Automatizovaný návrh pravidel pro integraci dat

Martin Řimnáč, Zdeňka Linková

Ústav informatiky AV ČR, v.v.i.

ITAT 2007
21.- 27.09. 2007

Formalismus

- Popis objektů universa - elementy $e \in \mathscr{E}$
- Element $\mathscr{E} = \mathscr{A} \times \mathscr{D}$
- Záznam $t \in \mathscr{T}$ definován $t \subseteq \mathscr{E}; \forall A \in \mathscr{A} |(A, \star) \in t| \leq 1$
- Data uložena pomocí vztahů
- Dvě úrovně:
  1. Instance - $i \in \mathscr{I} \subseteq \mathscr{E} \times \mathscr{E}$ - implikace mezi elementy
  2. Funkční závislosti -$f \in \mathscr{F} \subseteq \mathscr{A} \times \mathscr{A}$- zobecněné implikace na atributové úrovni

Formalismus

- Formalismus binárních matic
  1. Matice úložiště

$$\Phi = [\phi_{ij}], \quad \phi_{ij} = \begin{cases} 1 & \text{pokud } e_i \rightarrow e_j \in \mathscr{I} \\ 0 & \text{jinak} \end{cases}$$

  2. Matice funkčních závislostí

$$\Omega = [\omega_{ij}], \quad \omega_{ij} = \begin{cases} 1 & \text{pokud } A_i \rightarrow A_j \in \mathscr{F} \\ 0 & \text{jinak} \end{cases}$$

- Vztah (transformace)

$$\Omega = \Delta^T \Phi \Delta \qquad \Phi' = \Phi \odot \Delta \Omega \Delta^T$$

pomocí matice aktivních domén atributů

$$\Delta = [\delta_{ij}], \quad \delta_{ij} = \begin{cases} 1 & \text{pokud } e_i = (A_j, v_\star) \in \mathscr{E} \\ 0 & \text{jinak} \end{cases}$$

Příklad - Zdroj 1

$$\Phi_1 = \left[\begin{array}{ccc|cc|cc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{array}\right] \begin{array}{l} \text{mesto, Praha} \\ \text{mesto, Brno} \\ \text{mesto, Vídeň} \\ \hline \text{zeme, ČR} \\ \text{zeme, Rakousko} \\ \hline \text{mena, CZK} \\ \text{mena, EUR} \end{array}$$

$$\Omega_1 = \left[\begin{array}{c|c|c} 1 & 0 & 0 \\ \hline 1 & 1 & 0 \\ \hline 1 & 1 & 1 \end{array}\right] \begin{array}{c} \text{mesto} \\ \hline \text{zeme} \\ \hline \text{mena} \end{array}$$

$$\vec{y} = \Phi\vec{x}$$

Příklad - Zdroj 2

$$\Phi_2 = \left[\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}\right] \qquad \begin{array}{l} z_2\text{:hlavni\_mesto} = \text{Praha} \\ z_2\text{:hlavni\_mesto} = \text{Vídeň} \\ z_2\text{:stat} = \text{ČSFR} \\ z_2\text{:stat} = \text{Rakousko} \end{array}$$

$$\Omega_2 = \left[\begin{array}{c|c} 1 & 1 \\ \hline 1 & 1 \end{array}\right] \begin{array}{c} \text{hlavni\_mesto} \\ \hline \text{stat} \end{array}$$

Integrace dat - formalizace

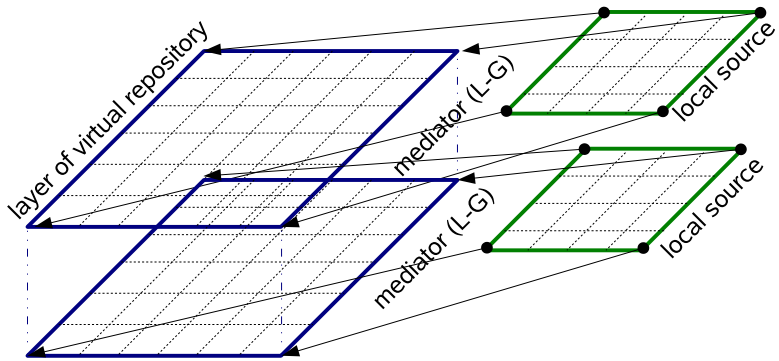
- každý element (lokálního) zdroje $S_l \in \mathscr{S}$ je mapován na globální element:

$$\Gamma_{S_l} = [\gamma^l_{ij}];\ \gamma^l_{ij} = \begin{cases} 1 & \text{když } e_i \sim e_j,\ \ e_i \in \bigcup_{\forall S \in \mathscr{S}} \mathscr{E}_S, e_j \in \mathscr{E}_{S_l} \\ 0 & \text{jinak} \end{cases}$$

- Řešení:
  1. Centralizovaně
  2. Decentralizovaně

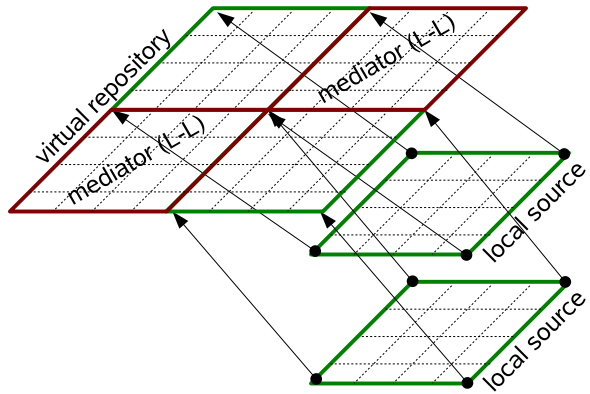
Virtuální globální matice úložiště

Centralizovaně



$$\Phi_{\mathscr{S}} = \sum_{\forall S_l \in \mathscr{S}} \Gamma_{S_l} \Phi_{S_l} \Gamma_{S_l}^T$$

Decentralizovaně



$$\Phi_{\mathscr{S}} = \begin{bmatrix} \Phi_1 & \Psi_{12} & \cdots & \Psi_{1|\mathscr{S}|} \\ \Psi_{21} & \Phi_2 & \cdots & \Psi_{2|\mathscr{S}|} \\ \vdots & & \ddots & \\ \Psi_{|\mathscr{S}|1} & \cdots & \cdots & \Phi_{|\mathscr{S}|} \end{bmatrix}$$

$$\Psi_{ij} = \Gamma_{S_i}^T \Gamma_{S_j}$$

Mapování - úroveň atributů - ekvivalence

1. Binární:

$$\Pi_{kl} = \Delta_{S_l}^T \Psi_{kl} \Delta_{S_k} \qquad \Psi'_{kl} = \Psi_{kl} \odot \Delta_{S_l} \Pi_{kl} \Delta_{S_k}^T$$

2. Vážené (semiautomatický návrh):
   - Míra - symetrická - překryv domén

$$\Pi_{kl} = \Pi_{lk}^T = [\pi_{ij}],\ \pi_{ij} = \frac{|\mathscr{D}_\alpha^{S_k}(A_i) \cap \mathscr{D}_\alpha^{S_l}(A_j)|}{|\mathscr{D}_\alpha^{S_k}(A_i) \cup \mathscr{D}_\alpha^{S_l}(A_j)|}$$

   - semiautomatický návrh:
     preference pravidel s maximální podporou
   - jednoznačné přiřazení na podmínky disjunktních "globálních" domén

Mapování - úroveň atributů - hierarchický vztah

- Míra - nesymetrická - podřazenost domén:

$$\Pi_{kl}^{\sqsubset} = [\pi_{ij}],\ \pi_{ij} = \frac{|\mathscr{D}_{\alpha}^{S_k}(A_i) \cap \mathscr{D}_{\alpha}^{S_l}(A_j)|}{|\mathscr{D}_{\alpha}^{S_k}(A_i)|}$$

- semiautomatický návrh:
  preference pravidel s maximální podporou
- nutná restrikce:
  nepřípustná dvojice mapování $A_q^{S_l} \sqsubset A_p^{S_k} \sqsubset A_o^{S_l}$
  neboť vede na neexistující "instance" $A_o^{S_l} \rightarrow A_q^{S_l}$

Příklad výpočtu ohodnocení pravidel

1. Pro hierarchická pravidla dostáváme:

$$\Pi_{12}^{\sqsubset} = \begin{bmatrix} \frac{2}{3} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \quad \Pi_{21}^{\sqsubset} = \begin{bmatrix} \frac{2}{2} & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \end{bmatrix}$$

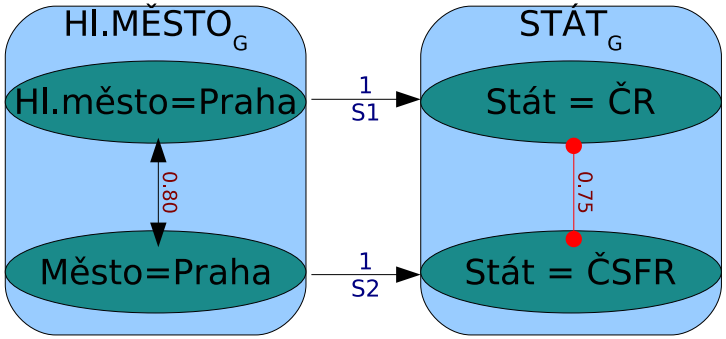
2. Podporu pro ekvivalenci:

$$\Pi_{12} = \Pi_{21}^{T} = \begin{bmatrix} \frac{2}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \end{bmatrix}$$

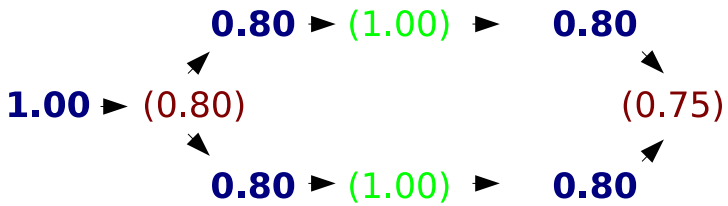
3. Seřadíme-li pravidla podle priority, dostáváme:

$z_2 : \text{hlavni\_mesto} \sqsubset z_1 : \text{mesto}$	1.00	$\oplus$
$z_2 : \text{hlavni\_mesto} \sim z_1 : \text{mesto}$	0.66	$\ominus$
$z_1 : \text{mesto} \sqsubset z_2 : \text{hlavni\_mesto}$	0.66	$\ominus$
$z_2 : \text{stat} \sqsubset z_1 : \text{zeme}$	0.50	$\oplus$
$z_1 : \text{zeme} \sqsubset z_2 : \text{stat}$	0.50	$\oplus$
$z_2 : \text{stat} \sim z_1 : \text{zeme}$	0.33	$\ominus$

Příklad - Nekonzistence a vážení pravidel



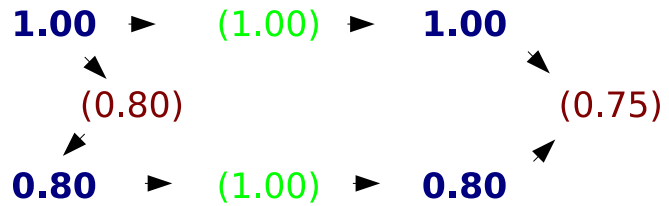
Centralizované



Nekonzistence:

$$\nu = y_j y_{j'} \psi_{jj'} = 0.80 \cdot 0.80 \cdot 0.75 = 0.48$$

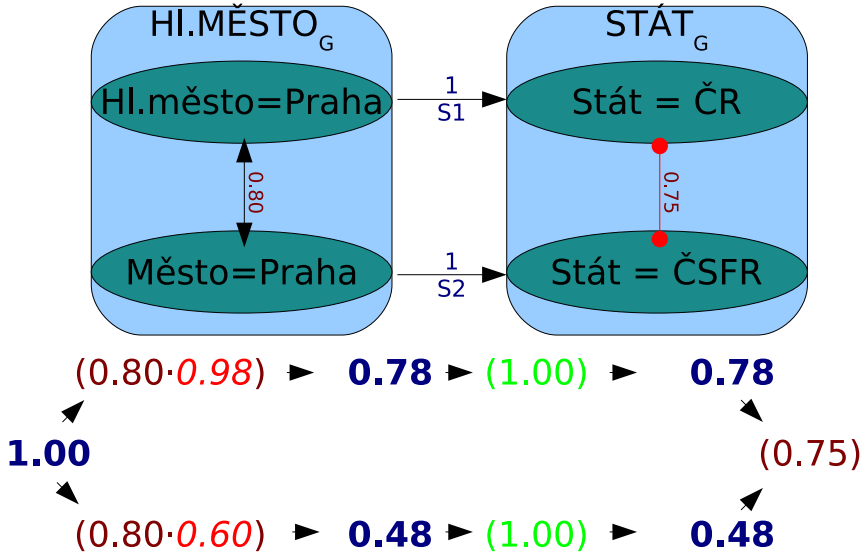
Decentralizované - výběr zdroje



Nekonzistence:

$$\nu = y_j y_{j'} \psi_{jj'} = 1.00 \cdot 0.80 \cdot 0.75 = 0.60$$

Další míry - Důvěryhodnost zdrojů



$$\nu = y_j y_{j'} \psi_{jj'} = 0.78 \cdot 0.48 \cdot 0.75 = 0.28$$

Přínos

- Přístup k integraci dat
  - centralizovaný (klasický)
  - decentralizovaný (webové zdroje dat)
- Umožnění odhadu integračních pravidel
  - tam, kde není možný lidský faktor
  - výběr nej... kandidátů na základě nepřímých měr
    - symetrická - ekvivalence
    - nesymetrická - hierarchie
  - Míra - ochrana lokálního zdroje před okolím