Sistemas de Recomendación

Recomendaciones no personalizadas, basadas en contenido y en conocimiento

Agenda

- **Recomendaciones no personalizadas**
- Recomendaciones basadas en contenido
- El problema de la búsqueda
- Vectores y predicción

Recomendaciones no personalizadas

- Se usan cuando no se sabe nada del usuario que va a “consumir” la recomendación
  - Cuando Guía Óleo dice “Este restaurante tiene 29/30 en servicio” es para todos los usuarios
  - Funciona para los usuarios a los que le gusta “lo popular”
- Aproximaciones:
  - Opiniones agregadas
  - Productos asociados

Opiniones agregadas (1)

- Agregaciones estadísticas de las opiniones
- Fáciles de calcular
- Se calculan una vez para todos los usuarios
- Ejemplos:
  - Ventas: “Los 40 principales” + “las pelis más vistas de la semana”
  - Opiniones explícitas: promedio de estrellas
  - Opiniones implícitas: lo más leído en diarios

Opiniones agregadas (2)

- Promedios: buenos para saber qué sienten los usuarios
- No tienen contexto
- Si podemos usar el contexto
  - Qué producto está mirando el usuario (contexto)
    → Recomendación de productos asociados

Productos asociados (1)

- A los usuarios que les gustó X también les gustó Y
- Porcentaje de usuarios que compraron X que tambien compraron Y: (X & Y) / X
- Problema:
  - X no es un producto popular
  - Y es un producto muy popular

Productos asociados (2)

- Ajustamos la fórmula con el incremento de X que está asociado a Y
- Miramos si X hace que Y sea más probable que si usara no X (!X)
  - [(X & Y) / X] / [(!X & Y) / !X]

Problemas

- Sesgo de selección
  - Ítems mediocres con ratings buenos
  - La gente que no le gustó deja de ir y no opina más
  - La gente que le encantó va y sigue opinando
- Diversidad de experiencias
  - Ítems excelentes con ratings mediocres
  - No todas las películas son para todos ya que se especializan en un segmento de los usuarios

Agenda

- Recomendaciones no personalizadas
- **Recomendaciones basadas en contenido**
- El problema de la búsqueda
- Vectores y predicción

Preferencias basadas en contenido

- Contenido: conjunto de atributos que describen los ítems (o los usuarios)
- Prefiero…
  - Películas de acción a románticas
  - Comida dulce a salada… pero si es pizza prefiero salada a dulce (con ananá)
  - Hoteles cerca del subte A en el centro

Recomendaciones basadas en contenido

- No dependen directamente de información de otros usuarios para hacer recomendaciones
- Usan la información que describe los ítems:
  - Noticias: Prefiero noticias de tecnología, política y economía
  - Ropa: Prefiero ropa casual, de algodón, azul.
  - Películas: Prefiero comedias o ciencia ficción, de los últimos 20 años
- No uso la popularidad de los ítems

Perfil del usuario

- Perfil del usuario
  - Unión de acciones/opiniones explícitas para inferir preferencias del usuario
- Creación
  - El usuario puede crearlo
  - Inferido de las acciones del usuario
  - Inferido de las opiniones explícitas del usuario

Recomendaciones basadas en casos

- Tenemos una base de datos de casos descriptos por una serie de atributos
- Se consulta usando un ejemplo o atributos para obtener casos relevantes
  - Me gustan las películas como “Matrix”
  - Me gustan las películas de acción y coches
- La consulta se puede armar en base a preguntas: ¿Con quién vas a ver la película? ¿Acción, romance o drama?

Recomendaciones basadas en conocimiento

- Es un sistema basado en casos con una interfaz inteligente
- Se pueden navegar los casos y ofrecerle al usuario opciones: “Quiero uno más barato, de cocina peruana…”
- Sirven para ofrecer ítems similares
- Idea: *bot* para recomendaciones basadas en conocimiento

Beneficios, desafíos y problemas

- Fáciles de explicar
- Dependen en atributos bien estructurados que se alinean con preferencias (problema: arte)
- Dependen en tener los atributos razonablemente distribuidos en los ítems (y al revés)
- No puede encontrar conexiones sorprendentes
- Fácil encontrar sustitutos de un ítem, no su complemento

Agenda

- Recomendaciones basadas en contenido
- El problema de la búsqueda
- Vectores y predicción

El problema de la búsqueda

- Tenemos el perfil del usuario como vector de palabras clave
- Queremos buscar ítems similares:
  - ¿Devolver todos los ítems que tienen los mismos valores en los atributos?
  - ¿Si un término ocurre mucho, es mejor?
- Tenemos que considerar:
  - La frecuencia de los términos tiene que ser significativa
  - No todos los términos son igualmente relevantes

Ponderación con TF-IDF

- TF-IDF = Frecuencia del término * Frecuencia inversa en los documentos
  - Frecuencia del término = # de veces que aparece un término en un documento
  - Frecuencia inversa en los documentos = # de documentos tienen un término. Usualmente log(#documentos/#documentos con el término)

¿Qué hace TF-IDF?

- Baja los términos muy usuales y *stopwords*
- Sube los términos centrales
- Falla:
  - Con términos muy raros
  - Las consultas son malas

¿Cómo se aplica TF-IDF a FBC?

- Se puede usar TF-IDF para crear un perfil de un documento o ítem
  - Una película puede ser descripta como un vector ponderado de sus etiquetas
- Y se combina con los ratings para crear un perfil del usuario y poder emparejarlo con futuros documentos

Agenda

- Recomendaciones basadas en contenido
- El problema de la búsqueda
- Vectores y predicción

Vector de palabras clave

- Cada palabra clave es una dimensión
- Las dimensiones definen un espacio
- Vectores:
  - Ítem
  - Perfil de preferencias de cada usuario
- La coincidencia entre las preferencias de los usuarios y los ítems se mide de acuerdo a que tanto se alinean los vectores

Representando ítems como vectores

- Simplemente 0/1 → no hay intensidad
- Cantidad de ocurrencias → hay intensidad
- TFIDF → hay intensidad pero además qué tanto me ayudan a distinguir

- Cuidado: “me gusta” != “importante” → aparece mucho en las cosas que me gustan, pero no es importante para mi

Creando perfiles

- Agrego los vectores de ítems por usuario
- ¿Todos los vectores tienen la misma importancia?
  - ¿Podemos usar el rating?
  - Sólo los de más de 3.5 estrellas
  - Rating negativo para los ratings bajos
- ¿Cómo actualizo los perfiles?
  - No lo hago / Ventana / Decaimiento

Calcular predicciones

- Coseno del ángulo entre los dos vectores (perfil, ítem)
- Producto escalar de vectores normalizados
- Varía de -1 a 1, cercano a 1 mejor

Fortalezas y desafíos de CBF

- Fortalezas:
  - Se basa por completo en contenido
  - El perfil creado es entendible
  - Se calcula fácil
  - Se puede integrar con buscadores
- Desafíos:
  - Saber los pesos de las palabras clave: más es más o repetición

Resumiendo

- Filtrado basado en contenido se basa en evaluar el perfil del contenido de cada ítem
- Los perfiles de usuarios se pueden crear agregando perfiles ponderados de los ítems de los que opinó
- Se evalúan los ítems no vistos haciendo coincidir el perfil del ítem contra el perfil del usuario (coseno vectorial)

Próxima clase (práctica)

- Aprender a crear perfiles de usuarios
- Mejorar las recomendaciones usando los datos de las películas
- No usar los datos de otros usuarios
  - Lo veremos más adelante

¿Preguntas?