Ingeniería Informática

FACULTAD DE INGENIERIA
UNIVERSIDAD NACIONAL DE JUJUY

OPTATIVA

Recuperación Avanzada de Información

Dr. J. Federico Medrano
@jfedemedrano

Unidad N° 1 – Parte 3

Temas a desarrollar

- ~~*Organización de la asignatura*~~
- ~~*Introducción.*~~
- ~~*Documentos electrónicos.*~~
- ~~*Modelos de recuperación de información.*~~
- ~~*Algoritmos y estructuras básicas.*~~
- ~~*La recuperación de información en Internet.*~~
- ~~*La recuperación multilingüe.*~~
- ~~*Sistemas hipertextuales y recuperación documental.*~~
- ~~*La recuperación de documentos multimedia o no textuales.*~~
- La recuperación basada en la citación.
- Sistemas de filtrado y recomendación

La Recuperación basada en la citación

Minería Web

- Uno de los sistemas con mayor publicación de datos sin restricciones de su contenido y de acceso libre es la Web. Ésta, tiene características únicas (Bing Liu, 2007) como (a) la existencia de distintos tipos de datos (audios, videos, textos, etc), (b) la información en las páginas Web es muy variada, dinámica y tiene ruido, (c) ***una pequeña cantidad de información está enlazada*** y (d) es de servicios y también es una sociedad virtual en la cual extraer información útil conlleva a una serie de problemas de índole multi-disciplinaria.

Minería Web

- El proceso de Minería Web (MW) puede ser definido formalmente como “el proceso global de descubrir información o conocimiento potencialmente útil y previamente desconocido a través de los datos de la Web” (Jeria and Escobar, 2007).
- La MW tiene como objetivo descubrir información útil o el conocimiento tanto del contenido de documentos Web, como también de la estructura de hipervínculos Web y los datos de uso.

Minería Web

- Las tareas de MW se clasifican en tres categorías: MW de contenido, MW de estructura y MW de uso.

1. ***La primera extrae información del contenido en los documentos Web***.
2. La segunda, MW de estructura, descubre un modelo a partir de la topología de enlaces de la red. Este modelo puede ser útil para clasificar o agrupar documentos. **ARS**
3. Y por último MW de uso extrae información (hábitos, preferencias, etc. de los usuarios o contenidos y relevancia de documentos) a partir de las sesiones o registros de uso en la Web (minería de web Logs).

Citas y Referencias bibliográficas

- Una *cita* es la mención a un texto, idea o frase ajena, envía al lector a la fuente de donde se sacó la información y está presente en la *referencia bibliográfica*.
- En el apartado de la bibliografía, todas las *referencias* se presentan ordenadas alfabéticamente por apellido sin diferenciar en papel o electrónicas, sin viñetas, con sangría francesa y, especialmente si es extensa, con cuerpo de letra menor.

Citas y Referencias bibliográficas (Ejemplo)

Modelo de Espacio Vectorial

Es el modelo de IR quizás más conocido y el más utilizado. Plantea la necesidad de utilizar una función de similitud entre el documento y la consulta introduciendo un ranking en los documentos recuperados (Bing Liu, 2007).

Referencias

Liu, B. (2007). "Web Data Mining, Exploring Hy-perlinks, Contents, and Usage Data" First Edition. Chicago: University of Illinois.

Tolosa, G. y Bordignon, F. (2008). "Introducción a la Recuperación de Información" UNLu.

Google Scholar / Google Académico

Information retrieval on the web [PDF] acm.org

M Kobayashi, K Takeda - ACM Computing Surveys (CSUR), 2000 - dl.acm.org

In this paper we review studies of the growth of the Internet and technologies that are useful for **information** search and **retrieval** on the Web. We present data on the Internet from several different sources, eg, current as well as projected number of users, hosts, and Web sites ...

☆ 99 Cited by 883 Related articles All 18 versions

Information retrieval on the web

☐ Search within citing articles

Inverted files for text search engines [PDF] acm.org

J Zobel, A Moffat - ACM computing surveys (CSUR), 2006 - dl.acm.org

The technology underlying text search engines has advanced dramatically in the past decade. The development of a family of new index representations has led to a wide range of innovations in index storage, index construction, and query evaluation. While some of ...

☆ 99 Cited by 1400 Related articles All 26 versions

Personalized e-learning system using item response theory [PDF] psu.edu

CM Chen, HM Lee, YH Chen - Computers & Education, 2005 - Elsevier

Personalized service is important on the Internet, especially in Web-based learning. Generally, most personalized systems consider learner preferences, interests, and browsing behaviors in providing personalized services. However, learner ability usually is neglected ...

☆ 99 Cited by 551 Related articles All 10 versions

Microsoft Academic | Coronavirus

FILTER BY:

Time

FROM 1970 TO 2020

Top Topics

- Coronavirus
- Virology

Showing 1-10* of 10513 (0.356 seconds)

VIEW SORT BY RELEVANCE

Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia

702 citations*

2020 THE NEW ENGLAND JOURNAL OF MEDICINE

Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou see all 45 authors

Virology | Transmission (mechanics) | Surgery View More (5+)

Abstract Background The initial cases of novel coronavirus (2019-nCoV)–infected pneumonia (NCIP) occurred in Wuhan, Hubei Province, China, in December 2019 and January 2020. We analyzed data on the...

REFERENCES CITED BY RELATED

1-10* of 393

VIEW SORT BY RELEVANCE

Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China

416 citations*

2020 JAMA

Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu see all 14 authors

Wuhan University

Surgery | Oxygen therapy | Moxifloxacin View More (8+)

Importance In December 2019, novel coronavirus (2019-nCoV)–infected pneumonia (NCIP) occurred in Wuhan, China. The number of cases has increased rapidly but information on the clinical characteristics of affected patients is limited. Objective To describe the epidemiological and clinical characte... View Full Abstract

Algunas aplicaciones

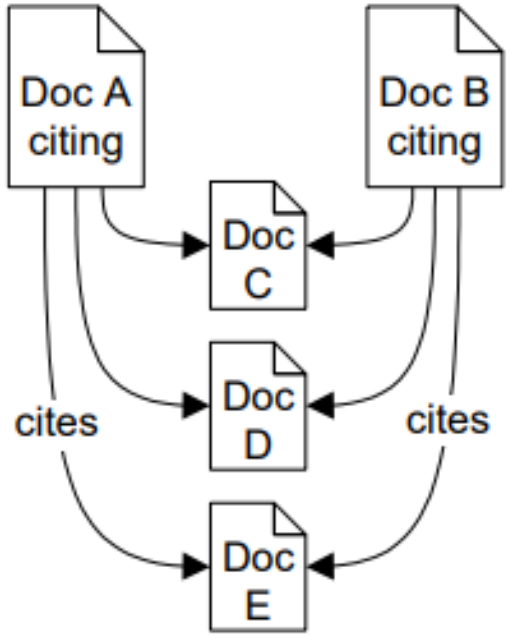


Figure 1: Bibliographic coupling

Documents are bibliographically coupled if they cite one or more documents in common. Figure 1 illustrates this approach: Papers A and B are related because they both cite papers C, D and E.

In contrast, two documents are "co-cited" when at least one paper cites both. This approach is illustrated in Figure 2: Papers A and B are related because they are both cited by papers C, D and E. The more co-citations two papers receive, the more related they are [6].

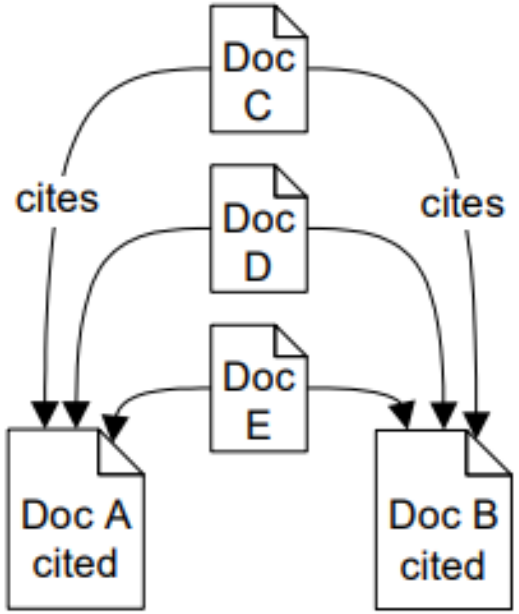
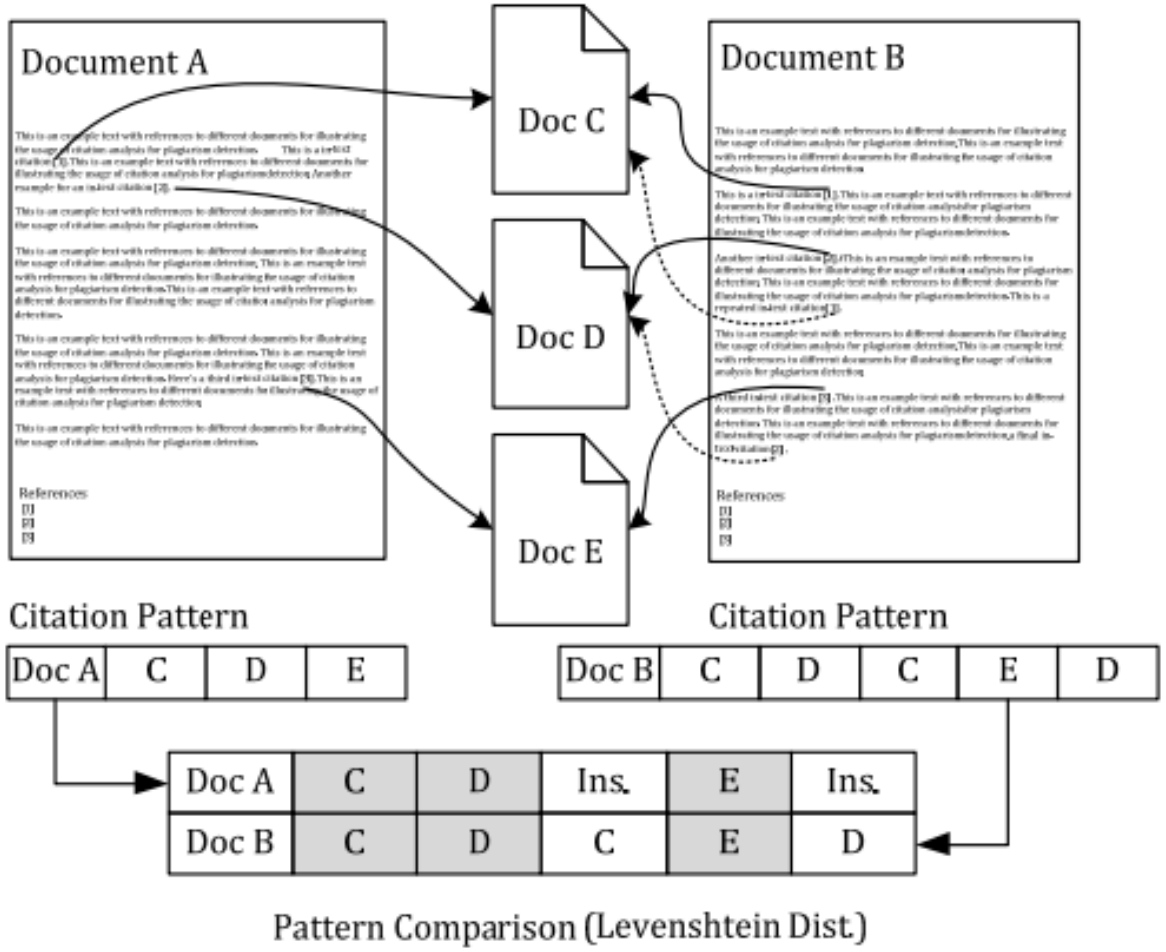


Figure 2: Co-citation analysis

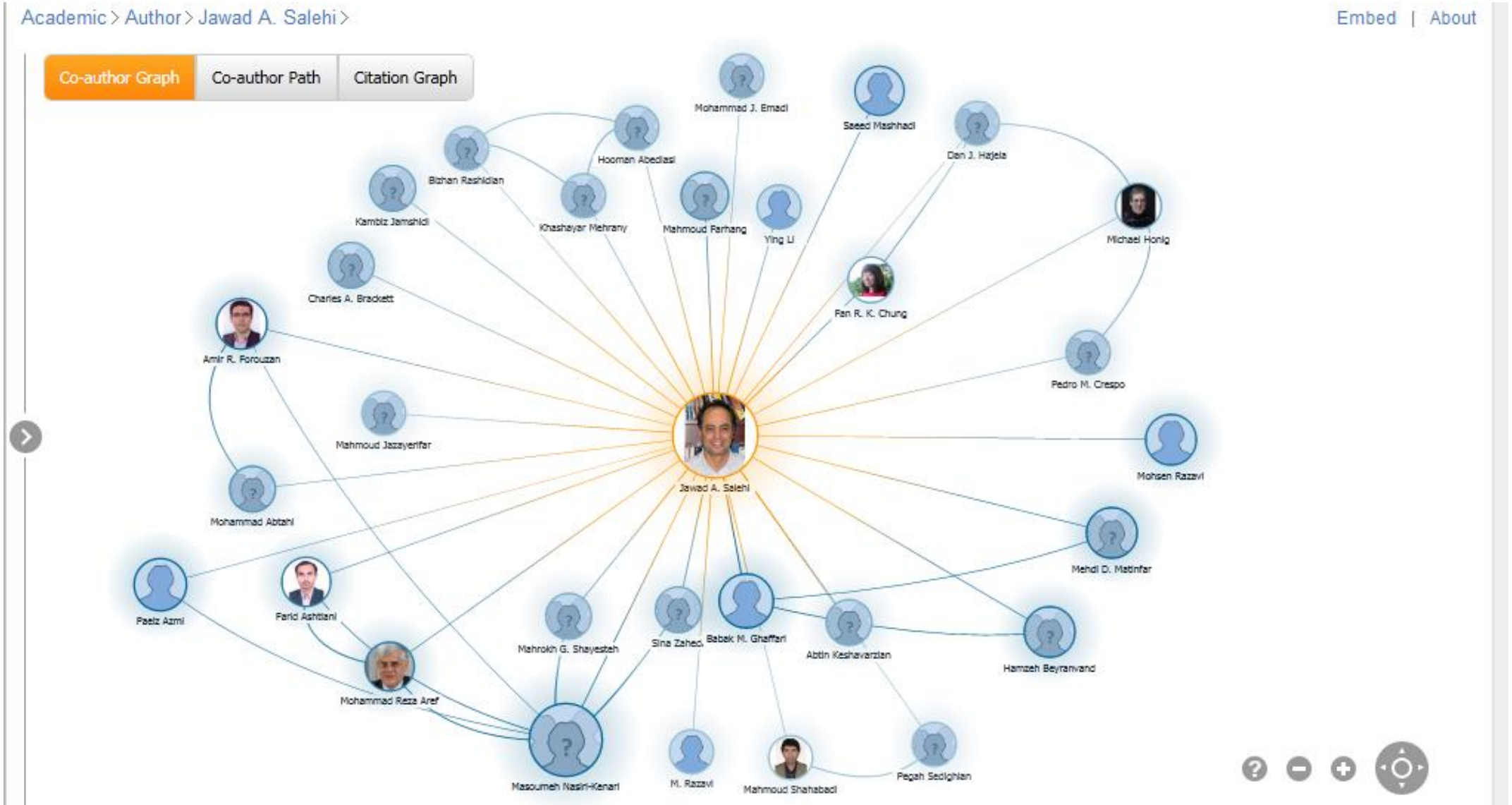
- Citation Proximity Analysis (CPA)

Algunas aplicaciones

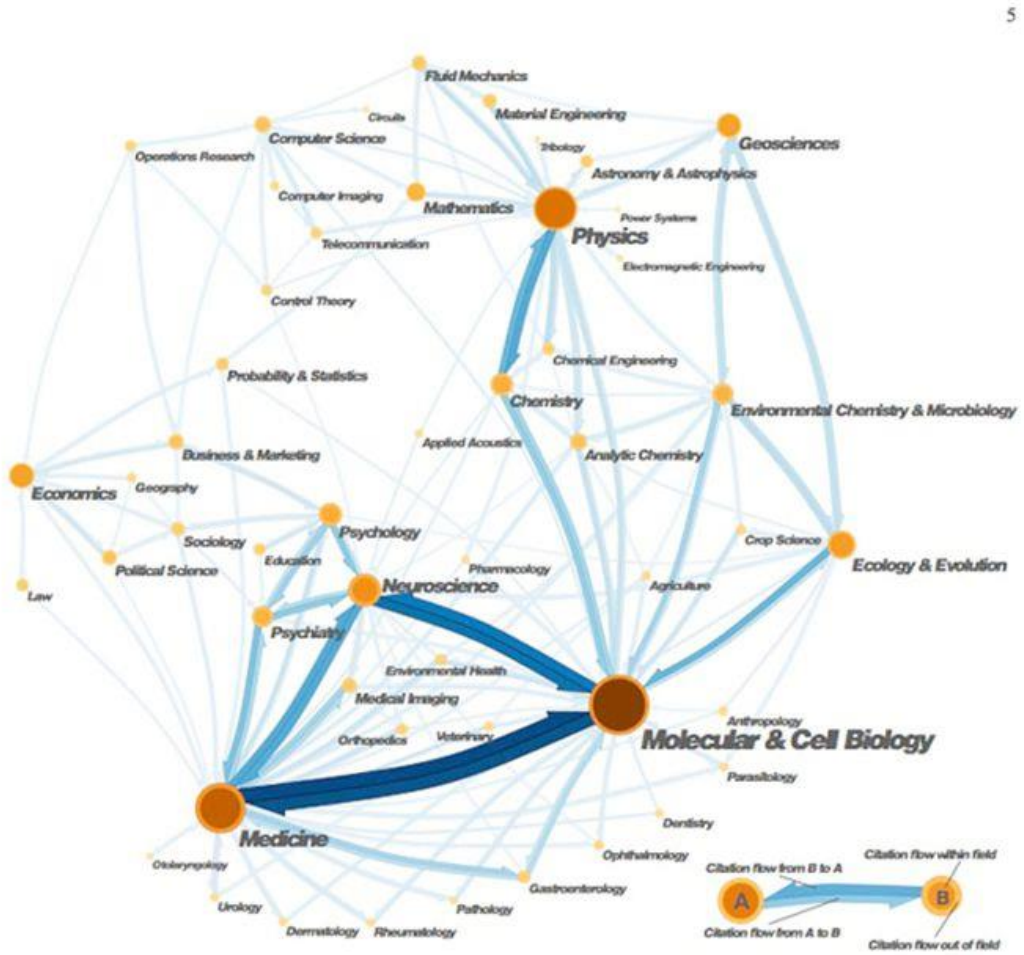


- [Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection]

Algunas aplicaciones

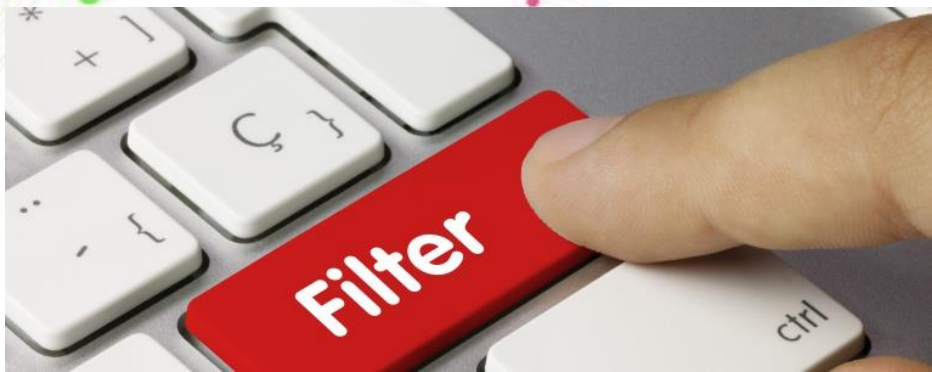


Algunas aplicaciones



Sistemas de filtrado y recomendación


CUSTOMERS WHO BOUGHT THIS ITEM:
ALSO BOUGHT:
piccolo


Filter
ctrl

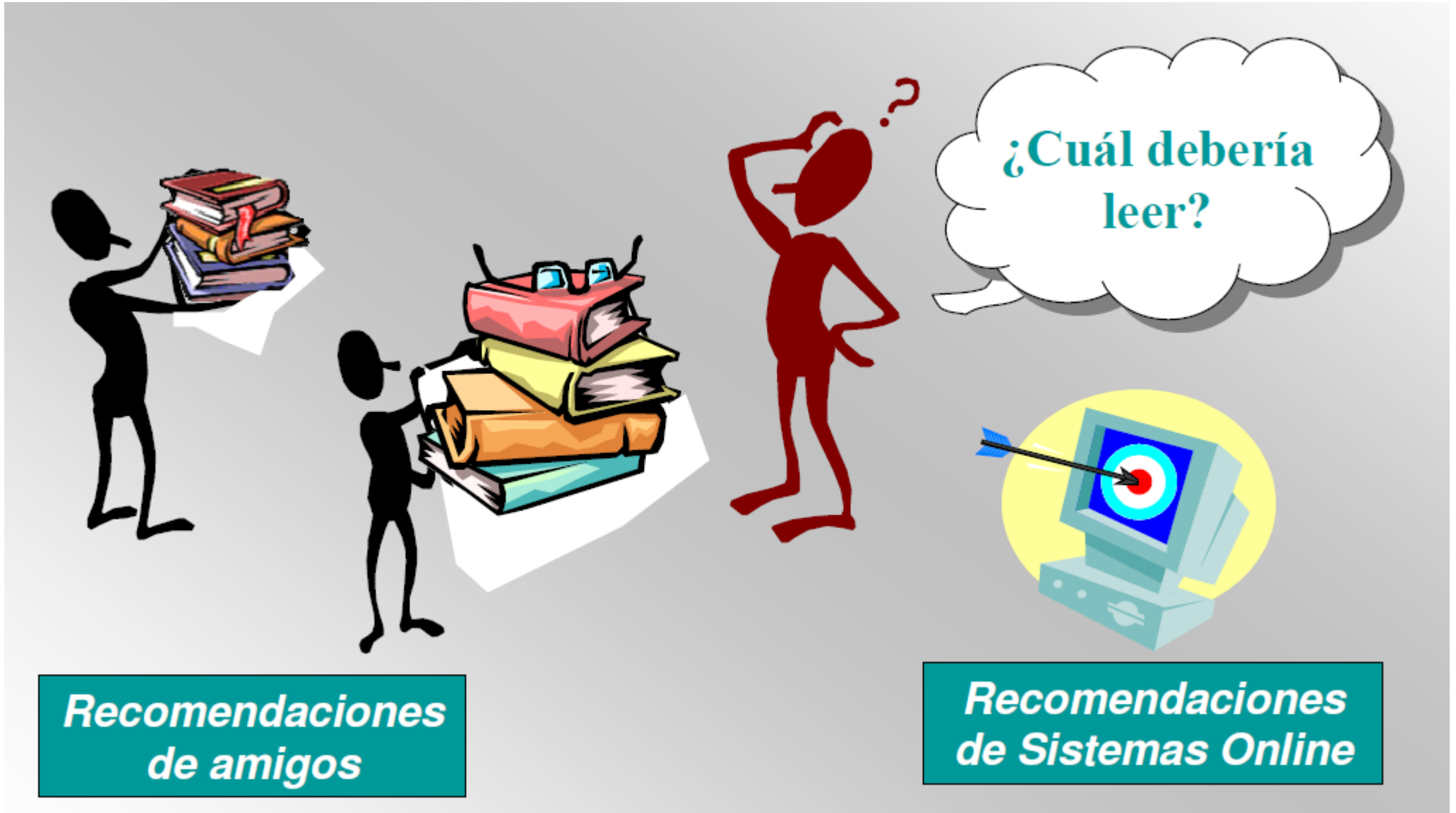
Filtrado Colaborativo

- Permite hacer predicciones automáticas sobre los intereses de un usuario mediante la recopilación de las preferencias o gustos de información de muchos usuarios
- Desventajas: *Cold start, new-item, transparencia*

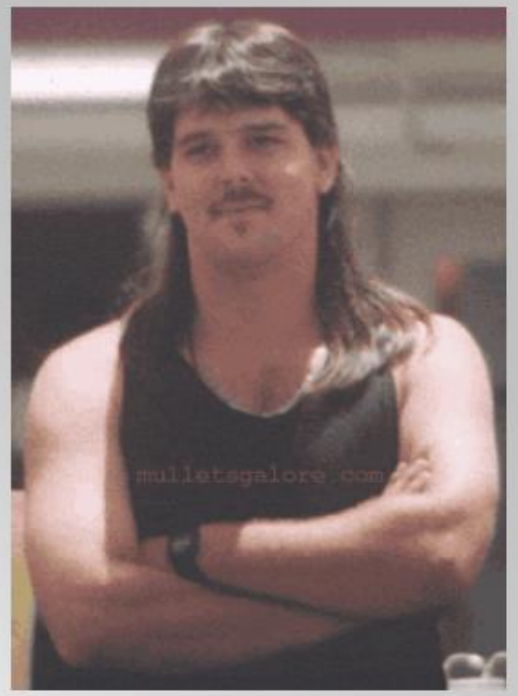
Filtrado Basado en el Contenido

- El contenido desempeña un papel principal en el proceso de recomendación
- Los elementos están representados por características: palabras sueltas, frases o n-grams

Sistemas de recomendaciones: una propuesta para la mejora de la recuperación de información



Sistemas de recomendaciones: una propuesta para la mejora de la recuperación de información



Usuario/Cliente A

- Compra CD de Metalica
- Compra CD de Megadeth

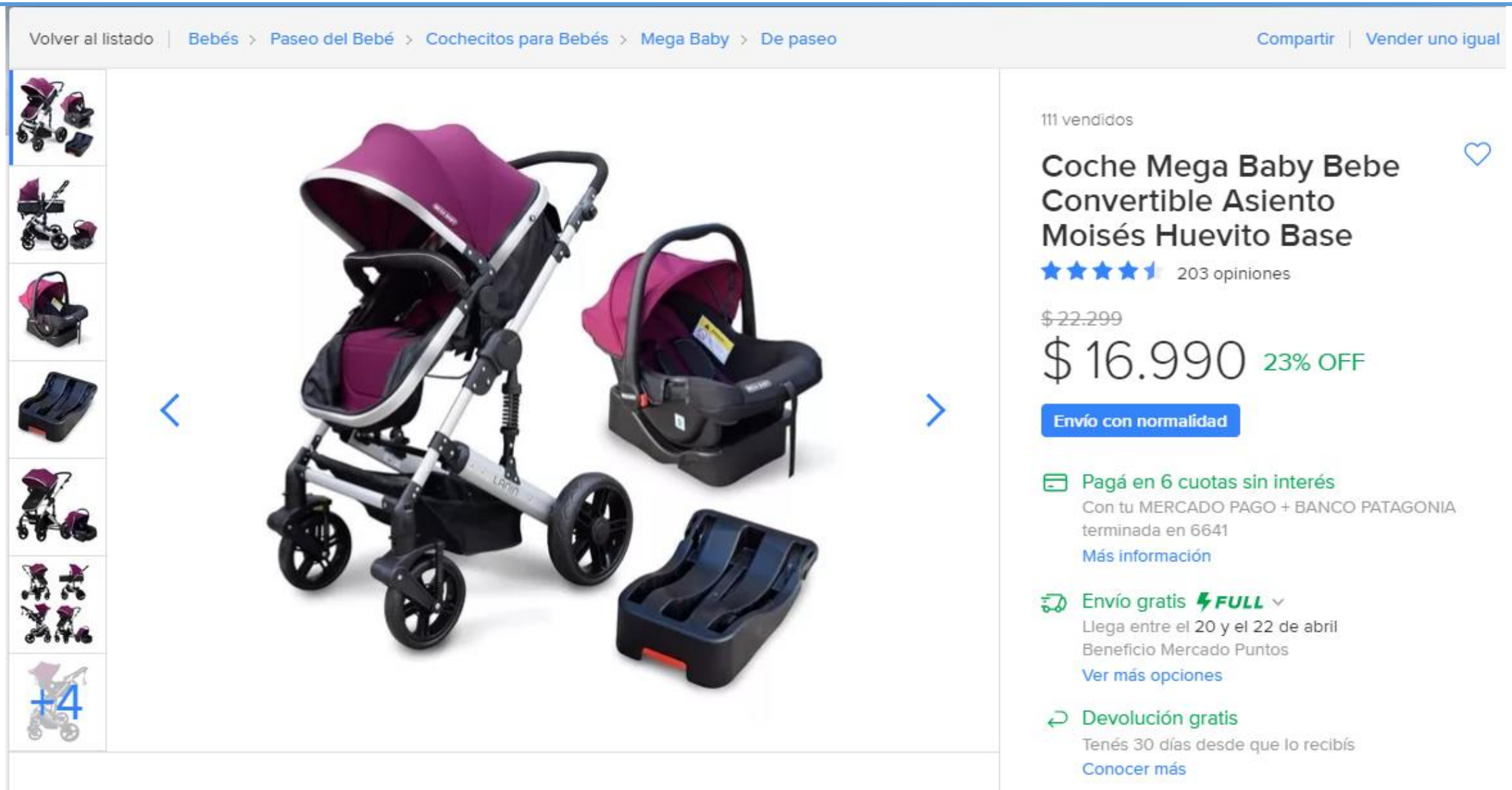
Usuario/Cliente B

- Busca acerca de Metalica
- El sistema de recomendación le sugiere Megadeth a partir de los datos recogidos del usuario A

Sistemas de recomendación

- El objetivo de un sistema de recomendaciones es guiar a un usuario mediante recomendaciones a aquellos productos y/o servicios más atractivos para él.
- De esta forma, el usuario empleará menos tiempo en encontrar lo que necesita, lo hará de una forma más rápida y cómoda, y posiblemente encontrará otros productos y/o servicios interesantes para él. El éxito de muchas tiendas de comercio electrónico (ej: Amazon www.amazon.com) se han basado en la utilización de este tipo de herramientas.

Sistemas de recomendación



Sistemas de recomendación

Quienes compraron este producto también compraron

$1.490

Booster Elevador 15 A 36 Kg Corato Belluno Baby Sin

$6.790 28% OFF

Envío gratis FULL

Butaca Silla Para Auto De Bebé Mega Baby 0 A 25 Kg

$2.500

18x $ 138[89] sin interés

Envío gratis

Mochila Bolso Maternal Felix Happy Little Moments Oficial

$5.029

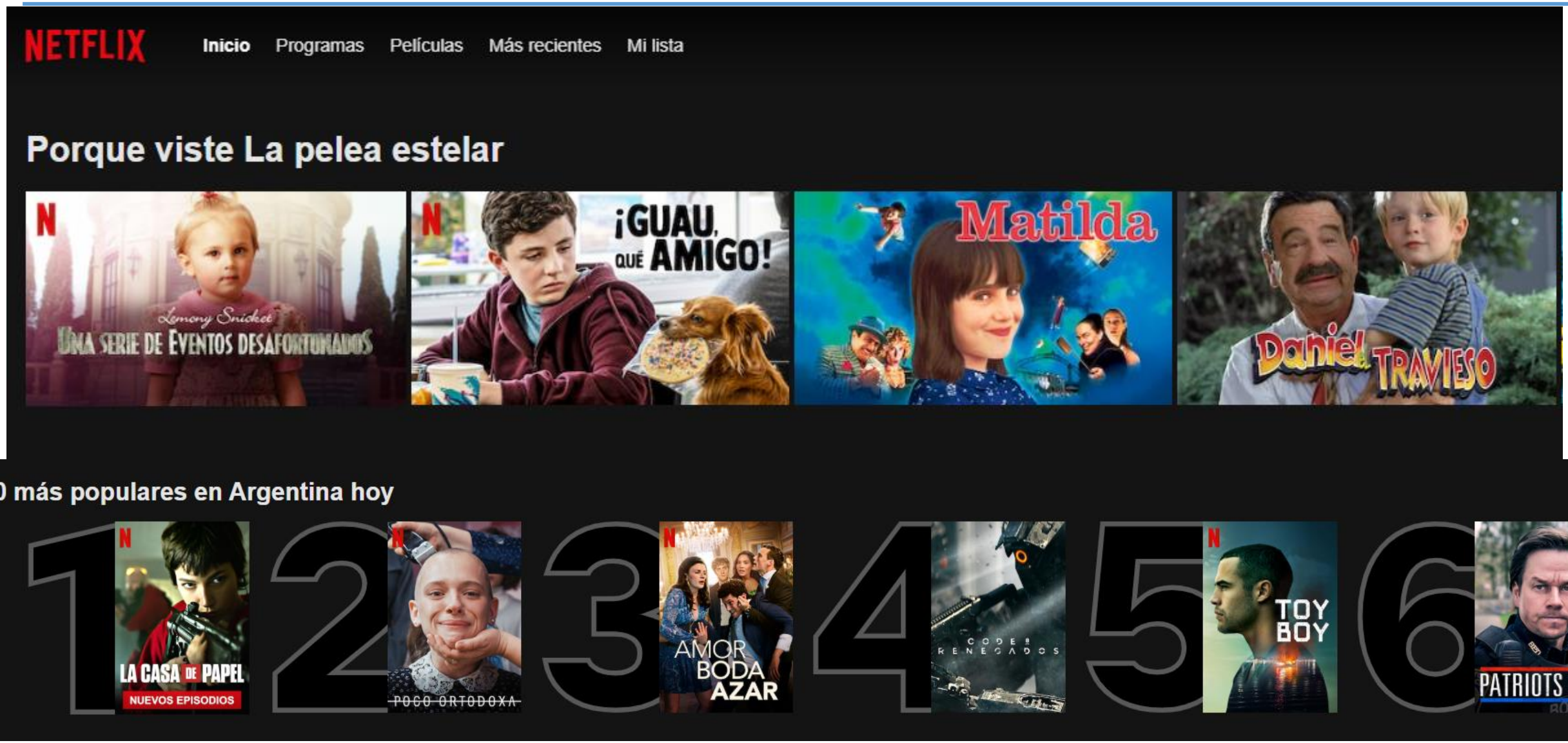
Envío gratis

Butaca Silla Auto Bebe Booster De 9 A 18 Kg Reclinable

$1.950

Bolsos Maternales Lunares Mamaflora + Cambiador/bebe

Sistemas de recomendación



Sistemas de recomendación

- Los **sistemas de recomendaciones** en el ámbito de la **RI** son herramientas que **asisten a los usuarios** en sus procesos de búsqueda de información, ayudando a **filtrar los ítems** de información recuperados y proponiendo **recomendaciones** sobre esos ítems a partir de las preferencias y **opiniones** proporcionadas por otros usuarios o bien a partir de las **preferencias** del usuario objeto de la recomendación (o usuario activo).

Sistemas de recomendación

- **Los primeros algoritmos que se propusieron y los más sencillos para generar recomendaciones basadas en el filtrado colaborativo se denominan *Memory based* (basados en la memoria):**
- **Paso 1: Representar los datos de entrada**
- **Paso 2: Encontrar los vecinos más cercanos:**
  - **2.1. Calcular la similaridad entre vectores: Se mide la similitud de todos los usuarios con respecto al usuario activo.**
  - **2.2. Seleccionar a los vecinos: Se selecciona un conjunto apropiado de usuarios, según la similitud de los mismos con respecto al usuario activo.**
- **Paso 3: Generar las recomendaciones**
  - **Se usan las valoraciones de los vecinos para generar la valoración que, supuestamente, ofrecería el usuario activo.**
  - **Se normalizan las puntuaciones de los distintos usuarios y se calcula una predicción a partir de algún tipo de combinación ponderada de las puntuaciones asignadas al ítem por los usuarios seleccionados en el paso anterior.**

Sistemas de recomendación

	u_1	u_2	u_3	u_4	u_5	u_6
$item_1$	5	1	5	4	0	3
$item_2$	3	3	1	1	5	1
$item_3$	0	1	0	2	1	4
$item_4$	1	1	4	1	1	2
$item_5$	3	2	5	0	0	3
$item_6$	4	3	0	0	4	0
$item_7$	0	1	5	1	1	1

Sistemas de recomendación: Problemas

- A veces hay que recurrir a incentivos para la provisión de recomendaciones y la creación de perfiles de intereses, puesto que los usuarios no suelen estar muy dispuestos a colaborar proporcionando información personal sobre sus preferencias: forzar al usuario a introducir sus preferencias a cambio de recibir recomendaciones u otro tipo de compensaciones.
- Si cualquiera puede recomendar, los propietarios de determinados productos/servicios podrían generar recomendaciones positivas de los mismos y negativas de otros.

Sistemas de recomendación: Problemas

- Aspectos de privacidad y debido a que algunas personas no quieren que se conozcan sus hábitos o preferencias, se permite la participación anónima o bajo un pseudónimo.
- El mantenimiento de un SR es costoso, financiación: pagar por usar, inclusión de publicidad, cobrar una cuota a los evaluados.
- Dotar a los SR de mejores técnicas de representación de las preferencias o recomendaciones de los usuarios que nos permitan captar verdaderamente su concepto del objeto recomendado y mejorar la interacción SR-usuario.