Nagyméretű adathalmazok kezelése

Szomszédság alapú ajánló rendszerek

Készítette: Szabó Máté

- A rendelkezésre álló adatmennyiség növelésével egyre nehezebb kiválogatni a hasznos információkat
- Megoldás: ajánló rendszer
- Olyan szoftver ami megkísérli előre jelezni, hogyan fog egy felhasználó értékelni egy eddig még nem ismert terméket.

Az ajánló rendszerek céljai

- Több termék eladása
- Többféle termék eladása
- Felhasználó elégedettségének növelése
- A felhasználó igényeinek feltérképezése

Elvárások

- Stabilitás
- Hatékonyság
- Pontosság
- Novelty
- Serendipity

Felhasználói visszajelzések

- Implicit
  - A felhasználó normális tevékenységének rögzítése
- Explicit
  - Unáris
  - Bináris
  - Skaláris

A probléma formálisan

- *U* - felhasználó
- *I* - árucikk
- *R* - a rendszer által már ismert értékelések
- *S* - a lehetséges érékelések értékkészlete

- Ekkor a feladat:

  az $f: U \times I \rightarrow S$ függvény becslése

Validálás

- Mean Absolute Error

$$\frac{1}{|R|}\sum_{r_{ui}\in R}|f(u,i)-r_{ui}|$$

- Root Mean Squared Error

$$\sqrt{\frac{1}{|R|}\sum_{r_{ui}\in R}(f(u,i)-r_{ui})^2}$$

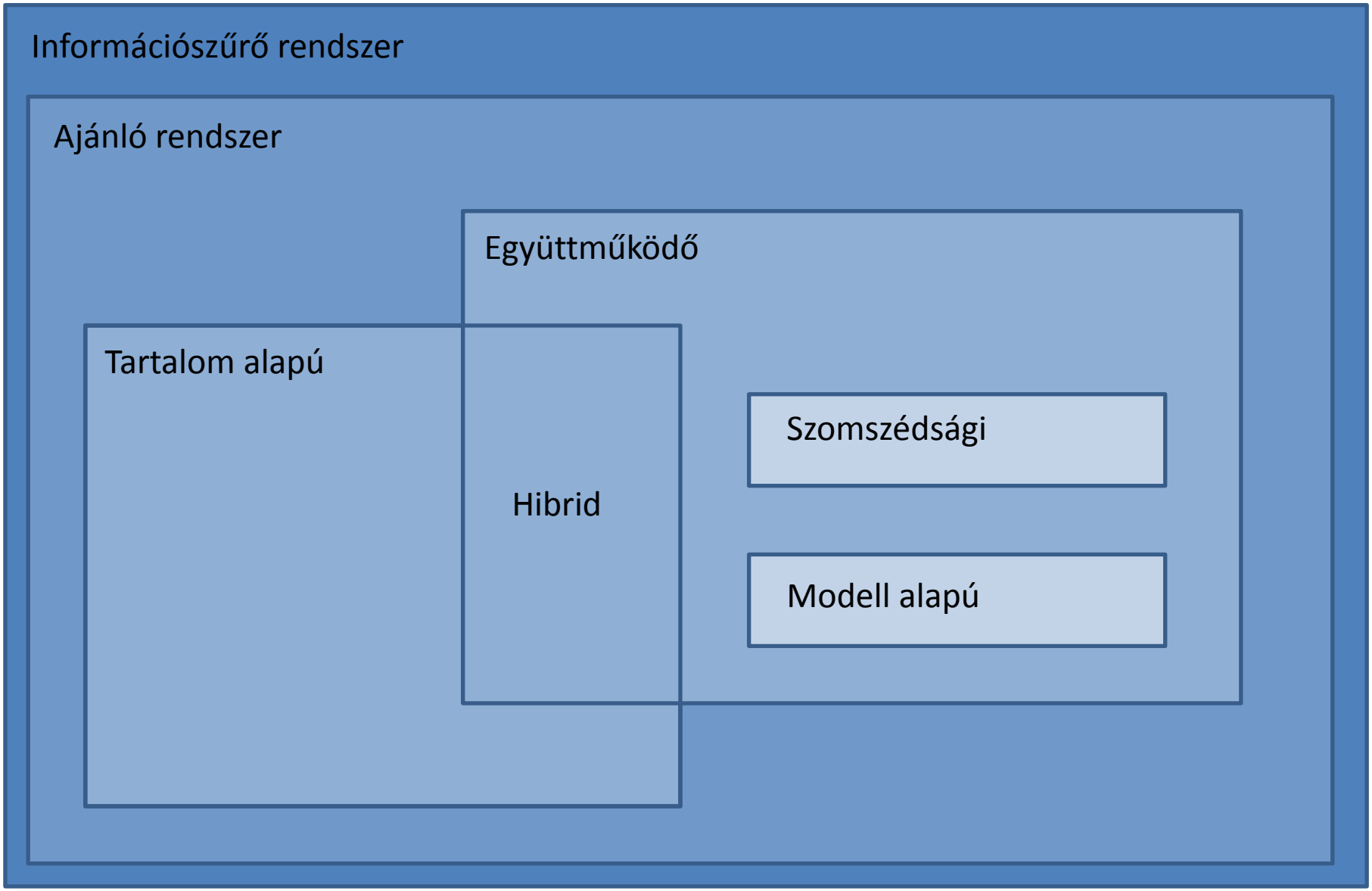
Validálás

- Precision

$$\frac{1}{|U|}\sum_{r_{ui}\in R}|L(u)\cap T(u)|/|L(u)|$$

- Recall

$$\frac{1}{|U|}\sum_{r_{ui}\in R}|L(u)\cap T(u)|/|T(u)|$$


Információszűrő rendszer
Ajánló rendszer
Együttműködő
Tartalom alapú
Hibrid
Szomszédsági
Modell alapú

Szomszédság alapú ajánló rendszer

- User alapú
- Item alapú

	a	b	c	d	e
A	2	1	4	3	2
B	5	3	1	1	4
C	1	3	3	4	2
D	5	2	3	?	4
E	4	2	2	1	3

User-based rating prediction

- Egyszerű

$$\hat{r}_{ui} = \frac{1}{|N_i(u)|} \sum_{v \in N_i(u)} r_{vi}$$

- Súlyozott

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i(u)} w_{uv} r_{vi}}{\sum_{v \in N_i(u)} |w_{uv}|}$$

- Súlyozott és normalizált

$$\hat{r}_{ui} = h^{-1}\left(\frac{\sum_{v \in N_i(u)} w_{uv} h(r_{vi})}{\sum_{v \in N_i(u)} |w_{uv}|}\right)$$

User-based classification

- Súlyozott

$$v_{ir} = \sum_{v \in N_i(u)} \delta(r_{vi} = r) w_{uv}$$

- Normalizált és súlyozott

$$\hat{r}_{ui} = h^{-1}\left(argmax \sum_{v \in N_i(u)} \delta(h(r_{vi}) = r) w_{uv} \right)$$

Különbség a két módszer között

- Ha az értékelések folytonosak, akkor inkább az előbbit, ha diszkrétek, akkor az utóbbit alkalmazzuk
- Gondoljunk arra az esetre ha egy terméket mindenki vagy 1 vagy 10 pontosra értékelt

Item-based rating prediction

- Súlyozott és normalizált

$$\hat{r}_{ui} = h^{-1}\left(\frac{\sum_{j \in N_u(i)} w_{ij} h(r_{uj})}{\sum_{j \in N_u(i)} |w_{ij}|}\right)$$

Item-based classification

- Súlyozott és normalizált

$$\hat{r}_{ui} = h^{-1}\left(argmax \sum_{j \in N_u(i)} \delta(h(r_{uj}) = r) w_{ij} \right)$$

- Az hogy item vagy user alapú megoldást választunk, alapvetően a felhasználók és az árucikkek számának arányától függ

	a	b	c	d	e
A	1	5	4	3	2
B	1	4	5	3	1
C	4	3	2	2	1

- Ne feledkezzünk meg a tár- és számításigényről sem

Szomszédság alapú ajánló rendszerek általános működése

- Normalizálás
- Hasonlósági súlyok kiszámítása
- Szomszédok kiválasztása

Normalizálás

- A felhasználók azonos skálán értékelnek, de nem azonos szempontok szerint
- Az egyes értékeléseket nem lehet összevetni
- Normalizálni kell

◯	a	b	c	d
A	5	5	5	4
B	3	2	2	4

Normalizálás

- Mean centering

$$h(r_{ui}) = r_{ui} - \bar{r}_u$$

◯	a	b	c	d
A	0,25	0,25	0,25	-0,75
B	0,25	-0,75	-0,75	1,25

- Z-score normalization

$$h(r_{ui}) = \frac{r_{ui} - \bar{r}_u}{\sigma_u}$$

◯	a	b	c	d
A	0,5	0,5	0,5	-1,5
B	0,26	-0,78	-0,78	1,31

Hasonlósági súlyok kiszámítása

- Koszinusz hasonlóság

$$\frac{\sum r_{ui} r_{vi}}{\sqrt{\sum r_{ui}^2 \sum r_{vi}^2}}$$

- Pearson korreláció

$$\frac{\sum (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum (r_{ui} - \bar{r}_u)^2 \sum (r_{vi} - \bar{r}_v)^2}}$$

- Még sok más

Hasonlósági súlyok kiszámítása

- Mean squared difference

$$\frac{I_{uv}}{\sum_{i \in I_{uv}} (r_{ui} - r_{vi})^2}$$

- Spearman rank korreláció

$$\frac{\sum (k_{ui} - \bar{k}_u)(k_{vi} - k_v)}{\sqrt{\sum (k_{ui} - \bar{k}_u)^2 \sum (k_{vi} - \bar{k}_v)^2}}$$

Szomszédok kiválasztása

- Nem gazdaságos minden értéket eltárolni, valamilyen előszűrést kell alkalmazni
  - Top-N filtering
  - Threshold filtering
  - Negative filtering
  - A fentiek keveréke

Szomszédok kiválasztása

k	MSD	SRC	PC
5	0.7898	0.7855	0.7829
10	0.7718	0.7636	0.7618
20	0.7634	0.7558	0.7545
60	0.7602	0.7529	0.7518
80	0.7605	0.7531	0.7523
100	0.7610	0.7533	0.7528

Christian Desrosiers, George Karypis - A comprehensive survey of neighborhood-based recommendation methods

A szomszédsági alapú rendszerek gyengéi

- A felhasználók nem feltétlenül adnak értékelést ugyanazokra a termékekre
- Gyakorlatban az értékelés eloszlása nem egyenletes
- Sok termékhez nincs, vagy nem elég az értékelés (pl. újonnan a rendszerhez adott tétel)

Lehetséges megoldások

- Dimenzió csökkentés
- Gráf alapú megoldások
  - Legrövidebb út
  - Véletlen bolyongás

A szomszédság alapú módszerek előnyei

- Egyszerűség
- Hatékonyság
- Igazolhatóság
- Stabilitás
- Véletlenszerű felfedezés

Köszönöm a figyelmet!