Mapování PMML a BKEF dokumentů

Obsah prezentace

- Úvod do řešené problematiky
- Příbuzné problémy
- Návrh vlastního řešení
- Konkrétní implementace

Úvod do problematiky

- SEWEBAR
  - Výsledky dataminingu – PMML
    - LISp-Miner
  - Znalostí expertů – BKEF

- Potřeba vytvořit mapovací soubor FML
  - PMML ⇔FML⇔ BKEF
  - PMML ⇔FML⇔ PMML

Nároky na řešení

- Cílem je poloautomatická podpora nabízející uživateli vhodná mapování
- Integrace do CMS Joomla! 1.5

Příbuzné problémy

- Mapování schémat relačních databází
- Mapování ontologií

Přístup k datům v PMML/BKEF

- Zpracování: dvojice tabulek z relačních DB
  - Převod použitelných informací do uni-formátu
  - (Meta)atributy představují sloupce

Úroveň mapování

- Dvoustuňový proces
  - Mapování „sloupců“
  - Mapování hodnot

Techniky pro určení podobnosti

- Podobnost sloupců
  - Podobnost názvů
    - similar_text()
  - Podobnost hodnot
    - kategoriální sloupce
      - n-gramy
    - numerické sloupce
      - zahrnutí hodnot do intervalu
      - podobnost intervalů
  - Předchozí zkušenosti

Podobnost sloupců – předchozí zkušenosti

- Ukládání úspěšných namapování po každém dokončení mapování
  - identifikace prostřednictvím názvů sloupců
  - rozdílné ohodnocení uživatelem potvrzených/namapovaných sloupců a mapování jen tolerovaných

Podobnost hodnot

- Neřešíme pro numerické sloupce
- Moduly
  - Totožné hodnoty
  - Nejpodobnější (similar_text)
  - Nejpodobnější (n-gramy)

Režim mapování

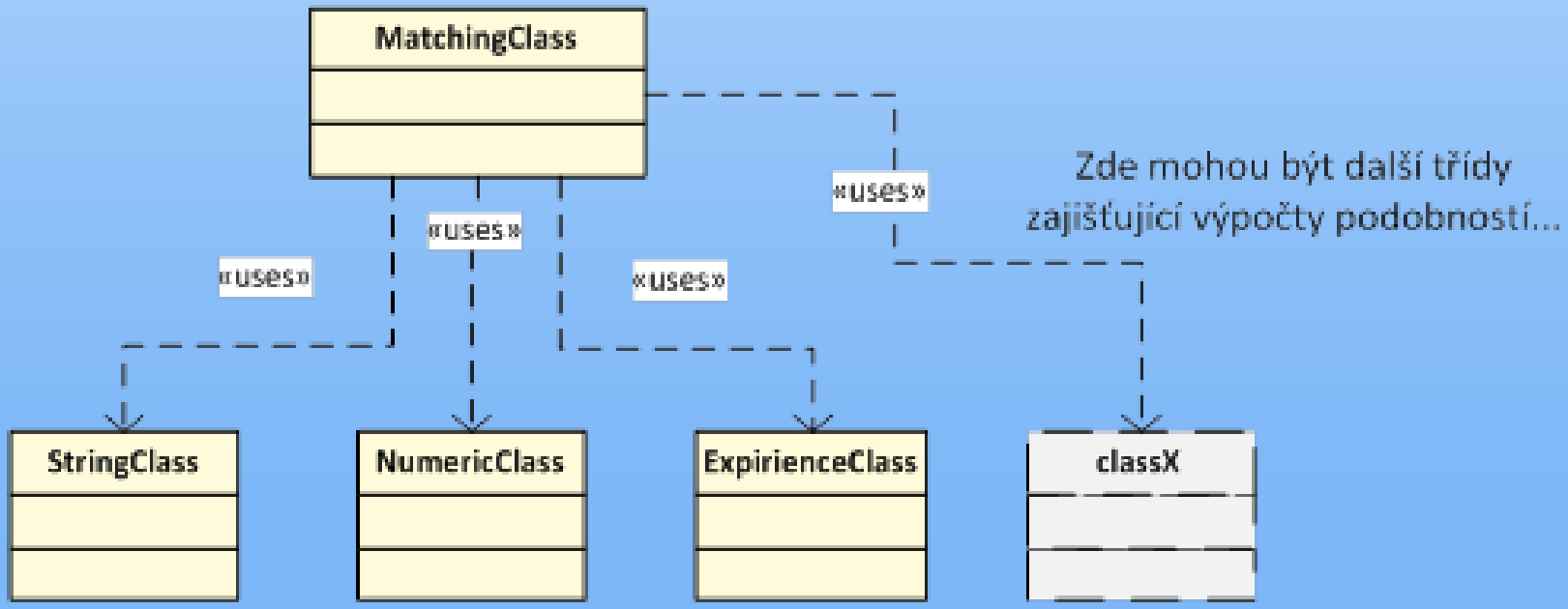
- Mapování sloupců: 1:1, N:1
- Mapování hodnot: M:N

Algoritmy pro výběr nejlepší kombinace sloupců

- Automatický návrh
  - 1:1
    - Globální maximum
    - Vlastní heuristický algoritmus
  - N:1
    - Nejpodobnější položky
- Manuální mapování

Implementace

- Řešení v jazyce PHP (objektově)
- Modulární řešení



Implementace – CMS Joomla!

- Kompatibilita s verzí 1.5
  - 1.6 má jinou strukturu organizace článků
  - Jinak funkční i ve verzi 1.6

Hodnocení úspěšnosti mapování

- Přesnost

$$p = \frac{\text{správně namapované sloupce}}{\text{všechny namapované sloupce}}$$

- Úplnost

$$r = \frac{\text{správně namapované sloupce}}{\text{všechny namapovatelné sloupce}}$$

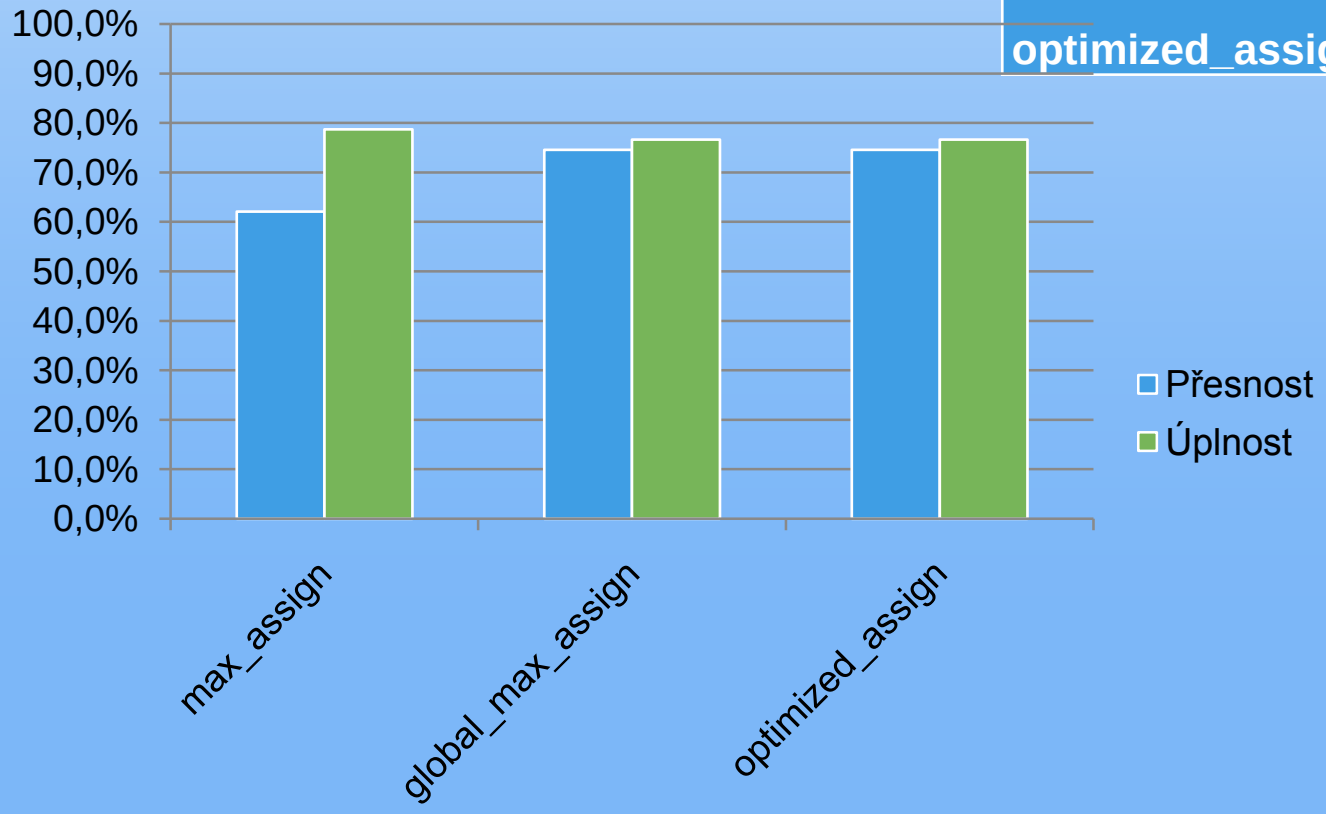
Testovací data

- Illinois Semantic Integration Archive
  - http://pages.cs.wisc.edu/~anhai/wisc-si-archive/
  - Data o kurzech vyučovaných na univerzitách

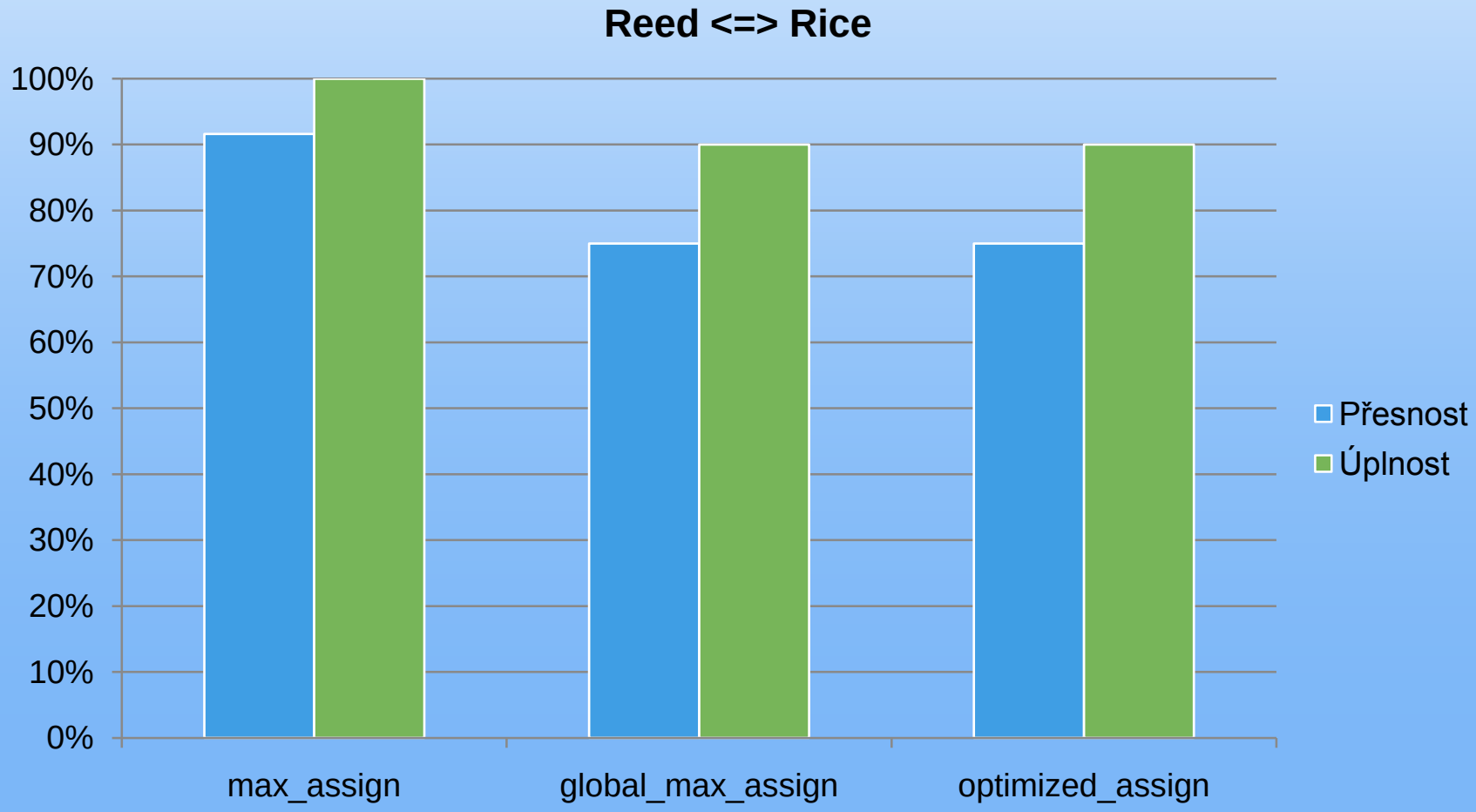
Univerzita	Počet řádků
Reed	703
Rice	1409
UWN	1642
Washington	3904
WSU	5493

Hodnocení úspěšnosti mapování

	Přesnost	Úplnost
max_assign	62,1%	78,7%
global_max_assign	74,6%	76,6%
optimized_assign	74,6%	76,6%



Hodnocení úspěšnosti mapování – s modulem předchozích zkušeností



Možnosti dalšího vývoje

- Ošetření ignorovaných sloupců
- Optimalizace formátu FML
- Integrace s BKEF editorem
- Integrace s aplikací LISp-Miner