Sustavi za preopruke

Od oskudice do izobilja

- Polica je prostor ograničen resurs tradicionalnih trgovaca
  - Slično: TV reklame, kina, ...
- Web omogućuje gotovo besplatno širenje informacija o proizvodu
  - Nema problema sa ograničenošću prostora
  - Čak štoviše, informacije o proizvodu mogu biti personalizirane
- Što je više informacija, potrebno je više filtera
  - Sustavi za preporuke

Tipovi preporuka:

- Urednički i ručno odabrani
  - Lista favorita
  - Lista 'moram pogledati' stvari
- Jednostavne agregacije
  - Top 10
  - Najpopularniji
  - Nedavni
  - ...
- Personalizirane preporuke
  - Amazon, Ebay, YouTube, ...

FORMALNI MODEL

- X = skup klijenata
- S = skup itema
- Uslužna funkcija u: X x S → R
- R = skup rangova
- R je totalno uređen skup
- Npr. 0-5 zvjezdica, realni broj [0,1]

Glavni izazovi

- Prikupljanje poznatih rangova za matricu
  - Kako prikupiti početne podatke?
- Koristiti poznate ocjene kako bi procijenili / predvidjeti nepoznate
  - Fokus na visokim nepoznatim rangovima
  - Ideja je da se radije izvući informacije što pojedini korisnik voli nego što ne voli

- Ocijeniti metode predviđanja
    - Kako mjeriti točnost i performanse metoda za preporuku

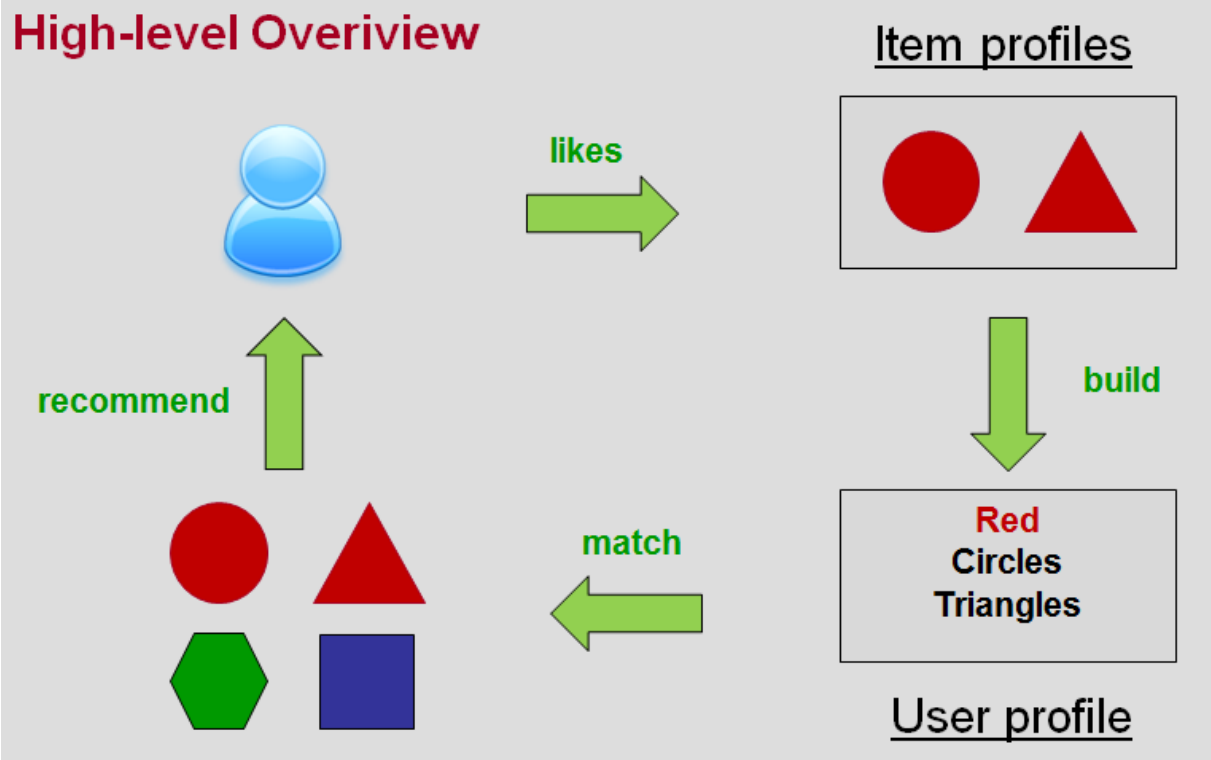
(1) Prikupljanje početnih podatake
- Eksplicitno
    - Pitati ljude da ocjene item
    - Nije efktivno u praksi, jer ljudi nemaju naviku dati povratnu informaciju eksplicitno
- Implicitno
    - Naučiti rangiranje iz korisničkih događaja
        - Npr. kupnja implicira visoko rangiranje
    - Kako se može implicitno detektirati nisko rangiranje?

(2) Koristiti postojeće rangove kako bi predvidjeli one koje nedostaju
- Ključni problem: Matrica korisnosti U je rijetka
    - Obični korisnik rangira samo ograničen podskup itema
- Problem teškog početka:
    - Novi korisnici nemaju povijest rangiranja
    - Novi itemi nemaju rangove
- Tri pristupa sustavima za preporuke
    - Bazirani na sadržaju
    - Suradnička otkrivanja
    - Bazirani na latentnom faktoru

Bazirani na sadržaju

Glavna ideja:

- Preporučiti iteme korisniku A slične prijašnjim itemima koje je korisnik A visoko rangirao
- Primerji: Isti glumci u filmovima, žanr, stranice sa istim sadržajem, ...



- Profili itema
  - Za svaki item, napravi profil
  - Profil je skup (vektor) značajki
    - Filmovi: autor, naziv, redatelj, ...
    - Tekst: skup ključnih riječi u dokumentu
  - Kako odabrati bitne značajke
    - Uobičajena heuristika iz rudarenja teksta je TD-IDF (uobičajenost pojma * inverz uobičajenosti dokumenta)
    - Pojam →značajka
    - Dokument →item

TD-IDF

- f_{ij} = frequency of term (feature) i in doc (item) j

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}}$$

- n_i = number of doc that contain term I
- N = total number of docs

$$IDF_i = log \frac{N}{n_i}$$

- **TF-IDF score** = $w_{ij} = TF_{ij} \times IDF_i$

- **Doc profile** = set of words with highest TF-IDF scores, together with their scores

Korisnički profili i prepostaka

- Mogućnosti korisničkih profila
    - Ponderirani prosjek rangiranih profila itema

Prediction heuristic

- Estimate $u(x, i)$ for user x and item i:

$$u(x, i) = \cos(x, i) = \frac{x \cdot i}{\|x\| \cdot \|i\|}$$

Prednosti sustava za preporuke baziranih na sadržaju:

- Nema potrebe za podacima na drugim korisnicima
- Može preporučiti korisnicima sa jedinstvenim ukusom
- Može preporučiti nove i nepopularne iteme
- Transparentnost
    - Može osigurati transparentno objašnjenje o značajkama sadržaja koje uzrokuju preporuku stavke

Nedostaci:

- Nalaženje adekvatnih značajki je teško i zahtjeva poznavanje specifične domene
- Ne može preporučiti novim korisnicima
- Previše specijalizirani
    - Preporuča samo iteme koji se podudaraju s profilom korisnika
    - Ljudi mogu imati više interesa
    - Ne može koristiti rangove drugih korisnika