Sistemas de recomendación con Surprise

Antonio David Pérez Morales
Rafael Haro Ramos

PyCon ES
ALICANTE 2019

@adperezmorales
@rafa_haro

adperezmorales
rafa_haro

Agenda

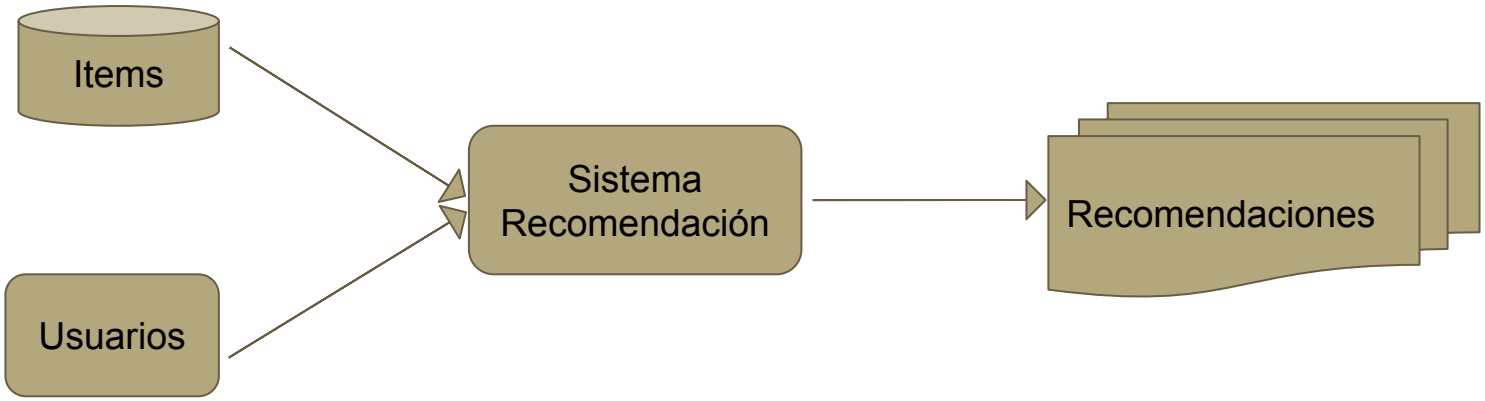
- Sistemas de Recomendación
- Tipos de Sistemas de Recomendación
  - Basados en contenido
  - Basado en filtrado colaborativo
    - Medidas de similitud
    - Cálculo de la predicción
    - Métodos de evaluación
  - Filtrado demográfico
  - Híbridos
- Librería Surprise
- Conclusiones

Sistemas de Recomendación

- Una de las aplicaciones de la ciencia de datos más usada y fáciles de entender
- Interés y demanda en este área aún vigente debido al auge de datos y sobrecarga de información
- Necesidad de ayudar a los usuarios a lidiar con la sobrecarga de información y proporcionar recomendaciones, contenido y servicios

Sistemas de Recomendación

- **Sistema inteligente** que proporciona a los usuarios una serie de **sugerencias personalizadas** (recomendaciones) sobre un determinado **tipo de elemento** (ítems) de un dominio particular



Sistemas de Recomendación

- Compara el perfil del usuario con algunas características de referencia del dominio de los elementos y busca predecir el “ranking” o puntuación que el usuario le daría a un elemento que aún el sistema no ha considerado para ese usuario.
- Características pueden estar basadas en la relación del usuario con el elemento o en el ambiente social del mismo usuario.
- Retroalimentación Información de Usuario: implícita, explícita, contextual

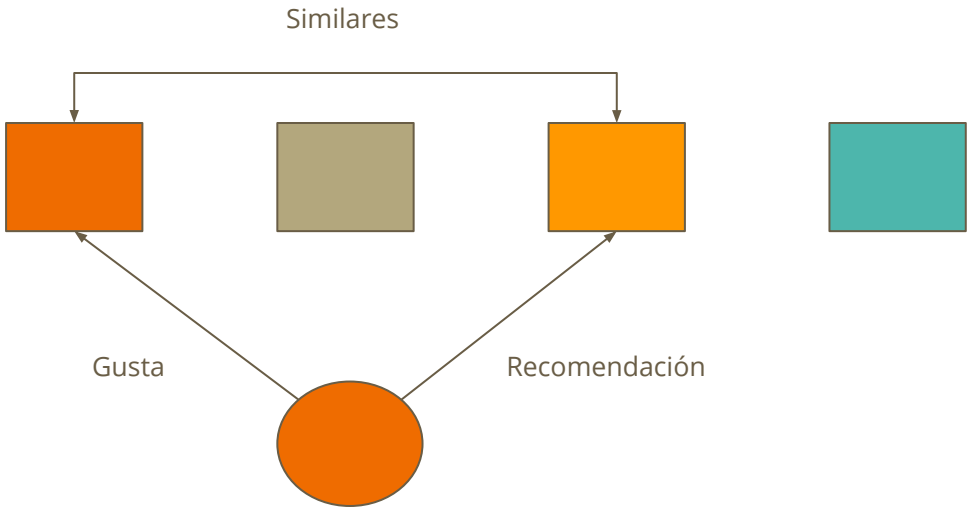
Tipos de Sistemas de Recomendación

Construcción de un Sistema de Recomendación

- Analizar los elementos
- Analizar los usuarios
- Analizar las relaciones entre elementos y usuarios

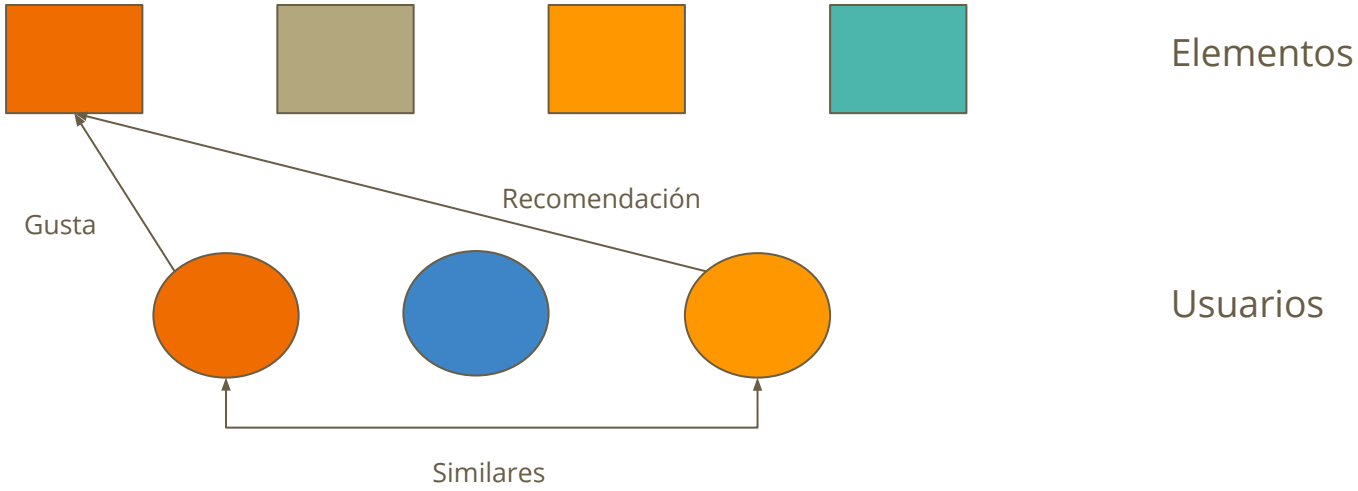
Tipos de Sistemas de Recomendación

- **Análisis de los elementos**

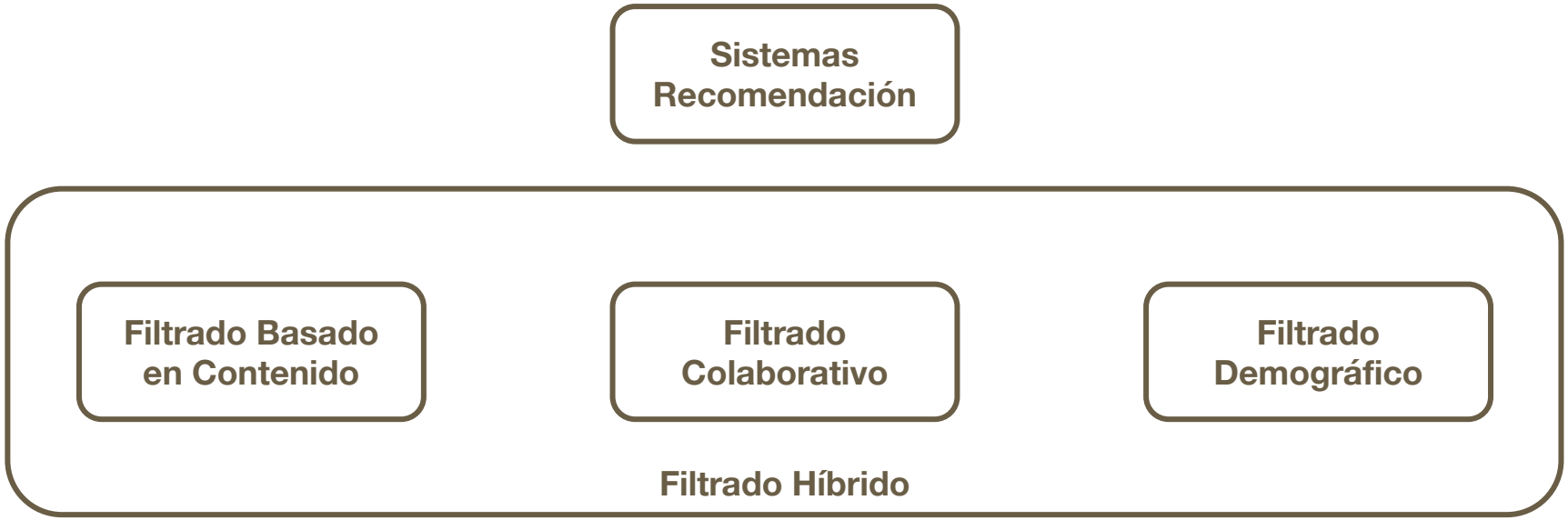


Tipos de Sistemas de Recomendación

- **Análisis de los usuarios**

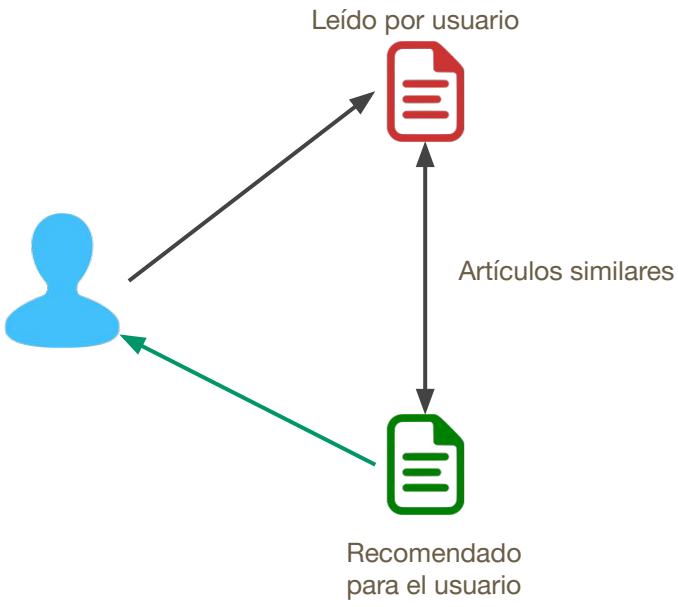


Tipos de Sistemas de Recomendación

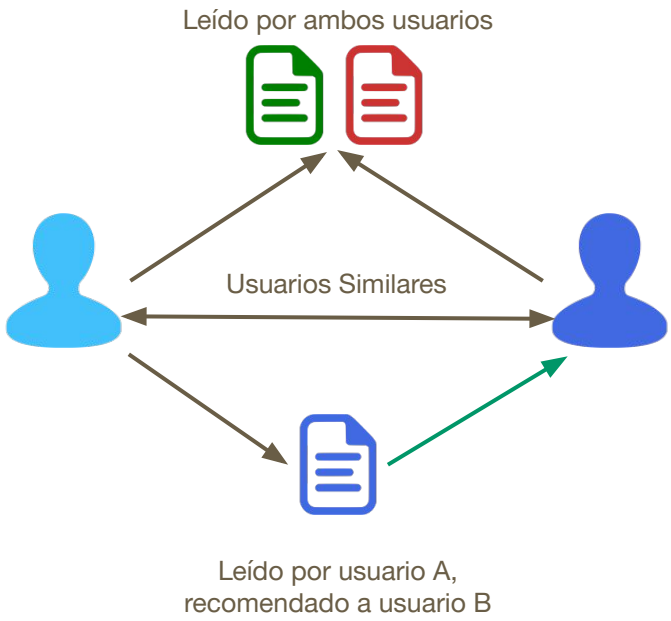


Tipos de Sistemas de Recomendación

Filtrado Basado en Contenido



Filtrado Colaborativo



Filtrado Basado en Contenido

- Recomendaciones basadas en el conocimiento que se tiene sobre los elementos que el usuario ha valorado (implícita o explícitamente), recomendando elementos similares.
- **More Like This**
- Elementos definidos en función de sus características:
  - Atributos de películas como género, año, director, actor, etc o contenido textual y características de artículos extraídos utilizando técnicas de NLP
- Perfil de usuario ⟶ Valoración de características
- Ejemplo: YouTube

Filtrado Basado en Contenido

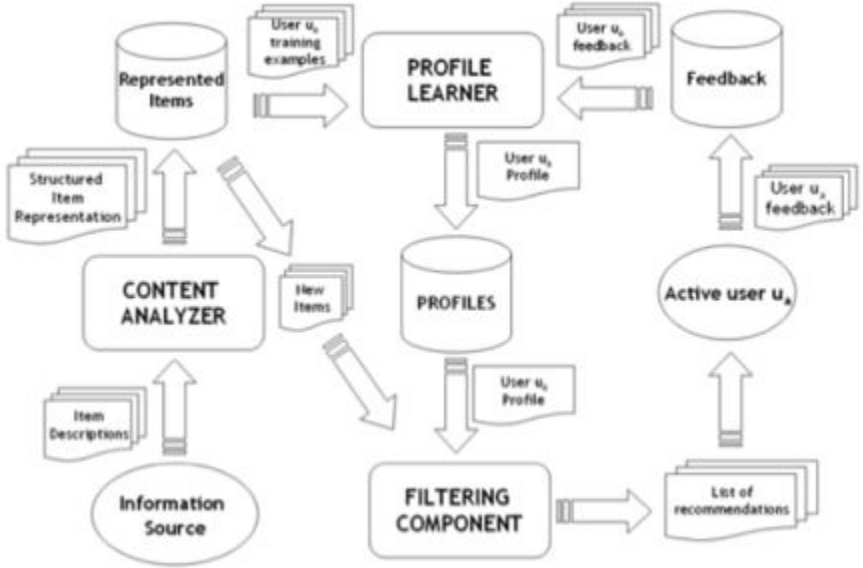


Fig. 3.1: High level architecture of a Content-based Recommender

Filtrado Basado en Contenido

Ventajas

- Elementos con suficientes características → Evita problema del ”nuevo elemento” (cold-start)
- Representaciones variadas en función de las técnicas usadas para extraer características
- Fácil hacer un sistema transparente

Desventajas

- Sobre-especialización (Overfitting)
- Filtro Burbuja
- Menos preciso que los métodos basados en Filtrado Colaborativo

Filtrado Colaborativo

- Utilizan preferencias históricas de usuarios sobre un conjunto de elementos
- Suposición principal: Usuarios que coincidieron en el pasado, tienden a coincidir en el futuro
- Idea: Muéstrame cosas que le han interesado a usuarios similares a mí
- Preferencias de usuario explícitas (valoración) o implícita (visita, click, ...)
- Matriz valoraciones usuario-elemento
- Basado en usuario (memoria) o elemento (modelo)

Filtrado Colaborativo

Ventajas

- Fácil de implementar
- Independiente del contexto
- Más precisos que los basados en contenido
- Permite recomendar contenidos difíciles de analizar
- Permite realizar recomendaciones válidas pero no esperadas, lo cual puede ser de utilidad

Desventajas

- Dispersión: % usuario-rating bajo
- Escalabilidad
  - Más vecinos = mejor clasificación
  - Más vecinos = mayor coste encontrar vecinos similares
- Arranque en frío (Cold-Start)
  - Nuevos usuarios con poca o nula información
  - Nuevos elementos con pocas valoraciones
- Problema de sinonímia
- Problema de subjetividad

Filtrado Colaborativo: Medidas Similitud

- Decidir cómo de similares son dos usuarios o elementos:
  - Similitud **Coseno** (Cosine)
  - Similitud **Correlación de Pearson** (Pearson)
  - Similitud **Coseno Ajustada** (Adjusted Cosine)
  - Similitud **Euclídea**
  - Similitud **Jaccard** (Jaccard Distance)

Filtrado Colaborativo: Cálculo de la Predicción

- Una vez se han calculado las similitudes, se obtiene la predicción que un usuario realizaría sobre los elementos que no ha valorado:
  - **Media Ponderada (Weighted Sum)**
  - **Regresión (Regression)**

Filtrado Colaborativo: Métodos de Evaluación

- Historial para evaluar las predicciones sobre las valoraciones
- Medidas de error
  - **Root Mean Square Error (RMSE)**
  - **Mean Absolute Error (MAE)**
- La tarea de recomendación suele verse como una tarea de recuperación de información
- Recuperar (recomendar) los elementos que han sido considerados como buenos por el sistema de recomendación
- **Precisión, Recall, F1 Score**
- Ejemplo TP: Buenas películas recomendadas
- Ejemplo FN: Buenas películas

Filtrado Colaborativo Basado en Usuarios (Memoria)

- Los algoritmos basados en usuario (memoria – User-Based) emplean las valoraciones que los usuarios dan a los elementos y métricas de similitud entre usuarios para determinar el conjunto de usuarios más similares a uno dado (usuario activo) y a continuación combinan las preferencias de esos vecinos para producir recomendaciones para dicho usuario (top-N)
- Asume que usuarios similares tendrán gustos similares
- Algoritmos:
  - K-Nearest Neighbourhood: algoritmo del vecino más cercano para determinar por similitud que elementos podrían interesar a un usuario, generando una valoración para un elemento analizando las valoraciones de los usuarios considerados vecinos

Filtrado Colaborativo Basado en Elementos (Modelo)

- Emplea algoritmos de ML y DM para predecir las valoraciones de un usuario, creando un modelo pre-procesado o aprendido
- Usan las semejanzas entre objetos para las recomendaciones: Valoración de un elemento basado en valoraciones de elementos similares
- Algoritmos:
  - K-Nearest Neighbourhood con similitud entre elementos
  - Factorización de matrices
  - SVD: Singular Value Decomposition
  - NMF: Non-Negative Matrix Factorization
  - Slope One
  - Co-Clustering
  - Deep Learning

Filtrado Colaborativo: Usuario VS Elemento

Usuario (Memoria)

- Es necesario usar toda la matriz de puntuaciones para encontrar los vecinos y luego realizar la recomendación
- En la práctica, más usuarios que objetos: métodos poco escalables
- Matrices de puntuación muy dispersas: pocas puntuaciones comunes entre usuarios

Elemento (Modelo)

- Se aprende el modelo (pre-procesado o aprendido)
- Únicamente realiza la recomendación en tiempo real
- Los modelos se actualizan periódicamente
- La construcción del modelo puede ser bastante costosa computacionalmente

Filtrado Demográfico

- Estas recomendaciones se realizan en función de las características de los usuarios
- Clasifica a los usuarios en grupos o categorías y hace recomendaciones de acuerdo con el grupo: ciudad, equipo, departamento, sexo, profesión, etcétera

Filtrado Híbrido

- Mezclan alguno de los tres filtrados mencionados anteriormente para realizar recomendaciones e incluso lo combinan con alguna otra técnica de inteligencia artificial como pueda ser Fuzzy Logic o la computación evolutiva.

Librería Surprise

- **Surprise** es una librería de Python para crear sistemas de recomendación basados en filtrado colaborativo
  - Arquitectura modular proporcionando un conjunto de algoritmos y datasets con los que comenzar a prototipar
- **Algoritmos**
  - KNN, Matrix Factorization, Slope One, CoClustering
- **Módulo de similitud**
  - Coseno, MSD, Pearson
- **Módulo de precisión/evaluación**: herramientas para calcular métricas de precisión sobre un conjunto de predicciones
  - RMSE, MAE, FCP

Surprise Notebook

Conclusiones

- Mayor notoriedad de los SR debido a sus numerosas aplicaciones
- Los usuarios no pueden gestionar toda la información disponible en internet
  - Necesidad de ayudar al usuario
  - También permite a las empresas sugerir contenidos que el usuario puede estar interesado en adquirir
- Posibles mejoras aplicables a los sistemas de recomendación:
  - Mejorar la seguridad respecto a valoraciones o usuarios falsos
  - Aprovechar las redes sociales
  - Explorar el grafo social y la interacción para determinar similitudes entre usuarios o elementos
  - Mejorar los métodos de evaluación cuando no hay valoraciones
  - Sistemas de recomendación proactivos

Referencias

- http://surpriselib.com/
- https://en.wikipedia.org/wiki/Recommender_system
- https://en.wikipedia.org/wiki/Collaborative_filtering
- https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26
- https://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf
- http://www.cs.carleton.edu/cs_comps/0607/recommend/recommender/memorybased.html
- http://www.cs.carleton.edu/cs_comps/0607/recommend/recommender/itembased.html

Q&A

GRACIAS!

PYCON ES
ALICANTE 2019

@adperezmorales
@rafa_haro

adperezmorales
rafa_haro