Recommender System for a Dating Service

Diplomant: Lukáš Brožovský
Vedoucí práce: RNDr. Václav Petříček
Oponent: RNDr. Tomáš Skopal, Ph.D.

Studijní obor: Informatika, Softwarové systémy

Motivace

- “information overload”
  - příliš mnoho dat v určitém oboru
  - existují oblasti, kde není jasné, co hledáme
- typický příklad = (online) seznamka
  - nyní: složité dotazníky + často fixní požadavky
  - ... a nasledné hodiny při probírání nabídky
  - čas + cena!

Doporučovací systémy

- snaha o jiný přístup k vyhledávání
- není založeno na explicitním dotazu
- hledá “objekty” **zajímavé** pro daného uživatele
- pracuje s uživatelskými “profily”
- v praxi kombinace s klasickým vyhledáváním
  - předfiltrování dat

Profil uživatele

- data o “životě uživatele v systému”

- explicitní
  - hodnocení některých objektů v systému
  - jednotlivě / vzájemně
  - (dotazníky)
- implicitní
  - sledování chování uživatele
  - které objekty si prohlíží, jak dlouho apod.

Kolaborativní filtrování

- princip mnoha doporučovacích systémů
- hledá podobné profily
- kombinuje “knowhow” podobných uživatelů
- **hlavní myšlenka = dosud podobní uživatelé se budou chovat podobně i v budoucnu**
- příklad = nákupní portál (CD)

Konkrétní model

- vzor = webová aplikace LíbímSeTi
- 1...N uživatelů
  - typicky je uživatelů relativně hodně
- matice hodnocení R (N x N)
  - $r_{i,j} \in R$ ~ jak uživatel “i” hodnotil uživatele “j”
  - matice velmi řídká
- profil uživatele ≡ odpovídající řádek matice R

Konkrétní model - pokračování

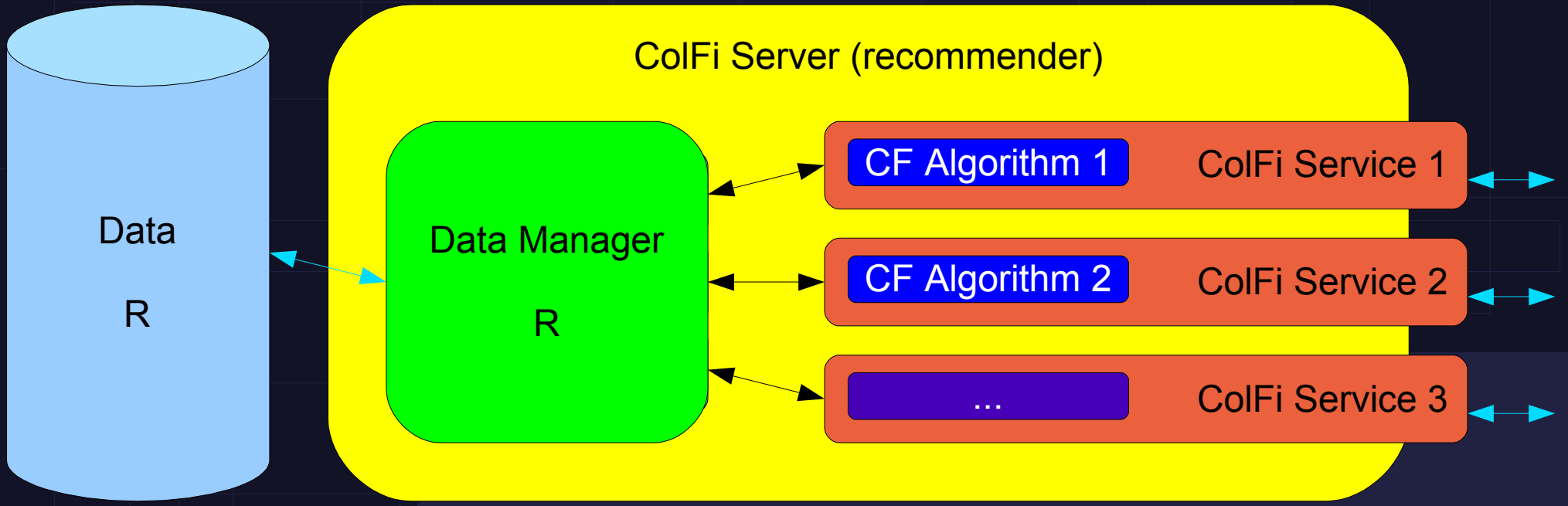
- “doporučení” = “top-K” uživatelů, které by uživatel hodnotil nejlépe
  - ze všech / jen z dosud nehodnocených
- omezíme se na algoritmy produkující atomické predikce
- obecný výpočet = po jednom dopredikujeme řádek matice a vybereme “top-K”

Algoritmy predikce

- 2 triviální pro srovnání
  - náhodný, průměr
- algoritmus (k)-nejbližších sousedů (kNN)
  - parametrizovatelný *(mCV, minN, maxN)*
- item-item varianta kNN
- obecný návrh - lze snadno rozšiřovat o další typy algoritmů
  - clusterovací, pravděpodobnostní modely...

ColFi Systém

- implementováno v Javě (Sun JVM, Java 5.0)
- stavebnicový model



Použitá data k testování

- 2 klasické datasety použité ve většině prací
  - MovieLens a Jester
  - transformace do tvaru čtvercové matice
  - pro srovnání
- 2 datasety přímo z online seznamek
  - ChceteMě a LíbímSeTi
  - reálná data
- v práci jsou datasety analyzovány a porovnávány

Testování

- 2 testy zaměřené především na absolutní přesnost/chybu predikce
  - AllButOne a GivenRandomX
- "simulační" test
  - simulace produkčního prostředí
  - testuje i výkon (vč. multi-uživatelského přístupu)
- "empirický" test
  - snaha ověřit kvalitu výsledného "doporučení"

Výsledky testů

- podobné výsledky na známých datasetech
  - chyba kolem 18-20% (NMAE)
- na datech seznamky dokonce mírně lepší!
  - chyba 14-15% (NMAE)
  - překvapivý úspěch triviálního algoritmu
    - univerzální preference, fotky

Závěr

- **slibný a v seznamkách dosud nevyužitý přístup**
- **robustní snadno rozšiřitelný open-source doporučovací systém**
- nové obsáhlé veřejné datasety
- stavební kámen pro další práci a výzkum
  - distribuované řešení, hybridní algoritmy, vzájemné párování, (ne)úmyslné útoky...