



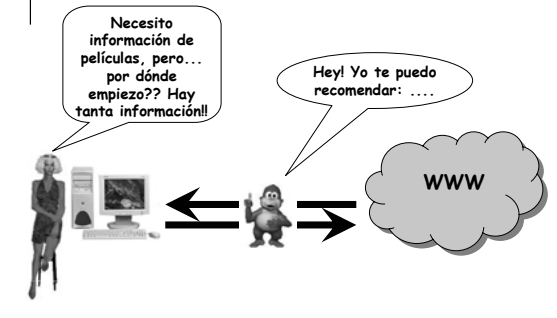
DCIC 

Sistemas de Recomendación y Personalización

Viviana L. Matos
 MINERÍA DE LA WEB-2006
 Universidad Nacional del Sur



SR: Situación de Aplicabilidad



Necesito información de películas, pero... por dónde empiezo?? Hay tanta información!!

Hey! Yo te puedo recomendar:

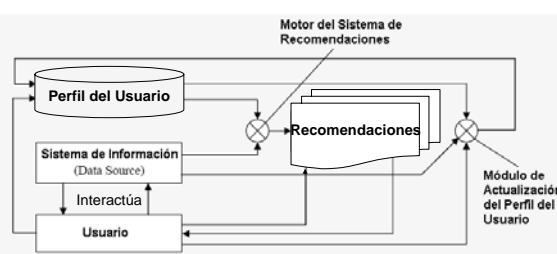
WWW

SR: Concepto

Un Sistema de Recomendación es una herramienta que ayuda al usuario a obtener la información que necesita de acuerdo a sus preferencias.

- Un SR toma decisiones basado en **información disponible**.
- Un SR requiere de un **método de filtrado de la información**.
- Un SR se aplica en algún dominio determinado(Ej: películas, música, restaurantes, email, e-commerce,..)

Arquitectura de un Sistema Recomendador



Motor del Sistema de Recomendaciones

Perfil del Usuario

Sistema de Información (Data Source)


Interactúa

Usuario

Recomendaciones

Módulo de Actualización del Perfil del Usuario

SR: Una Taxonomía

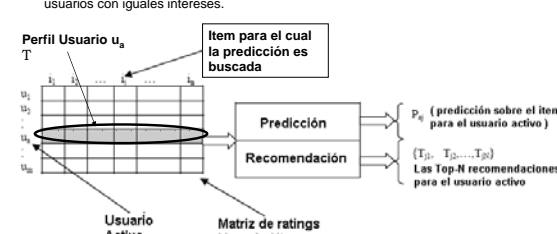


Sistema Recomendador

- Tipo Sistema Recomendador**
 - Colaborativo
 - Basado en Contenido
 - Demográfico
 - Basado en Conocimiento
 - Híbrido
- Objetivo de la Recomendación (Target)**
 - Individual
 - Grupal
- Tecnologías Asociadas**
 - Datamining
 - Information Retrieval
 - Grafos
 - Clustering
 - Aprendizaje Mecánico
- Áreas de Aplicación (Dominios)**
 - Música
 - Películas
 - Noticias
 - Páginas Web
 - E-commerce
- En el Contexto de Otras Tecnologías Relacionadas**
 - Information Retrieval
 - Filtro de Información
- Perfil Del Usuario**
 - Representación
 - Mantenimiento
 - Explotación

Tipos de SR: Sistemas Colaborativos

- Mapea usuarios con intereses similares y luego crea recomendaciones sobre esa base.
- Sugiere nuevos ítems o predice la utilidad de un cierto ítem para un usuario particular basado en los gustos previos del usuario y en las opiniones de otros usuarios con iguales intereses.



Perfil Usuario u_a

Item para el cual la predicción es buscada

Matriz de ratings Usuario / Item

Predicción

Recomendación

Usuario Activo

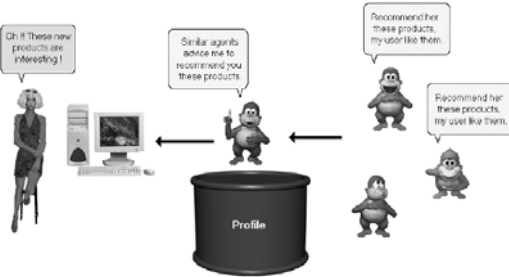
P_{ij} (predicción sobre el ítem para el usuario activo)

$\{I_1, I_2, \dots, I_N\}$ Las Top-N recomendaciones para el usuario activo

Fuente: (Sarwar et al., 2001).

Tipos de SR: Colaborativos

Gráficamente...

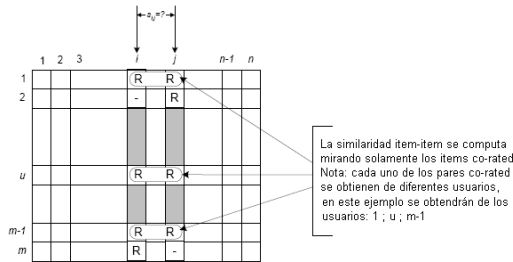


Tipos de SR: Colaborativos

- Un Filtro Colaborativo puede ser
 - Basado en Memoria
 - Utiliza el algoritmo del **vecino más cercano** para chequear por similitud a los usuarios con respecto al usuario activo. Y combina las preferencias para determinar una predicción o lista de recomendaciones para el usuario activo.
 - Basado en Modelo
 - No usan la matriz de ratings directamente. Usan un **esquema de aprendizaje mecánico** para construir un modelo de valoraciones de usuarios. Ej: **sistema de filtro colaborativo basado en ítem**. Un usuario podría estar interesado en ítems que le gustaron y evitaría los ítems que no le gustaron.

Filtro Colaborativo Basado en Ítem.

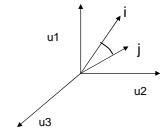
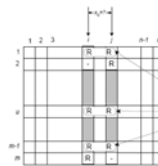
Sobre el conjunto de ítems que el usuario activo rankéó, computa cuán **similar** es el ítem recuperado con respecto a los ítems valorados por el u_i , y elige los k más similares (i_1, i_2, \dots, i_k) y al mismo tiempo sus correspondientes similitudes $\{s_{i_1}, s_{i_2}, \dots, s_{i_k}\}$.



Similitud entre ítems

Similitud Por Coseno

$$sim(i, j) = \cos(\vec{r}_i, \vec{r}_j) = \frac{\vec{r}_i \cdot \vec{r}_j}{\|\vec{r}_i\|_2 * \|\vec{r}_j\|_2}$$



Similitud Por Coseno Ajustada

$$sim(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_u)^2}}$$

Similitud basada en la Correlación de Pearson

$$sim(i, j) = corr_{i,j} = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_i)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_j)^2}}$$

Tipos de SR: Colaborativos

- Cálculo de la predicción

Weighted Sum

La predicción sobre un ítem i para un usuario u es la suma de los ratings dados por el usuario u a los ítems similares a i . Cada rating es pesado por la correspondiente similitud $s(i, j)$ entre los ítems i y j :

$$P_{u,i} = \frac{\sum_{\text{all similar items, } N} (s_{i,N} * R_{u,N})}{\sum_{\text{all similar items, } N} (s_{i,N})}$$

Luego la suma se normaliza para que la predicción caiga dentro de un rango definido

Desventajas Filtros Colaborativos

- Problema de Cold-Start
 - ✓ Problema del Usuario Nuevo (ó early rater)
 - ✓ Problema de Ítem Nuevo
- Problema de Dispersión (Sparsity)
 - ✓ Si el número de usuarios es pequeño en relación al volumen de información en el sistema, se corre el riesgo de que el cubrimiento de ratings se vuelva muy disperso. Achicando la colección de ítems recomendables.
- Problema de Escalabilidad
 - ✓ A medida que la cantidad de usuarios y de ítems crece, también crece la cantidad de cálculos de vecinos más cercanos para la determinación de usuarios similares, y como los cálculos se hacen en tiempo real, el sistema puede colapsar.

Desventajas Filtros Colaborativos

- Problema de la Oveja Gris
 - ✓ Existen usuarios donde sus perfiles caen entre clases existentes de usuarios, haciendo difícil determinar para ellos una recomendación adecuada.
- Problema de la Sinonimia
 - ✓ Se produce por la escasez de cualquier forma de interpretación semántica. Items similares no se tratarán de tal manera cuando se hagan las recomendaciones.
- Problema de la Subjetividad
 - ✓ Con respecto a la naturaleza de los ratings

Ventajas Filtros Colaborativos

- **Permite recomendar contenidos difíciles de analizar.**
- **Recomendar ítems basados en las preferencias del usuario**
- **Realizar recomendaciones válidas pero no esperadas, lo cual puede ser de gran utilidad.**

Sistemas Basados en Contenido

Matriz de Ratings - Escala:1-5

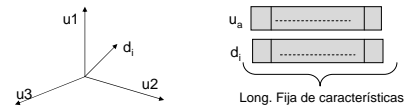
- **Recomienda ítems que son similares a los ítems que previamente valoró el usuario.**

	Matrix	The Net	I Robot	A.I.
Alice	?	2	5	4
Bob		4	5	5
Cindy	2	4	2	
David	3		5	2



Sistema de Recomendación Basado en Contenido

- Los ítems se definen según sus características. (Ej: palabras en el documento)
- El perfil de usuario se basa en cómo el usuario valora esas características.
- Este recomendador representa los documentos como las técnicas de IR, haciendo uso del espacio vectorial y los índices para representar usuarios y documentos.



Desventajas Sistemas Basados en Contenido.

- Sobreespecialización
 - ✓ El usuario está limitado a que le recomienden ítems similares a los que recomendó.
- Subjetividad de los Contenidos.
 - ✓ Dificultad en dominios con contenido difícil de analizar, (audio, gráficos, imágenes, video)
- Problema del Usuario Nuevo.
 - ✓ El usuario tiene que rankear un número suficiente de ítems para que el sistema pueda realmente entender sus preferencias.

Desventajas Sistemas Basados en Contenido.

- Representación del Perfil del Ítem
 - ✓ Para cada ítem se extraen ciertas características sobre las cuales se evalúa la similaridad.
- Efecto Portfolio.
 - ✓ Se da en dominios como recomendación de **noticias**, ya que es posible descartar noticias que pueden ser muy similares a previas, pero que al mismo tiempo presentan hechos nuevos e importantes.
- Problema Estabilidad vs Plasticidad.
 - ✓ Es difícil para el sistema aprender a adaptarse, a los cambios en el perfil del usuario hasta no haber recolectado un número suficiente de ratings actualizados.
- Tarea Onerosa de tener que rankear.

Ventajas de Sistemas Basados en Contenido.

- **Recomendación por contenido**

- ✓ Y no por opiniones subjetivas de otros usuarios.

- **Se Explica**

- ✓ El sistema puede generar explicaciones sobre la recomendación que hizo en base al historial del usuario.

- **No hay Dispersión (Sparsity)**

- ✓ Pues el modelado de la información está presente en las características del documento y no necesitan proveerlas otros usuarios.

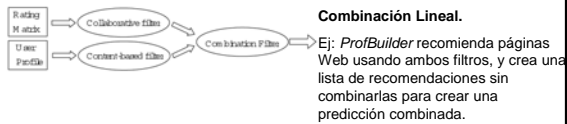
Sistemas Híbridos

- Los sistemas híbridos explotan características de los sistemas Basados en Contenido y Colaborativos, debido a la naturaleza complementaria de ambos. Lo que se busca es sobrellevar los inconvenientes de ambos sistemas para obtener mejores recomendaciones.

Para crear un **sistema híbrido colaborativo basado en contenido**, los perfiles de usuario se mantienen según el análisis de los contenidos de los ítems, y directamente se comparan esos perfiles para determinar las similitud entre usuarios para una recomendación colaborativa.

Sistemas Híbridos

Se han propuesto varios esquemas de sistemas Colaborativos basados en contenido, que pueden categorizarse en dos grupos:



Combinación Secuencial. Ej.: *FAB*



El Sistema FAB

- Nace en la Universidad de Stanford en 1994
- Basado en agentes inteligentes para descubrir nuevas páginas de interés.
- En 1997, se libera para uso público.
- En 1998, se introdujo la interfaz *Slider*, para inferir "feedback" a partir de las acciones drag/drop/clic de los usuarios
- Grupos de Investigación:
 - El Proyecto de Biblioteca Digital de Stanford
 - Grupo "MULTIAGENT"
 - Grupo HCI "Human-Computer Interaction"

(Fuente: Balabanovic and Yoav Shoham)

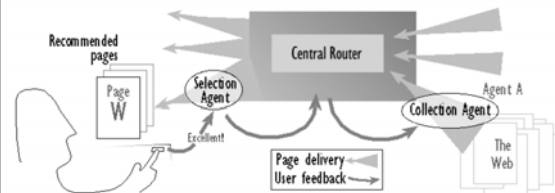
El Sistema FAB



Para este ejemplo vamos a particionar el proceso de recomendación en dos etapas: **colección** y **selección**

(Fuente: Balabanovic and Yoav Shoham)

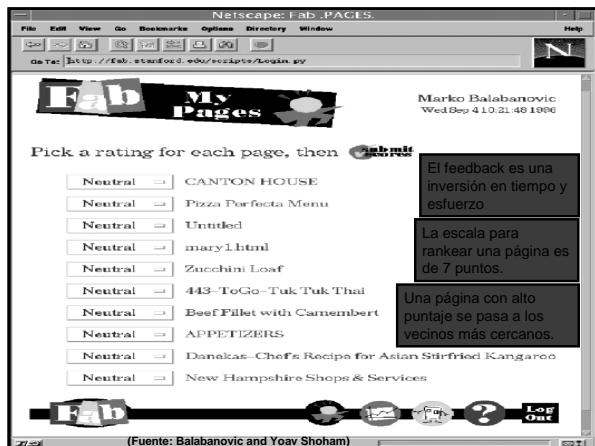
Arquitectura del Sistema FAB



Agentes de Colección: recolecta de la Web, páginas para un tópico específico

Agentes de Selección: recupera páginas para un usuario específico a partir del sistema de información.

(Fuente: Balabanovic and Yoav Shoham)



(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

- La construcción de **perfiles exactos** es una tarea importante.
- Los intereses de los usuarios los van a servir varios agentes de colección.
- Los agentes de colección se adaptan dinámicamente a la población de usuarios, luego las especializaciones de estos agentes deben ser dinámicas.

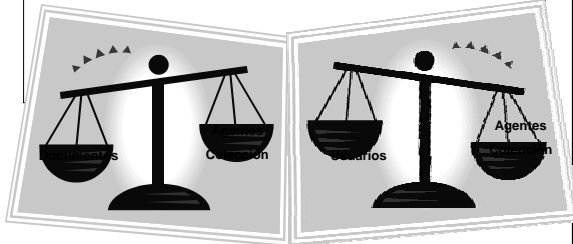


- Agentes poco exitosos por exitosos.

(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

FAB se compromete en dos actos diferentes y simultáneos de carga-balance reflejado en los dos conjuntos de vínculos cambiantes dinámicamente:



(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

FAB implementa distintos tipos de agentes de colección:

- **De Búsqueda:** realizan una *búsqueda best-first de la Web*, tratando de encontrar las páginas que mejor mapeen con sus perfiles.
- **De Índice:** construyen consultas para pasar a motores de búsqueda Web con indexado exhaustivo previamente realizado.
- Otros
 - Agentes que toman páginas al azar
 - Agentes que recolectan sitios "cool" del día
 - Agentes que intentan servir a un "usuario promedio".

(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

El sistema exhibe todas las ventajas de un sistema híbrido:

- Puede usar las experiencias de otros, en vez de un análisis impreciso e incompleto de contenidos. (vent. Colaborativa)
- Puede tratar con ítems no vistos por otros. (vent. B-Conten.)
- Permite hacer buenas recomendaciones a usuarios, usando los perfiles construidos a partir del contenido de ítems, aun si no hubieran usuarios similares.
- Los usuarios que no han calificado los mismos ítems se benefician de recomendaciones colaborativas.
- Al usar *feedback de grupo* se requieren potencialmente menos ciclos para alcanzar el mismo nivel de personalización.

(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

Además, la adaptación de los agentes de colección trae los siguientes beneficios:

- **Poder instanciar un número más chico que la población de usuarios, tal vez un número fijo.** Esto permitirá escalar la población de usuarios y documentos.
- **Los agentes de colección identifican comunidades emergentes de interés en la población de usuarios,** permitiendo a FAB soportar interacciones sociales entre usuarios similares y automáticamente proveer recomendaciones individuales y grupales.

(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

Experimentos

...Balabanovic et.al presenta tres experimentos: dos estadísticos y uno anecdótico.

Se utilizó un número pequeño de usuarios, y los test se realizaron sobre páginas Web del mundo real, con los tópicos: gráficos por computadora y programación de juegos - catalogación de biblioteca y clasificación - música post industrial - información de deportes y juegos - cultura americana nativa - cocina - música de los 60's - excursionismo - evolución.

Se utilizó una medida para medir distancia entre un perfil de usuario y los intereses reales del usuario:

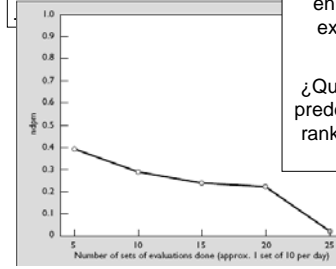
NDPM: (Normalized Distance-Based Performance Measure)

Capacidad de ordenar correctamente los items de interesante a no interesante

(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

Primer Experimento



La construcción de **perfiles exactos** es una cuestión importante para recomendaciones basadas en contenido. Con este experimento se quiere determinar:

¿Qué tan bien FAB puede predecir para un usuario, el ranking de un conjunto de items?

Distancia entre rankings actuales y predcidos, promediado sobre todos los usuarios en cada punto de evaluación.

(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

Hecho Anecdótico

A medida que se testeaba el sistema, hubo una evidencia anecdótica del sistema. En un caso claro de especialización automática, un agente se volvió experto en cocina: el 77% de las 400 top-terms en su perfil estaba relacionado con la cocina.

tablespoon	2.95	sprinkl	1.95	tomato	1.58
teaspoon	2.44	saut	1.92	cup	1.51
onion	2.16	chop	1.92	str	1.44
flour	2.13	parslef	1.92	preheat	1.37
minc	2.09	saucepan	1.79	pepper	1.34
garlic	2.06	sauc	1.71	parmesan	1.33
clove	2.00	butter	1.59		

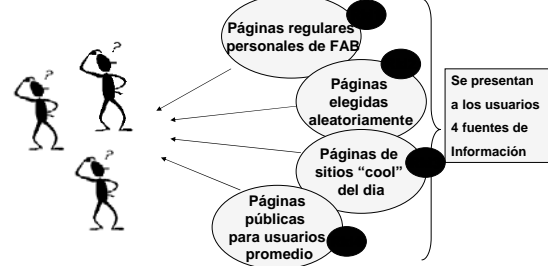
Conclusión: Los agentes se pueden especializar en tópicos específicos con el tiempo, y automáticamente converger a áreas de solapamiento entre los usuarios.

(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

Segundo Experimento

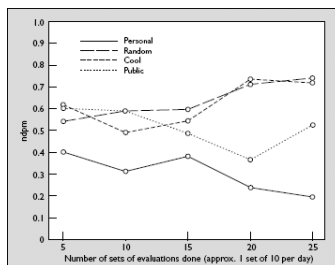
Mide la Performance General del sistema FAB.



(Fuente: Balabanovic and Yoav Shoham)

El Sistema FAB

Segundo Experimento



Ranking Ideal Rki de algún lote de páginas para la fuente S, es uno donde el usuario prefiere cada página de S para cada página que no es de S.

>> Preferencia por páginas de S, << va a ser la distancia ndpm entre el ranking real del usuario y el ranking ideal.

(Fuente: Balabanovic and Yoav Shoham)

SR: Dominios

NAME	REFERENCES	DOMAIN	
ACR News	[Molashar et al., 2000]	Netnews Filtering	
Amazon	[Amazon]	Comerc electronic	
Amalhasa	[Moulas, 1997]	Web Recommender	
Anatagonomy	[Sakagami et al., 1997]	Personalized Newspaper	
Beehive	[Huberman and Kaninsky, 1996]	Sharing News	
Belcore Video Recommender	[Hill et al., 1995]	Movie Recommender	
Casnar	[Berney and Ferneley, 1999]	Document Recommender	
CDNow	[CDNow]	Comerc electronic	
Fab	[Balabanovic and Shoham, 1997]	Web Recommender	
GroupLens	[Resnick et al., 1994]	Netnews Recommender	
IFWeb	[Manis and Tasso, 1996]	[Aucinar and Tasso, 1997]	Web Recommender
InfoFinder	[Krubwisch and Burkey, 1995]	Information Recommender	
INFormer	[Riordan and Sorensen, 1995]	[Sorensen et al., 1997]	Netnews Filtering
Krakatos Chronicle	[Kamba et al., 1995]	Personalized Newspaper	
LabalP	[Schwab et al., 2001]	Document Recommender	
Let's Browse	[Lieberman et al., 1999]	Web Recommender	
Letizia	[Lieberman, 1995]	Web Recommender	
LifeStyle Finder	[Krubwisch, 1997]	Purchase, Travel and Store Recommender	
MovieLens	[Good et al., 1999]	Movie Recommender	
News Dada	[Bilous and Pazzani, 1999]	Netnews Recommender	
NewsWeeder	[Lang, 1995]	Netnews Recommender	

SR: Dominios

System Name	References	Domain
Personal WebWatcher	[Mladenic, 1996]	Web Recommender
PSUN	[Sorpsen and McEligot, 1995]	Netnews Recommender
Re-Agent	[Boone, 1998]	E-mail Filtering
Recommender	[Bass et al. 1998]	Movie Recommender
Ringo - Firefly	[Shardanand and Maes, 1995], [Shardanand, 1994]	Music Recommender
SIFT Netnews	[Yan and Garcia-Molina, 1995]	Netnews Filtering
SiteF	[Stefani, and Strappavara, 1998]	Web Recommender
Smart Radio	[Hayes and Cunningham, 1999], [Hayes and Cunningham, 2000]	Music Recommender
Syskill & Webert	[Pazzani et al., 1996], [Pazzani and Billsus, 1997]	Web Recommender
Tapestry	[Goldberg et al., 1992]	E-mail Filtering
Webmate	[Chen and Sycara, 1998]	Web Recommender
WebSail	[Chen et al., 2000]	Web Search Filtering
WebSell	[Cunningham et al., 2001]	Purchase Recommender
WebSite	[Cooley et al., 1999]	Web Recommender
WebWatcher	[Armstrong et al., 1995], [Joschims et al., 1997]	Web Recommender

SR: Dominios

NAME	REFERENCES	DOMAIN
ACR News	[Mobasher et al., 2000]	Netnews Filtering
Amazon	[Amazon]	Commerce electronic
Amalblue	[Nofias, 1997]	Web Recommender
Anatagonomy	[Salganani et al., 1997]	Personalized Newspaper
Beehive	[Huberman and Kaminsky, 1996]	Sharing News
Bellcore Video Recommender	[Hill et al., 1995]	Movie Recommender
Casmir	[Berney and Ferneley, 1999]	Document Recommender
CDNow	[CDNow]	Commerce electronic
Fab	[Balabanovic, and Shoham, 1997]	Web Recommender
GroupLens	[Resnick et al., 1994]	Netnews Recommender
iWeb	[Minio and Tasso, 1996], [Assicar and Tasso, 1997]	Web Recommender
InfoFinder	[Krubwisch and Burkey, 1995], [Krubwisch and Burkey, 1996]	Information Recommender
INFORMer	[Riordan and Sorensen, 1995], [Sorensen et al., 1997]	Netnews Filtering
Krakatoa Chronicle	[Kamiba et al., 1995]	Personalized Newspaper
Labo'Cr	[Schwab et al., 2001]	Document Recommender
Let's Browse	[Lieberman et al., 1999]	Web Recommender
Letizia	[Lieberman, 1995]	Web Recommender
LifeStyle Finder	[Krubwisch, 1997]	Purchase, Travel and Store Recommender
MovieLens	[Good et al., 1999]	Movie Recommender
News Dale	[Billsus and Pazzani, 1999]	Netnews Recommender
NewsWeeder	[Lang, 1995]	Netnews Recommender

SR vs IR

VALOR DE REFERENCIA	SR	IR
Necesidad de satisfacer una Φ requerida	La Φ que el usuario necesita se especifica en forma de perfil de usuario.	Los usuarios especifican sus necesidades en forma de consultas.
Uso y evaluación del documento recuperado	El sistema aprende del comportamiento del usuario y actualiza su perfil (y por lo tanto las recomendaciones) adecuadamente.	La evaluación puede conducir a la modificación de la consulta en un proceso llamado <i>relevance feedback</i> .
Frecuencia de uso	Depende del sistema. Algunos SR (ej. de noticias) mantienen registros de perfiles de usuarios de largo término; otros, como web browsing, no.	Diseñado para un uso ad-hoc para usuarios con metas de búsqueda a one-time
Proceso de determinación de la info del user	El usuario no se involucra activamente en recuperar Φ relevante una vez que el perfil está definido	El usuario activamente especifica lo que necesita en una consulta

SR vs IR

VALOR DE REFERENCIA	SR	IR
Meta	Selección y Presentación de la Φ más relevante para el usuario a medida que él interactúa con el sistema.	Selección de datos relevantes a partir de una base de datos.
Origen de los datos	Dinámico	Relativamente Estático, aunque con el crecimiento de la Web, los motores de búsqueda necesitan y hacen monitoreo constantemente del contenido disponible en la Web.
Tipos de Usuarios	Conocido por el sistema por sus perfiles	No conocidos por el sistema
Cuestiones de Privacidad	Importante ya que los usuarios están registrados	No es de mucha importancia