

## Zadání zápočtového projektu:

### Kolaborativní filtrování

#### Zadání

Cílem projektu je implementovat webovou aplikaci pro doporučování produktů přihlášenému uživateli na základě shody jeho preferencí s preferencemi ostatních uživatelů v databázi. Na vstupu bude databáze uživatelů, produktů a uživatelských hodnocení těchto produktů. Na výstupu se pak danému uživateli zobrazí seznam doporučených produktů z databáze.

Maximální počet členů v týmu: 2

#### Podrobnosti

- V případě kolaborativního filtrování [1] doporučíme uživateli produkt, který by se mu mohl líbit, pokud ho kladně ohodnotí jeden nebo několik jiných uživatelů, kteří jsou danému uživateli podobní z hlediska svých preferencí. Modelovat podobnost mezi dvěma uživateli můžeme např. s využitím Spearmanova koeficientu [2], Pearsonova koeficientu [3, 4] nebo kosinové míry podobnosti [4, 5]. Např. Spearmanův koeficient vrací hodnoty z intervalu  $[-1, 1]$ , kde 1 znamená maximální podobnost dvou uživatelů a -1 představuje zcela opačné zájmy.
- Ukázkový příklad výpočtu Spearmanova koeficientu je možné nalézt na Wikipedii [2]. V případě modelování podobnosti dvou uživatelů ho můžeme vypočítat následujícím způsobem.  $X_i$  a  $Y_i$  jsou hodnocení daných produktů uživateli X a Y (např. počty udělených hvězdiček). Hodnoty  $x_i$  a  $y_i$  udávají pořadí hodnocení produktů daného uživatele od nejlepšího k nejhoršímu.  $d_i$  je rozdíl pořadí hodnocených produktů.

$X_i$	$Y_i$	Pořadí $x_i$	Pořadí $y_i$	$d_i$	$d_i^2$
10	8	1	3	2	4
5	6	4	4	0	0
6	9	3	2	1	1
8	10	2	1	1	1
3	2	5	5	0	0

- Výsledný koeficient  $\rho$  pak vypočteme pomocí následujícího vzorce, kde  $N$  je počet hodnocení.

$$\rho = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} = 1 - \frac{6 \times 6}{5 \times 24} = 0.7$$

- Pokud vybereme  $n$  uživatelů podobných aktivnímu (přihlášenému) uživateli  $a$ , můžeme jeho hodnocení  $j$ -té položky  $p_{a,j}$  v databázi (tedy, jak moc se mu daná položka bude líbit) predikovat např. pomocí následujícího vzorce (viz přednáška č. 7, slide 13), kde  $w(a, i)$  odpovídá podobnosti mezi uživatelem  $a$  a uživatelem  $i$ . Podobnost může být vypočtena různými způsoby uvedenými výše (např. Spearmanovým koeficientem) a pouze na položkách, které oba dva uživatelé ohodnotili.  $v_{i,j}$  je hodnocení položky  $j$  uživatelem  $i$ .  $\bar{v}_a$  resp.  $\bar{v}_i$  je průměrné hodnocení položek uživatelem  $a$  resp.  $i$  (těch, které daný uživatel hodnotil). Konstanta  $\kappa$  je koeficient pro vyvážení obou sčítanců.

$$p_{a,j} = \bar{v}_a + \kappa \sum_{i=1}^n w(a, i)(v_{i,j} - \bar{v}_i)$$

## Implementace

Aplikace by měla obsahovat následující součásti:

- webové rozhraní, které bude přihlášenému uživateli umožňovat procházení databáze produktů (resp. libovolných objektů, které lze nějakým způsobem hodnotit) a jejich hodnocení (např. daným počtem hvězdiček) a které mu bude doporučovat dosud nehodnocené produkty z databáze na základě shody jeho preferencí s preferencemi ostatních uživatelů v databázi
- administrátorské rozhraní pro výpočet podobnosti mezi uživateli a předpočítání doporučených produktů pro všechny uživatele; v rozhraní bude možné konfigurovat parametry výpočtu např. maximální počet podobných uživatelů zahrnutý do výpočtu, nastavovat funkci pro výpočet podobnosti mezi dvojicemi uživatelů (Spearmanův koeficient, Pearsonův koeficient, kosinová podobnost, apod.), nastavovat různé konstanty použité při výpočtu podobnosti, aj. Funkce pro výpočet podobnosti uživatelů a predikci doporučených produktů je třeba implementovat samostatně (tj. bez použití knihoven).

## Data

- Data jsou vhodnější reálná, stažená z libovolného otevřeného zdroje, např. anonymizovaný seznam hodnocení filmů.

## Experimenty

- V rámci projektu lze porovnat různé funkce pro výpočet podobnosti mezi dvojicemi uživatelů z hlediska kvality doporučování. Dále lze měřit dobu potřebnou pro výpočet doporučených produktů všech uživatelů v databázi pro různé nastavení např. velikost databáze (tj. různý počet produktů/uživatelů/hodnocení), maximální počet uživatelů zahrnutých do výpočtu, aj.

## Reference

- [1] Wikipedia, Recommender system, Collaborative filtering, [https://en.wikipedia.org/wiki/Recommender\\_system#Collaborative\\_filtering](https://en.wikipedia.org/wiki/Recommender_system#Collaborative_filtering)
- [2] Wikipedia, Recommender system, Spearman's rank correlation coefficient, [https://en.wikipedia.org/wiki/Spearman%27s\\_rank\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient)
- [3] Wikipedia, Recommender system, Pearson correlation coefficient, [https://en.wikipedia.org/wiki/Pearson\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)
- [4] Tomáš Skopal, BI-VWM.21 - Vyhledávání na webu a v multimediálních databázích, Přednáška č. 7: Personalized search and social context.
- [5] Tomáš Skopal, BI-VWM.21 - Vyhledávání na webu a v multimediálních databázích, Přednáška č. 3: Vector model of information retrieval.