Proximity

Tomáš Gregorovič

Proximity

Proximity

http://kdl.cs.umass.edu/proximity

- systém pro dobývání znalostí z relačních dat
- open-source, Java
- vyvíjí laboratoř pro dobývání znalostí na University of Amherst, Massachusetts

Proximity – správa dat

- správa dat – databázový server MonetDB

  http://www.monetdb.nl/

  - open-source, multiplatformní
  - vertikální databázový model

Proximity – reprezentace dat

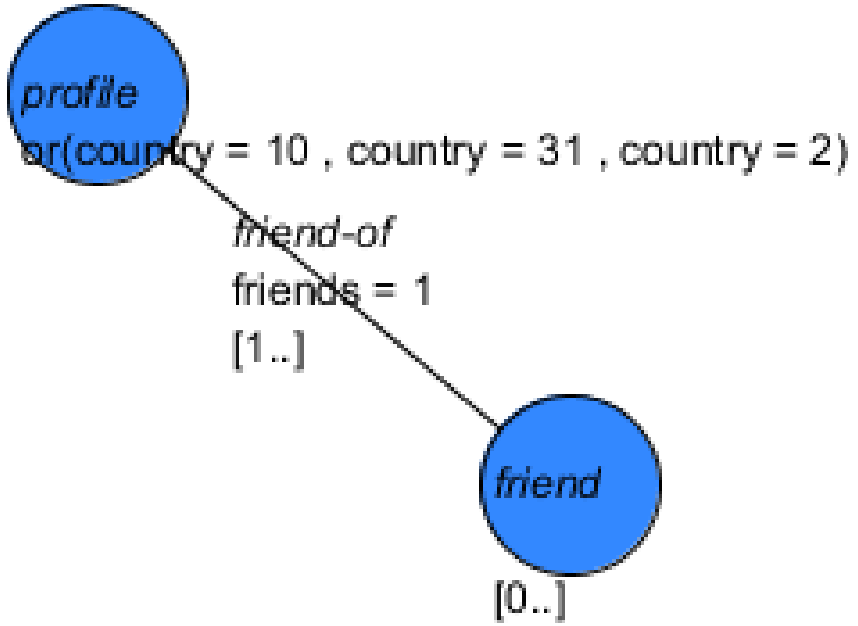
- relační databáze v podobě sítě
- objekty (Objects)
- vazby (Links) – orientované spojení mezi dvěma objekty
- kontejnery (Containers) – kolekce podgrafů sítě

Proximity - atributy

- pro objekty, vazby i kontejnery
- umožňují reprezentovat různé typy objektů a odkazů v rámci jedné sítě
- typy – celé a desetinné číslo, řetězec, datum
- vícehodnotové – pro zaznamenání relačních atributů typu 1:N

Proximity - QGraph

- grafický dotazovací jazyk, pro přípravu dat pro dolování
- označkovaný graf – uzly odpovídají objektům, hrany vazbám
- omezující podmínky, anotace
- výsledkem kontejner podgrafů splňujících dotaz



Proximity – modely pro dobývání znalostí

- pravděpodobnostní statistické modely
- skripty v Pythonu, Java API Proximity
- relační Bayesovský klasifikátor
- relační závislostní sítě
- relační pravděpodobnostní stromy RPT

Proximity – RPT

- pravděpodobnostní rozhodovací strom, klasifikace atributu
- nastavení:
  - soubor instancí pro učení – kontejner podgrafů
  - soubor instancí pro testování
  - třídní atribut centrálního objektu podgrafu
  - zvolení uvažovaných atributů pro učení
  - maximální hloubka, min. počet podgrafů v uzlu

Proximity – RPT učení I.

- propozicionalizace relačních dat

  => vytváření atributových tabulek pomocí agregátových funkcí:

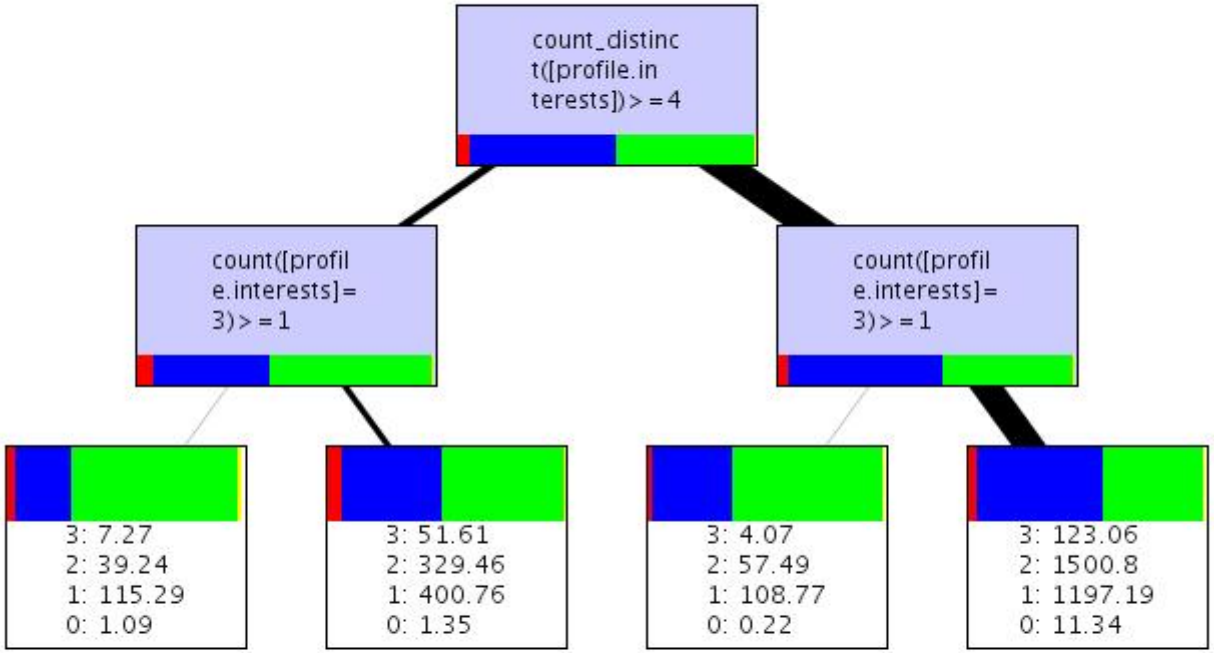
  - průměr, počet, stupeň, součet
  - minimum, maximum, majorita, proporce

Proximity – RPT učení II.

- výběr atributu a binárního dělení uzlu stromu
- podle statistické metriky chi-squared
- pro zjištění významnosti dělení – p-hodnota

Proximity – RPT výsledek

- rozhodovací strom



- celková úspěšnost (accuracy)

Proximity – RPT výpočetní náročnost

- podle experimentálního pozorování
- lineární vzhledem k počtu podgrafů
- lineární vzhledem k počtu vytvořených atributových tabulek
- obvykle pro počet tabulek podle typu atributu platí:

  obyčejný << vícehodnotový << atribut necentrálního objektu

Proximity – RPT nevýhody

- pouze binární větvení – růst hloubky, roste výpočetní náročnost, klesá přehlednost, těžší hledání zajímavých cest
- chybí automatické prořezávání výsledného stromu