,054	1.33%	[0.5-5]
Movielens 20M	138,493	27,278	20,000,263	0.52%	[0.5-5]
Jester	124,113	150	5,865,235	31.50%	[-10, 10]
Book-Crossing	92,107	271,379	1,031,175	0.0041%	[1, 10], and implicit
Last.fm	1892	17632	92,834	0.28%	Play Counts
Wikipedia	5,583,724	4,936,761	417,996,366	0.0015%	Interactions
OpenStreetMap (Azerbaijan)	231	108,330	205,774	0.82%	Interactions
Git (Django)	790	1757	13,165	0.95%	Interactions

División de los datos del data set

Training set

Datos con los que se entrena el sistema.
Simula la interacción de los usuarios con
el sistema, hasta el momento actual

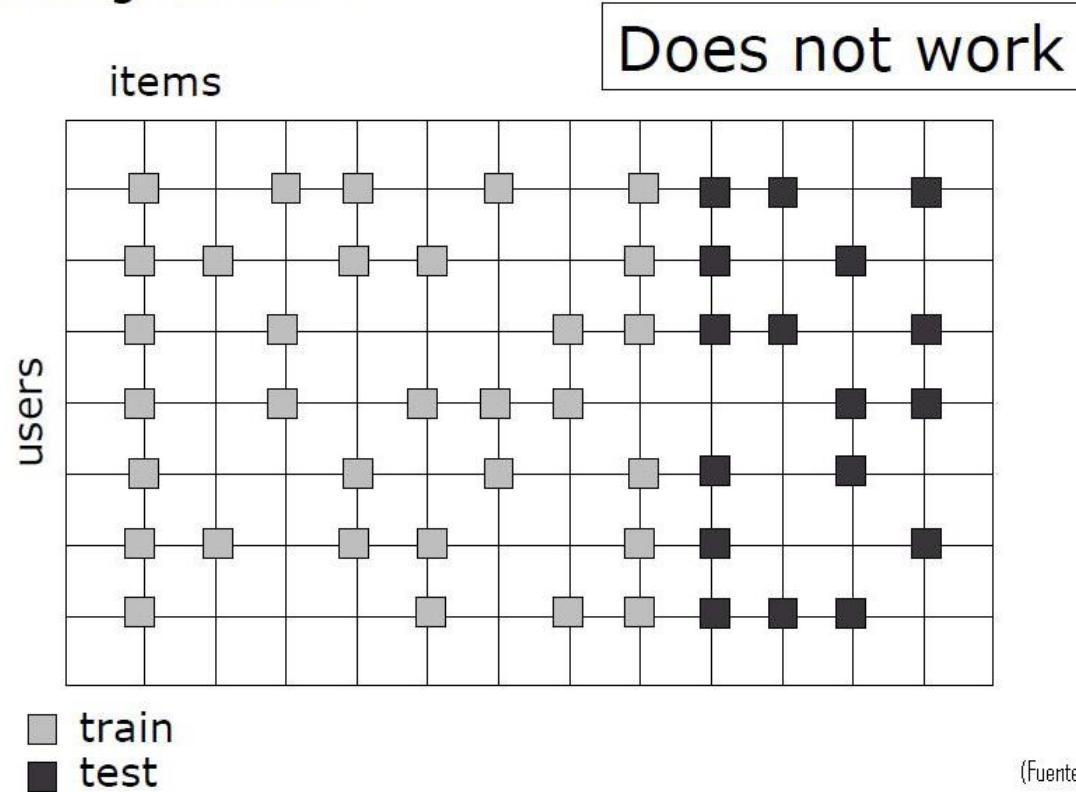
Test set

Usado para obtener resultados sobre la
recomendación ofrecida

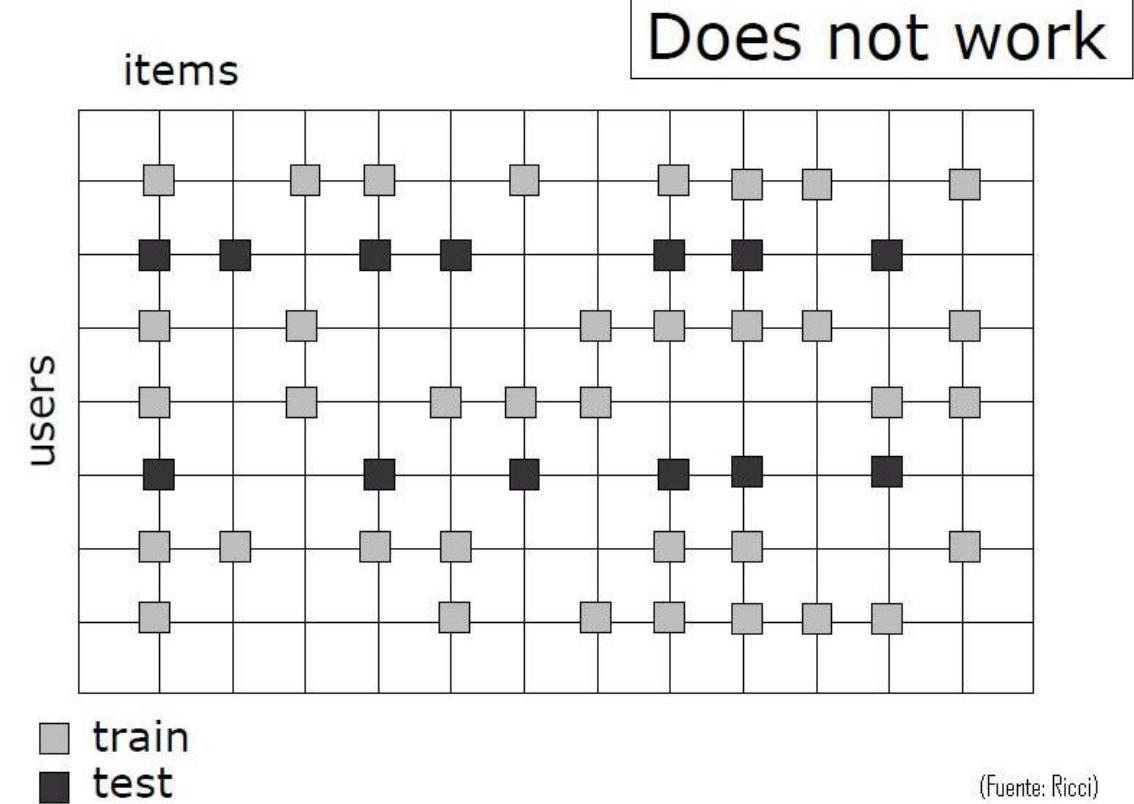


División de los datos del data set

Splitting the data



Splitting the data

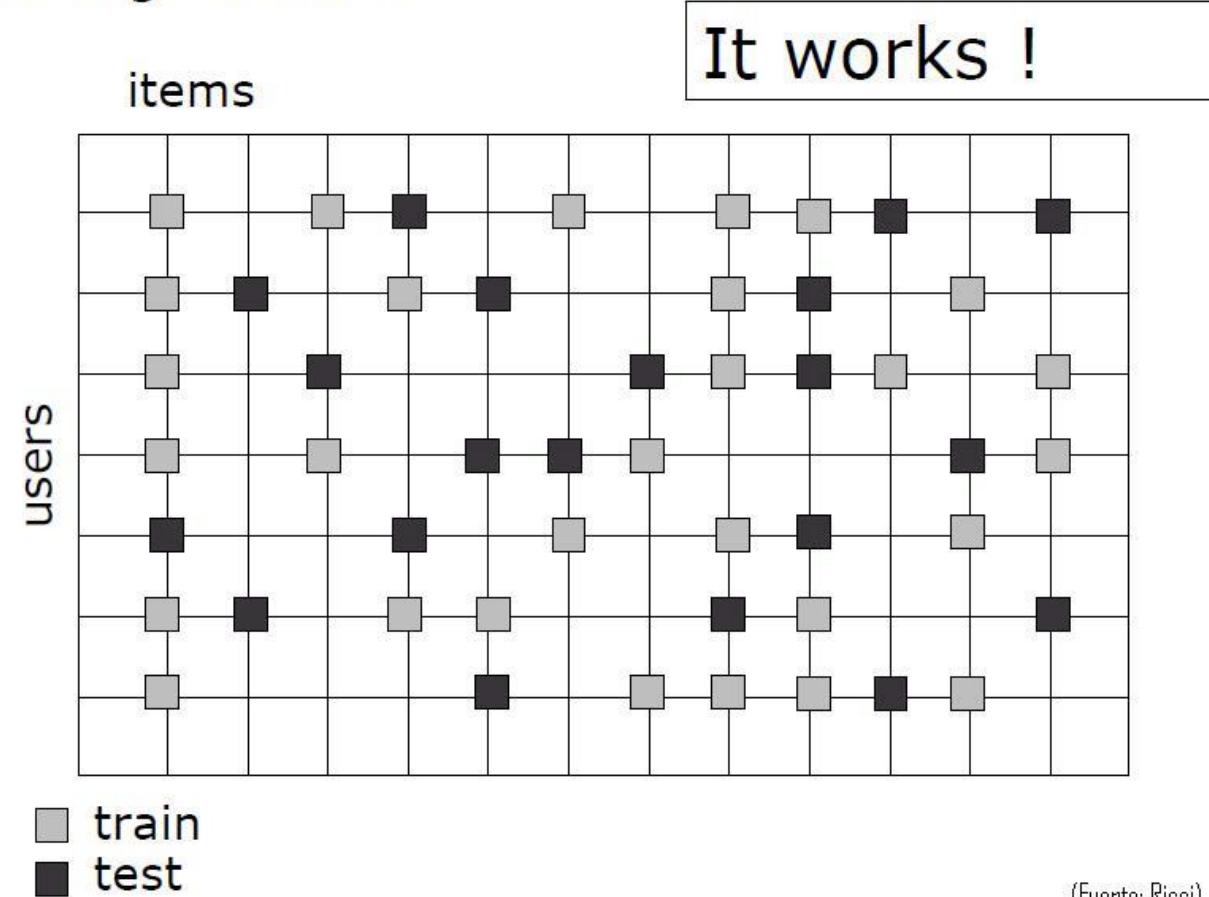


(Fuente Ricci)

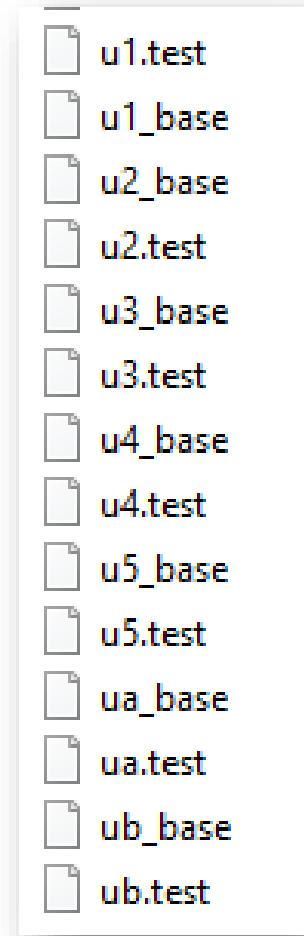
(Fuente: Ricci)

División de los datos del data set

Splitting the data



División de los datos del data set



u1_base: Bloc d...			
Archivo	Edición	Formato	Ver
1	1	5	
1	2	3	
1	3	4	
1	4	3	
1	5	3	
1	7	4	
1	8	1	
1	9	5	
1	11	2	
1	13	5	
1	15	5	
1	16	5	
1	18	4	
1	19	5	
1	21	1	
1	22	4	
1	25	4	
1	26	3	
1	28	4	
1	29	1	
1	30	3	
1	32	5	
1	34	2	

u1.test: Bloc d...			
Archivo	Edición	Formato	Ver
Ayuda			
1	6	5	
1	10	3	
1	12	5	
1	14	5	
1	17	3	
1	20	4	
1	23	4	
1	24	3	
1	27	2	
1	31	3	
1	33	4	
1	36	2	
1	39	4	
1	44	5	
1	47	4	
1	49	3	
1	51	4	
1	53	3	
1	54	3	
1	56	4	
1	60	5	
1	61	4	

El dataset contiene distintas divisiones de datos para que se pueda evaluar adecuadamente

Métricas de evaluación

Métricas



How good your recommender system is? A survey on evaluations in recommendation.

Thiago Silveira, Min Zhang¹ · Xiao Lin, Yiqun Liu, Shaoping Ma.

International Journal of Machine Learning and Cybernetics (2019) 10:813–831

Métricas

Comparan

Lista de ítems
recomendados al usuario

Obtenidos mediante el SR
que se quiere evaluar

Lista de ítems de test

Obtenidos de un data set o
mediante usuarios reales

Métricas

Ítems recomendados

Lista de ítems que se muestra en la interfaz (N)

Todos los ítems recomendados que superan un ratio

Todos los ítems recomendados

Ítems relevantes

Los ítems relevantes son aquellos que el usuario hubiese elegido

N ítems que el usuario hubiese elegido (los de mayor ratio)

Todos los que le gustan (superan un ratio)

Todos los ítems relevantes

Métricas

Lista de N ítems recomendados

Lista de todos los ítems recomendados que superan un ratio

Lista de todos los ítems recomendados



Se pueden hacer comparaciones con cualquiera de éstas combinaciones

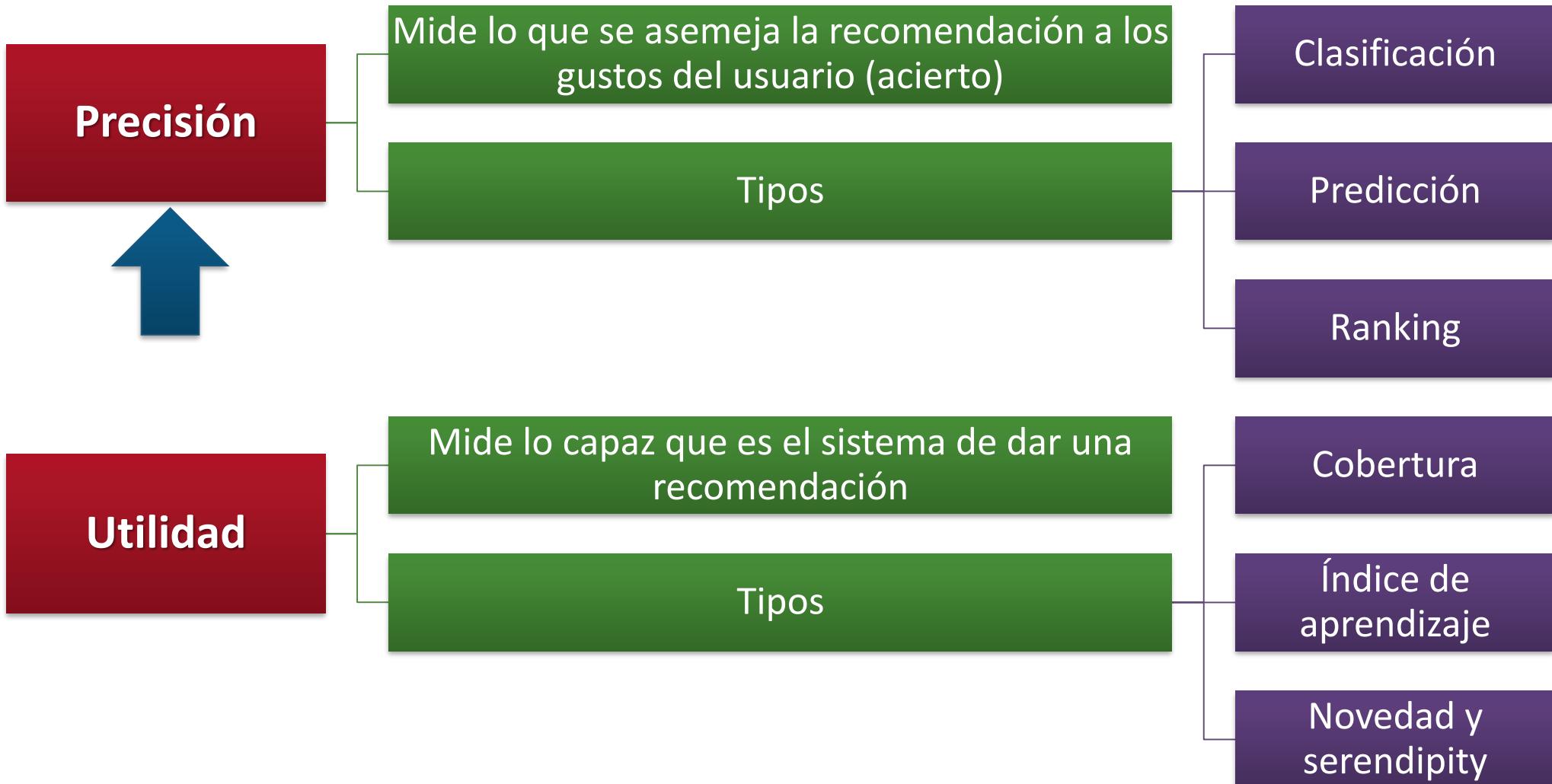
Lista de N ítems relevantes

Lista de todos ítems relevantes que superan un ratio

Depende de lo que nos interese

Lista de todos los ítems relevantes

Métricas



Métricas de precisión

Métricas de precisión

Mide lo que se asemeja
la recomendación a los
gustos del usuario
(acuerdo)

Tipos

- Clasificación
- Predicción
- Ranking (posición)

Métricas de precisión

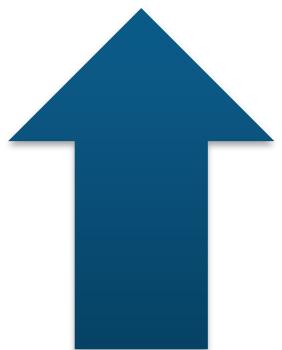
Precisión de la clasificación

- Compara los elementos incluidos en las listas



Precisión de la predicción

- Compara los ratios de ambas listas

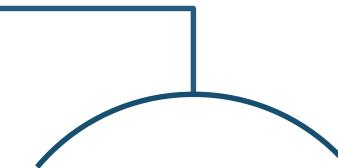
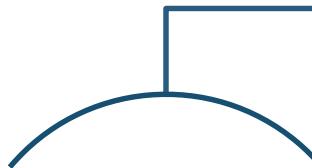


Precisión del ranking

- Compara el orden de las listas

Métricas de precisión: Clasificación

Mide la frecuencia con la que
el SR selecciona un elemento
de interés para el usuario



Métricas de precisión: Clasificación

Precisión

Porcentaje de ítems recomendados que son relevantes (de interés) para el usuario

Probabilidad de que un elemento seleccionado sea relevante

Recall

Porcentaje de ítems relevantes para el usuario que son recomendados

Probabilidad de seleccionar un elemento relevante

$$\text{Precision} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Recomendados}|}$$

$$\text{Recall} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Relevantes}|}$$



Ejemplo

Se recomiendan 5 ítems (rojos y verdes)

10 ítems relevantes para el usuario (azules y verdes)

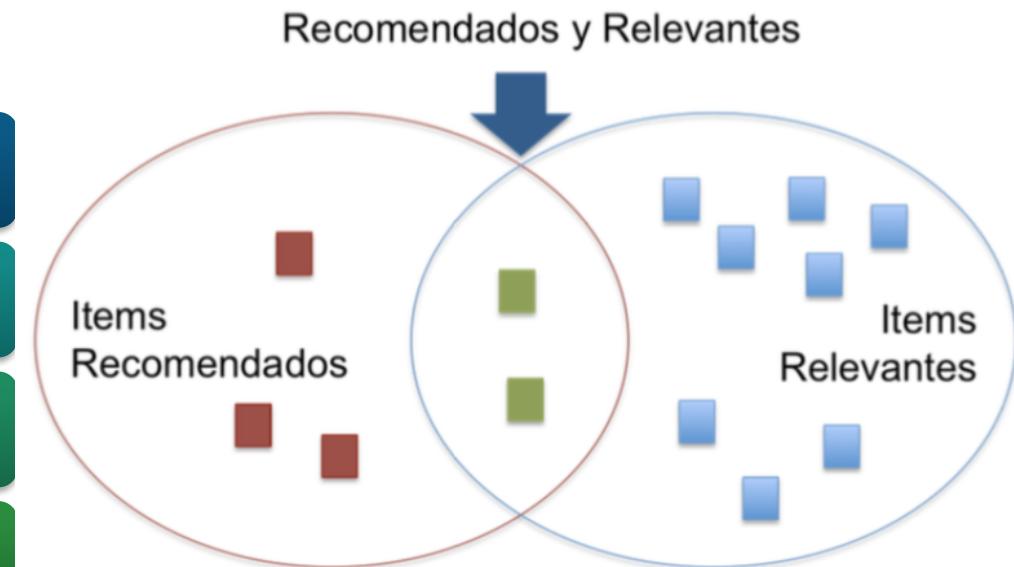
2 ítems recomendados son relevantes

Precisión

$$\bullet \frac{2}{5} = 0.4 \text{ (40\%)}$$

Recall

$$\bullet \frac{2}{10} = \frac{1}{5} = 0.2 \text{ (20\%)}$$



$$\text{Precision} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Recomendados}|}$$

$$\text{Recall} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Relevantes}|}$$

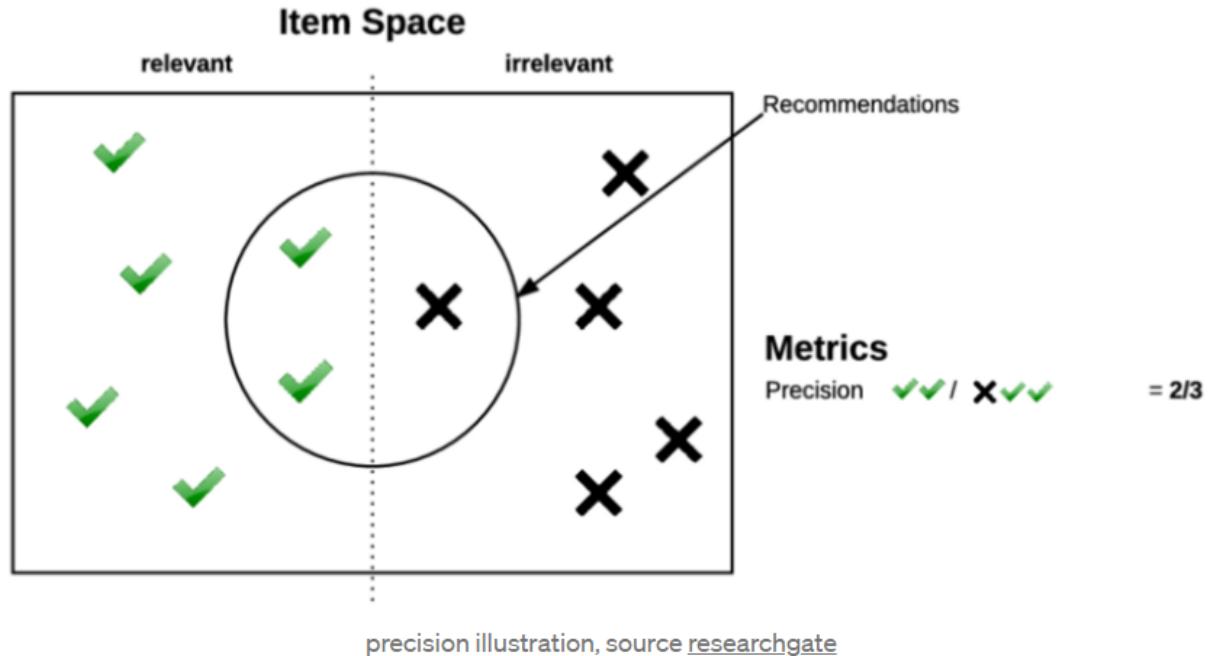
Precisión

- Porcentaje de ítems recomendados que son relevantes (de interés) para el usuario

Recall

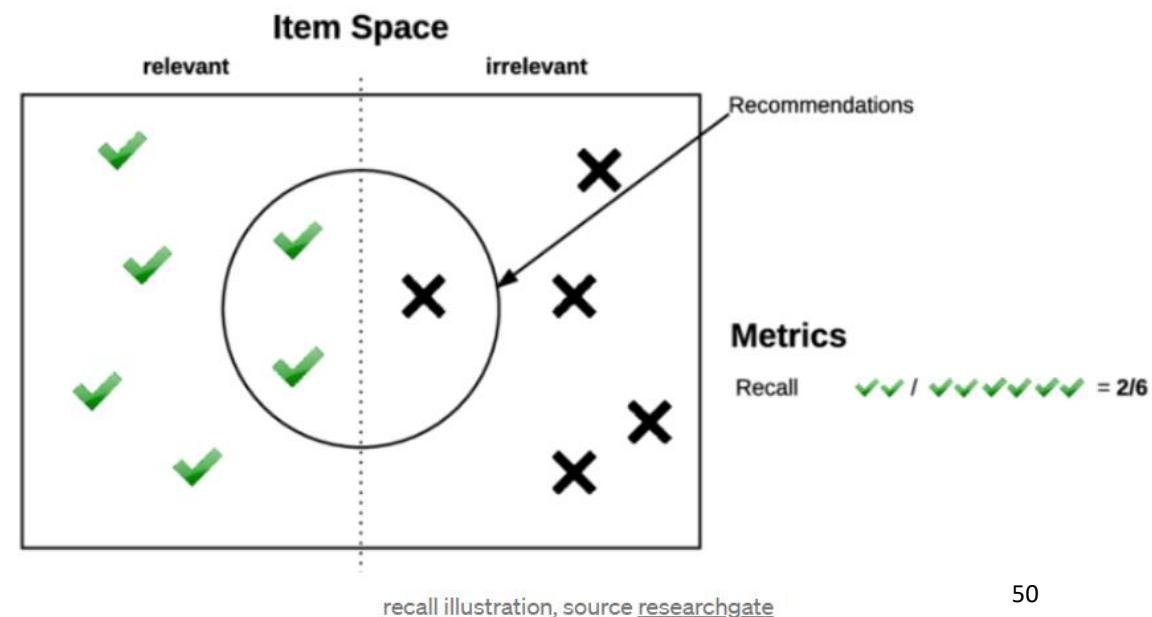
- Porcentaje de ítems relevantes para el usuario que son recomendados

Ejemplo

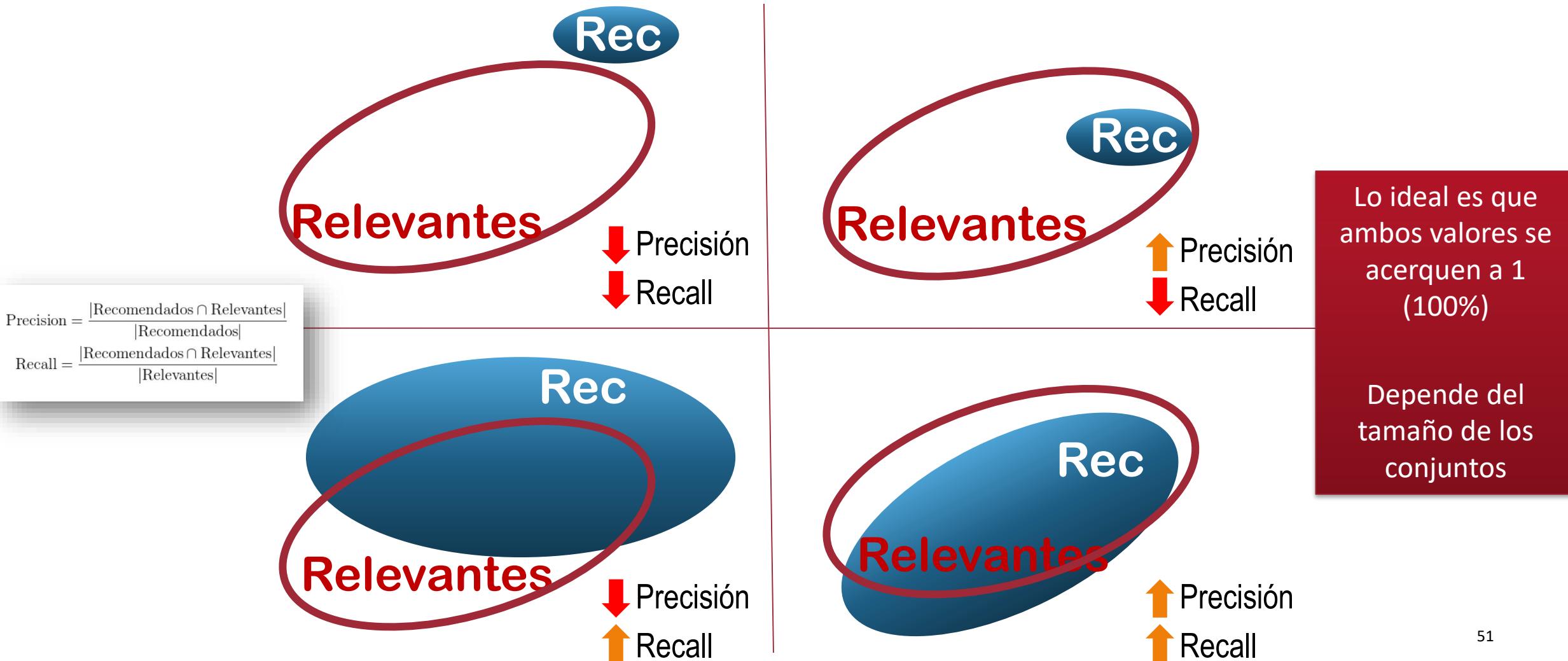


$$\text{Precision} = \frac{2}{3} = 0,67 \text{ (67\%)}$$
$$\text{Recall} = \frac{2}{6} = 0,33 \text{ (33\%)}$$

$$\text{Precision} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Recomendados}|}$$
$$\text{Recall} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Relevantes}|}$$



Métricas de precisión: Clasificación



Ejemplo

	Pañales	Toallitas	Chupete	Sonajero
Me gustan			*	
Me recomiendan		+	+	

$$\text{Precision} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Recomendados}|}$$

$$\text{Recall} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Relevantes}|}$$

RECALL

$$\frac{\#Gustan \& Muestran}{\#Gustan} = \frac{1}{1} = 1$$

Desde el punto de vista del Recall la situación es buena porque me recomienda el que me gusta

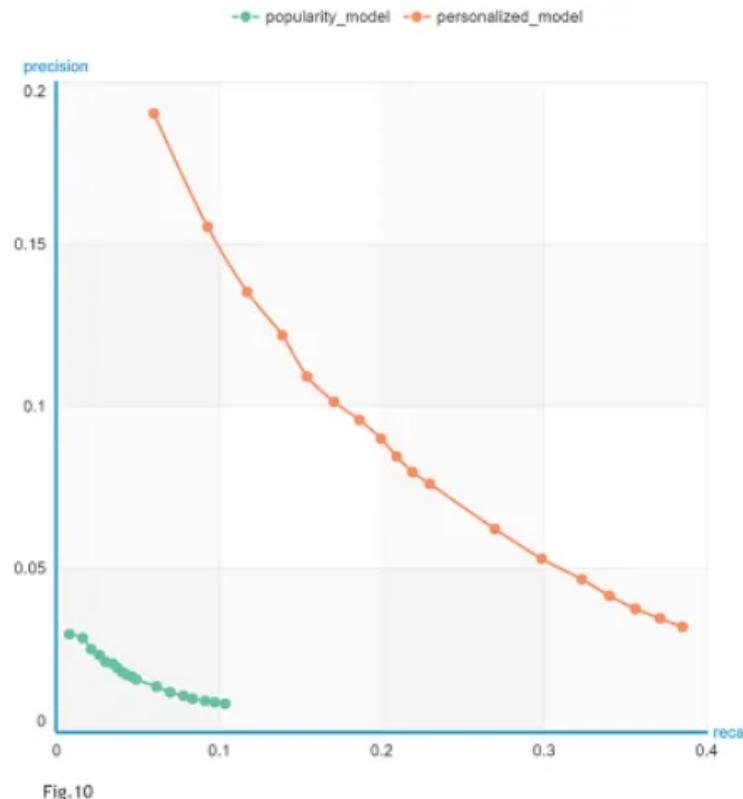
PRECISION

$$\frac{\#Gustan \& Muestran}{\#Muestran} = \frac{1}{2} = 0,5$$

Desde el punto de vista de Precision no tanto porque me muestran más de los que me gustan. Digamos que “me molestan” un poco.

Ejemplo

Por último, para comparar los recomendadores se grafica Precision vs Recall (en general para un grupo test – Fig10 – **Curva Precision-Recall**). Tal y como indicamos en la figura, la curva que está por encima nos indica cuál es el mejor recomendador.



La curva que está por arriba nos indica cuál es el mejor recomendador

Fig.10

Métricas de precisión: Predicción

Compara los ratios de los ítems de la lista de ítems recomendados con los ratios que habría dado el usuario

Mean absolute error MAE

- Error absoluto medio

Mean Squared Error MSE o MAE^2

- Error cuadrático medio

$$MAE = \frac{\sum_{i=1}^N | p_i - r_i |}{N}$$

$$MAE^2 = \frac{\sum_{i=1}^N (| p_i - r_i |)^2}{N}$$

p: predicción (ratio del recomendador)

r: ratio del usuario

N: nº ítems recomendados

Métricas de precisión: Predicción

Error absoluto medio MAE

Media de la diferencia entre el ratio predicho por el recomendador y el ratio dado por el usuario

Define lo lejos que está el ratio predicho respecto al que habría dado el usuario

Se usan valores absolutos porque sólo interesan las diferencias, sean positivas o negativas

**Cuanto menor es el MAE, mejor recomendación
(ideal MAE=0)**

$$MAE = \frac{\sum_{i=1}^N | p_i - r_i |}{N}$$

p: predicción (ratio del recomendador)

r: ratio del usuario

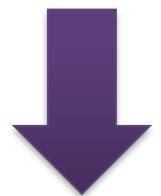
N: nº ítems recomendados

$$MAE = \frac{\sum_{i=1}^N |p_i - r_i|}{N}$$

Ejemplo

Película	Ratio usuario (p)	Ratio estimado (r)	Diferencia
Titanic	5	4.1	0.9
Pretty woman	3	3.5	-0.5
El señor de los anillos	5	4.8	0.2
Atrapado en el tiempo	5	4.5	0.5
La novia cadáver	3	4.1	1.1
Los juegos del hambre	4	3.7	0.3

MAE = 0.58



Película	Ratio usuario (p)	Ratio estimado (r)	Diferencia
Titanic	5	4.9	0.1
Pretty woman	3	2.8	0.2
El señor de los anillos	5	4.8	0.2
Atrapado en el tiempo	5	4.9	0.1
La novia cadáver	3	3.1	0.1
Los juegos del hambre	4	3.7	0.3

MAE = 0.13

Métricas de precisión: Predicción

Error cuadrático medio
 MAE^2

Enfatiza las
diferencias
entre los ratios

$$MAE^2 = \frac{\sum_{i=1}^N (| p_i - r_i |)^2}{N}$$

p: predicción (ratio del recomendador)

r: ratio del usuario

N: nº ítems recomendados

$$\sum_{i=1}^N (|p_i - r_i|)^2$$

$$MAE^2 = \frac{\sum_{i=1}^N (|p_i - r_i|)^2}{N}$$

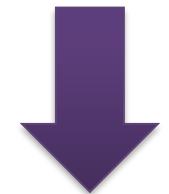
Ejemplo

Película	Ratio usuario (p)	Ratio estimado (r)	Diferencia ²
Titanic	5	4.1	0.81
Pretty woman	3	3.5	0.25
El señor de los anillos	5	4.8	0.04
Atrapado en el tiempo	5	4.5	0.25
La novia cadáver	3	4.1	1.21
Los juegos del hambre	4	3.7	0.09

$$MAE^2 = 0.44$$

$$MAE = 0.58$$

Película	Ratio usuario (p)	Ratio estimado (r)	Diferencia ²
Titanic	5	4.9	0.1
Pretty woman	3	2.8	0.04
El señor de los anillos	5	4.8	0.04
Atrapado en el tiempo	5	4.9	0.1
La novia cadáver	3	3.1	0.1
Los juegos del hambre	4	3.7	0.09



$$MAE^2 = 0.078$$

$$MAE = 0.13$$

Métricas de precisión: Ranking

Considerar la lista de ítems recomendados ordenada por ratio

Mide si la lista ordenada de recomendaciones se asemeja a la lista que el usuario habría seleccionado

- Considera la **posición** de los ítems en la lista

Métricas de precisión: Ranking

Cumulative gain CG

- Ganancia acumulada
- Suma los ratios de la lista de ítems recomendados
- **A mayor CG, mejor recomendación**
- Se suele sumar un “ranking” en vez del ratio:
 - Ítems relevantes 2p
 - Relevancia media 1p
 - Menos relevantes 0p
- No tiene en cuenta la posición de los ítems en la lista

Items Ranking	Relevancy Score
Movie 1	1
Movie 3	2
Movie 2	2
Movie 5	0
Movie 4	1

CG = 6

$$CG_p = \sum_{i=1}^p reli_i$$

Métricas de precisión: Ranking

Discounted cumulative gain DCG

- Ganancia acumulada descontada
- Penaliza los elementos que aparecen más abajo en la lista
 - Un ítem relevante que aparece abajo en la lista, es una mala recomendación (mal rendimiento)
- Divide el ratio del ítem por el logaritmo de su posición en la lista

Items Ranking	Relevancy Score
Movie 1	1
Movie 3	2
Movie 2	2
Movie 5	0
Movie 4	1

CG =	6
DCG =	12.1

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)}$$

Métricas de precisión: Ranking

Normalized discounted cumulative gain nDCG

- Ganancia acumulada descontada normalizada
- **Valor entre 0 y 1**
- El valor de DGC no es comparable si la lista tiene distinto número de ítems (no es lo mismo estar en la posición 5 de 10 elementos, que 5 de 6 elementos)
- Se ordena la lista de ítems recomendados por relevancia y se calcula su DGC (IDGC)

Items Ranking	Relevancy Score	Perfect Ranking	Relevancy Score
Movie 1	1	Movie 3	2
Movie 3	2	Movie 2	2
Movie 2	2	Movie 1	1
Movie 5	0	Movie 4	1
Movie 4	1	Movie 5	0

CG =	6	CG (p) =	6
DCG =	12.1	DCG (p) =	13.9

$$nDCG = \frac{DCG}{IDCG} = 0.87$$

$$nDCG_p = \frac{DCG_p}{IDCG_p},$$

Ejemplo

Dados dos conjuntos de ítems recomendados, A y B, según la ganancia acumulada, ambos son igualmente buenos

Según la ganancia acumulada descontada, B es mejor recomendación que A

$$CG_A = 2 + 3 + 3 + 1 + 2 = 11$$

$$CG_B = 3 + 3 + 2 + 2 + 1 = 11$$

$$Set\ A = [2, 3, 3, 1, 2]$$

$$Set\ B = [3, 3, 2, 2, 1]$$

$$DCG_A = \frac{2}{\log_2(1+1)} + \frac{3}{\log_2(2+1)} + \frac{3}{\log_2(3+1)} + \frac{1}{\log_2(4+1)} + \frac{2}{\log_2(5+1)} \approx 6.64$$

$$DCG_B = \frac{3}{\log_2(1+1)} + \frac{3}{\log_2(2+1)} + \frac{2}{\log_2(3+1)} + \frac{2}{\log_2(4+1)} + \frac{1}{\log_2(5+1)} \approx 7.14$$

$$DCG_A < DCG_B$$

Ejemplo

Si se usa la ganancia acumulada descontada normalizada, donde DCG considera el orden obtenido e IDCG el orden ideal

Cuanto más se acerca nDCG a 1, mejor es la recomendación

$$\text{Recommendations Order} = [2, 3, 3, 1, 2]$$

$$\text{Ideal Order} = [3, 3, 2, 2, 1]$$

$$DCG = \frac{2}{\log_2(1+1)} + \frac{3}{\log_2(2+1)} + \frac{3}{\log_2(3+1)} + \frac{1}{\log_2(4+1)} + \frac{2}{\log_2(5+1)} \approx 6.64$$

$$iDCG = \frac{3}{\log_2(1+1)} + \frac{3}{\log_2(2+1)} + \frac{2}{\log_2(3+1)} + \frac{2}{\log_2(4+1)} + \frac{1}{\log_2(5+1)} \approx 7.14$$

$$NDCG = \frac{DCG}{iDCG} = \frac{6.64}{7.14} \approx 0.93$$

Métricas de precisión: Ranking

Otras métricas

Mean reciprocal Rank, MRR

Average Precision

Spearman rank correlation

<https://towardsdatascience.com/an-exhaustive-list-of-methods-to-evaluate-recommender-systems-a70c05e121de>

<https://medium.com/@eng.saavedra/sistemas-de-recomendaci%C3%B3n-parte-4-evaluaciones-cfd1f96b887a>

Métricas de utilidad

Métricas de utilidad

Mide si el sistema es capaz de ofrecer recomendaciones

Cobertura

Índice de aprendizaje

Novedad y serendipity

Métricas de utilidad: Cobertura

Cobertura de ítems

- Mide el porcentaje de ítems que el sistema puede recomendar
- Proporción de los **ítems** que el sistema puede recomendar (ítems más comunes vs ítems de características peculiares)

Cobertura de usuarios

- Proporción de los **usuarios** a los que el sistema puede recomendar ítems (usuarios con gustos comunes vs ovejas negras)

Cobertura baja \Rightarrow limitaciones del sistema

Métricas de utilidad: Índice de aprendizaje



Mide si el sistema es capaz de aprender sobre los gustos del usuario o sobre como ofrecer la recomendación, durante sucesivas recomendaciones

Mide la **rapidez** con la que un SR puede producir **buenas** recomendaciones

“Cold-start” o problema de usuario o ítem nuevo

Métricas de utilidad

Se puede encontrar una explicación de las métricas de utilidad en

Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. Vargas, Saúl and Castells, Pablo. Proceedings of RecSys 2011

Evaluación centrada en el usuario

Evaluación centrada en el usuario



La evaluación de la recomendación se suele realizar de forma offline

Se basa en métricas que miden el acierto de los ítems recomendados y en los ratios

La precisión de los algoritmos es sólo uno de los factores que influencian la aceptación de las recomendaciones por parte de los usuarios

Introducir ciertos elementos de explicación de la recomendación, aumenta la satisfacción del usuario

- Se deberían medir otros factores que indiquen la satisfacción del usuario con la recomendación recibida
- Evaluación centrada en el usuario

Evaluación centrada en el usuario

La evaluación centrada en el usuario es un enfoque en el diseño y desarrollo de SR que se centra en entender y medir cómo el usuario interactúan con el sistema, con el objetivo de mejorar la experiencia del usuario y la satisfacción general

Se trata de poner al usuario en el centro del proceso de evaluación para asegurar que las recomendaciones satisfagan sus necesidades, expectativas y preferencias

Este enfoque implica recopilar datos sobre la experiencia del usuario, comprender sus comportamientos y opiniones, y utilizar esta información para mejorar continuamente el proceso de recomendación

Técnicas para la evaluación centrada en el usuario

Pruebas de usabilidad

- Observar a los usuarios mientras interactúan con el sistema para identificar problemas de usabilidad, dificultades de navegación o cualquier otro aspecto que pueda afectar negativamente la experiencia del usuario

Encuestas y cuestionarios

- Recopilar información directamente de los usuarios a través de encuestas y cuestionarios para entender sus necesidades, preferencias y opiniones sobre el sistema

Entrevistas en profundidad

- Realizar entrevistas en profundidad con usuarios representativos para obtener una comprensión más detallada de sus necesidades y motivaciones

Análisis de métricas de uso

- Analizar métricas de uso y comportamiento del usuario, como tiempos de sesión, clics, ..., para entender cómo los usuarios interactúan con el sistema y qué partes pueden necesitar mejoras

Elementos que aumentan la satisfacción con un SR

Transparencia

- Explicar cómo funciona el sistema

Escrutabilidad

- Dejar al usuario indicar que el sistema comete un error

Confianza

- Incrementar la confianza del usuario en el sistema

Efectividad

- Ayudar al usuario a tomar buenas decisiones

Persuasión

- Convencer a usuario a probar o a comprar

Eficiencia

- Ayudar a usuarios a tomar decisiones más rápido

Satisfacción

- Aumentar facilidad de uso o el disfrute en el sistema

Elementos que aumentan la satisfacción con un SR

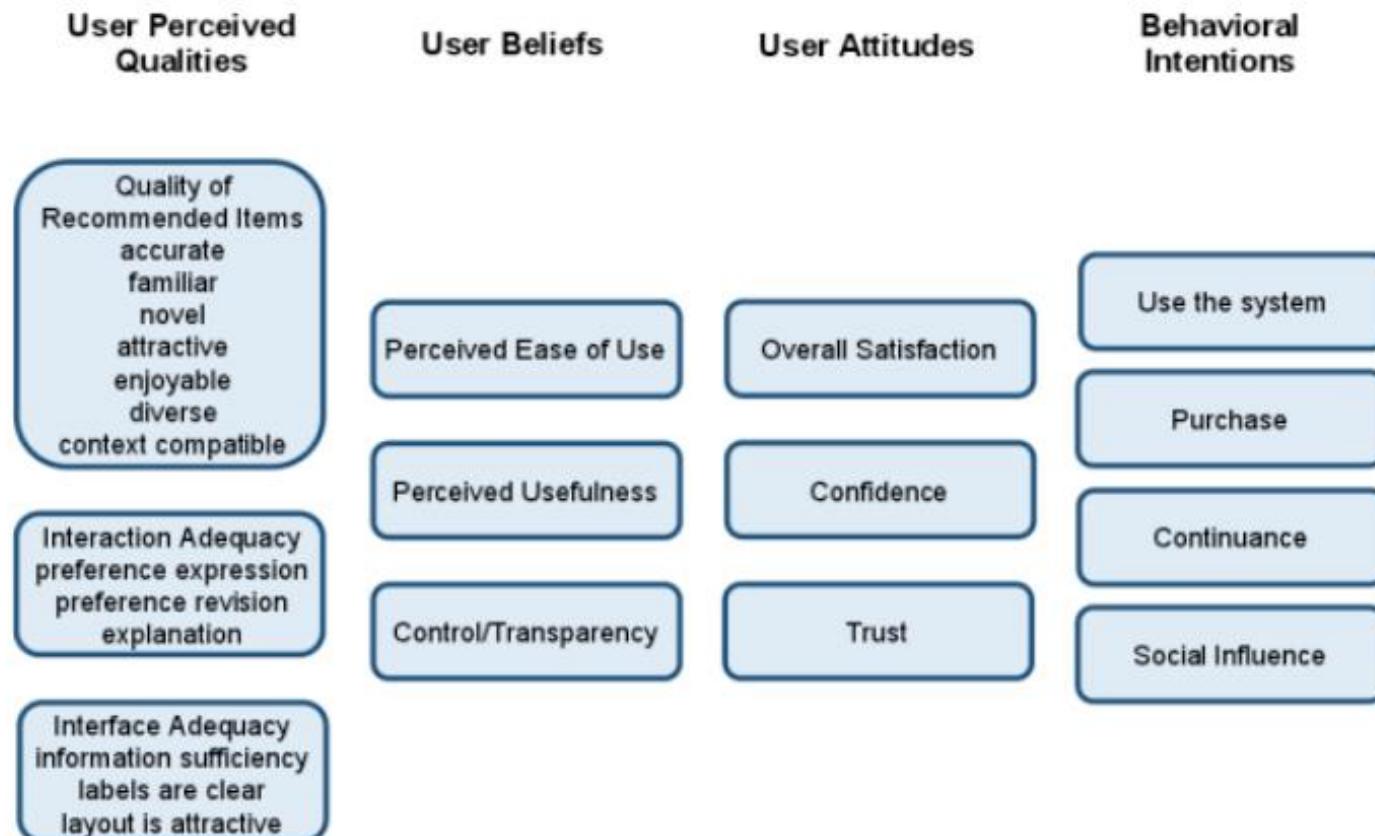


Figure 1: Constructs of an Evaluation Framework on the Perceived Qualities of Recommenders (ResQue).

Elementos que aumentan la satisfacción con un SR

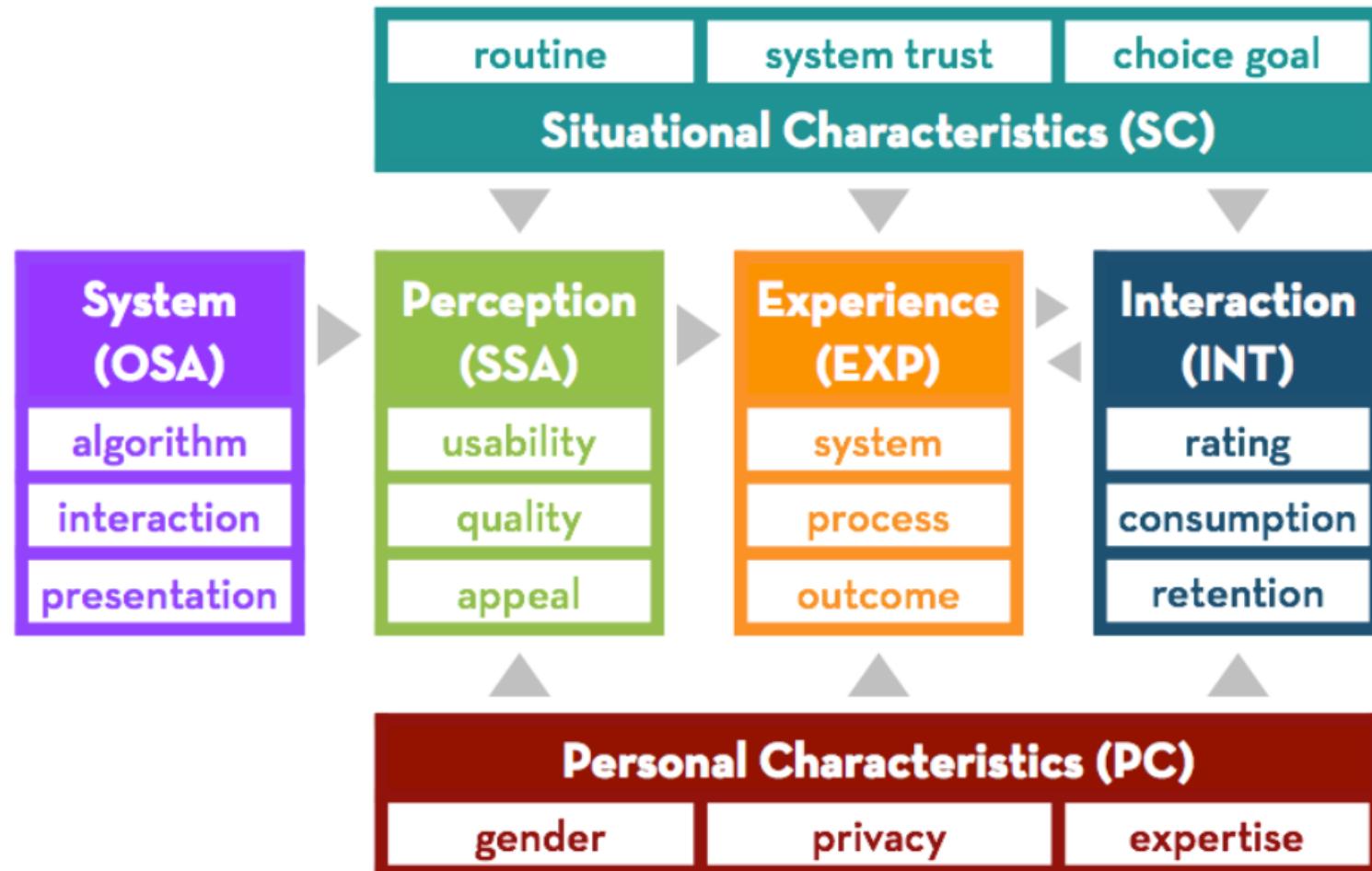


Fig. 1 An updated version of the User-Centric Evaluation Framework [61].

**Gracias por vuestra
atención...**