

Sémantický web: Vize globálního úložiště dat?

Martin Řimnáč, Zdeňka Linková, Roman Špánek

Ústav informatiky AV ČR, v.v.i.

DATAKON 2007
Brno 20.-23.10. 2007

- 1 Úvod, motivace a formalismus
 - Motivace
 - Formalismus
 - Fuzzyfikace
- 2 Integrace dat a reputace zdrojů
 - Integrace dat
 - Virtuální matice úložiště
 - Nekonzistence a vážení pravidel
 - Reputace zdrojů
 - Centralizované a decentralizované řešení
- 3 Závěr

Motivace

Sémantický web

- Data - vztahy mezi jednoznačně definovanými entitami
- Předmětem výzkumu posledních let
 - Deskripční logiky, Protege, Pellet, slovníky...
- Co však přinesl laickému uživateli?

Motivace

Sémantický web

- Data - vztahy mezi jednoznačně definovanými entitami
- Předmětem výzkumu posledních let
 - Deskripční logiky, Protege, Pellet, slovníky...
- Co však přinesl laickému uživateli?
 - Nové metody pro vyhledávání dat
 - Aktivní datové zdroje (obsah) - Web X.0

Motivace

Sémantický web

- Data - vztahy mezi jednoznačně definovanými entitami
- Předmětem výzkumu posledních let

Deskripční logiky, Protege, Pellet, slovníky...

- Co však přinesl laickému uživateli?
 - Nové metody pro vyhledávání dat
 - Aktivní datové zdroje (obsah) - Web X.0

Cíl:

- Představit vizi (webového) distribuovaného prostředí datových zdrojů
- Pojmenovat problémy a nástin řešení

Motivace

*Mějme datové zdroje S_1 , S_2 a S_3 poskytující data o cestování.
Dotazujme se nyní, které země je Praha hlavním městem.*

- zdroj S_1 odpoví Česká Republika
- zdroj S_2 odpoví Slovenská Republika
- zdroj S_3 odpoví Česká Republika

Jaká odpověď je správná?

Motivace

*Mějme datové zdroje S_1 , S_2 a S_3 poskytující data o cestování.
Dotazujme se nyní, které země je Praha hlavním městem.*

- zdroj S_1 odpoví Česká Republika
- zdroj S_2 odpoví Slovenská Republika
- zdroj S_3 odpoví Česká Republika

Jaká odpověď je správná?

Rozhodnutí není jednoduché...

- nejčtenější

Motivace

*Mějme datové zdroje S_1 , S_2 a S_3 poskytující data o cestování.
Dotazujme se nyní, které země je Praha hlavním městem.*

- zdroj S_1 odpoví Česká Republika
- zdroj S_2 odpoví Slovenská Republika
- zdroj S_3 odpoví Česká Republika

Jaká odpověď je správná?

Rozhodnutí není jednoduché...

- nejčtenější versus kompletní (preferenčně uspořádaná) odpověď

Motivace

*Mějme datové zdroje S_1 , S_2 a S_3 poskytující data o cestování.
Dotazujme se nyní, které země je Praha hlavním městem.*

- zdroj S_1 odpoví Česká Republika
- zdroj S_2 odpoví Slovenská Republika
- zdroj S_3 odpoví Česká Republika

Jaká odpověď je správná?

Rozhodnutí není jednoduché...

- nejčtenější versus kompletní (preferenčně uspořádaná) odpověď
- Inspirace sociálním chováním - zahrnutí uživatelských zkušeností s danými zdroji

Motivace

Předpoklad znalosti:

① Schémata zdrojů

integritní omezení, obzvláště funkční závislosti

② Integračních pravidel

Motivace

Předpoklad znalosti:

① Schémat zdrojů

integritní omezení, obzvláště funkční závislosti

② Integračních pravidel

Ty však jsou mnohdy nedostupné... Proto

① Odhad struktury dat - uložení dat

(funkční závislosti, aktivní domény atributů)

② Odhad integračních pravidel - integrace dat

(semiautomatické metody - kosinové míry)

③ Odhad důvěryhodnosti zdrojů

(reputační systémy, dynamické chování)

Formalismus

- Popis objektů universa - elementy $e \in \mathcal{E} \subseteq \mathcal{A} \times \mathcal{D}$
- Záznam $t \in \mathcal{T}$ definován $t \subseteq \mathcal{E}; \forall A \in \mathcal{A} | (A, \star) \in t | \leq 1$
- Koexistence dvou elementů $e_i, e_j \in t$ v záznamu může ukazovat na vztah mezi nimi
- Dvě úrovně:
 - 1 Instance - $i \in \mathcal{I} \subseteq \mathcal{E} \times \mathcal{E}$ - implikace mezi elementy
 - 2 Funkční závislosti - $f \in \mathcal{F} \subseteq \mathcal{A} \times \mathcal{A}$ - zobecněné implikace na atributové úrovni

Formalismus

- Formalismus binárních matic

- Matice úložiště

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

- Matice funkčních závislostí

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

- Vztah (transformace)

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

pomocí matice aktivních domén atributů

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

Příklad

$$\Phi = \left[\begin{array}{ccc|cc|cc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{array} \right] \begin{array}{l} \text{Town, Prague} \\ \text{Town, Brno} \\ \text{Town, Wien} \\ \hline \text{State, Czech Republic} \\ \text{State, Austria} \\ \hline \text{Currency, CZK} \\ \text{Currency, EUR} \end{array}$$

$$\Omega = \left[\begin{array}{c|c|c} 1 & 0 & 0 \\ \hline 1 & 1 & 0 \\ \hline 1 & 1 & 1 \end{array} \right] \begin{array}{l} \text{Town} \\ \hline \text{State} \\ \hline \text{Currency} \end{array} \quad \vec{y} = \Phi \vec{x}$$

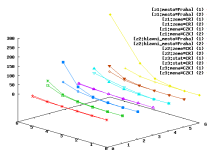
Fuzzyfikace

- Někdy není dobré něco tvrdit s naprostou jistotou. Proto zavádíme:

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

Váhy příspěvků

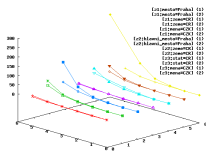
$$y_i = \sum_{\forall j} \phi_{ij} x_j$$



- jak moc je velká podpora
- mnohdy nestabilní

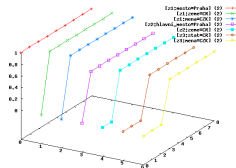
Váhy příspěvků

$$y_i = \sum_{\forall j} \phi_{ij} x_j$$



- jak moc je velká podpora
- mnohdy nestabilní

$$y_j = \max_{\forall j} \{ \phi_{ij} x_j \}$$



- jen jedno pravidlo s maximální podporou
- avšak neříká nic o četnosti

Integrace dat - formalizace

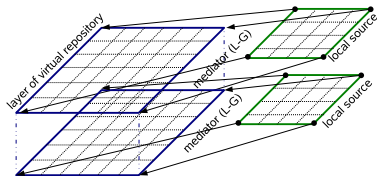
- každý element (lokálního) zdroje $S_i \in \mathcal{S}$ je mapován na globální element:

$$\Gamma_{S_i} = [\gamma'_{ij}]; \quad \gamma'_{ij} = \begin{cases} 1 & \text{když } e_i \sim e_j, \quad e_i \in \bigcup_{S \in \mathcal{S}} \mathcal{E}_S, e_j \in \mathcal{E}_{S_i} \\ 0 & \text{jinak} \end{cases}$$

- Řešení:
 - Centralizovaně
 - Decentralizovaně

Virtuální globální matice úložiště

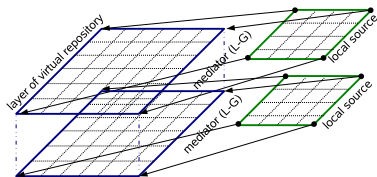
Centralizovaně



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

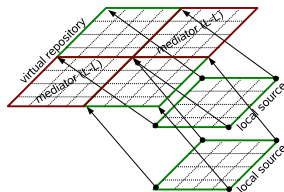
Virtuální globální matice úložiště

Centralizovaně



$$\Phi_{\mathcal{S}} = \sum_{\forall S_i \in \mathcal{S}} \Gamma_{S_i} \Phi_{S_i} \Gamma_{S_i}^T$$

Decentralizovaně



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

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

Decentralizované mapování na úrovni atributů

1 Binární:

$$\Pi_{kl} = \Delta_{S_l}^T \Psi_{kl} \Delta_{S_k}$$

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

Decentralizované mapování na úrovni atributů

1 Binární:

$$\Pi_{kl} = \Delta_{S_l}^T \Psi_{kl} \Delta_{S_k}$$

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

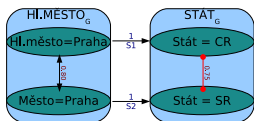
2 Vážené:

$$\Pi_{kl} = [\pi_{ij}], \pi_{ij} \in \langle 0, 1 \rangle$$

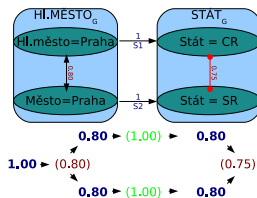
např.:

$$\pi_{ij} = \frac{|\mathcal{D}_\alpha^{S_k}(A_i) \cap \mathcal{D}_\alpha^{S_l}(A_j)|}{|\mathcal{D}_\alpha^{S_k}(A_i) \cup \mathcal{D}_\alpha^{S_l}(A_j)|}$$

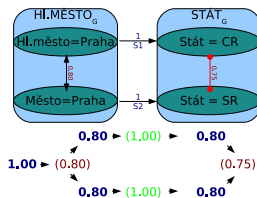
Nekonzistence a vážení pravidel



Nekonzistence a vážení pravidel

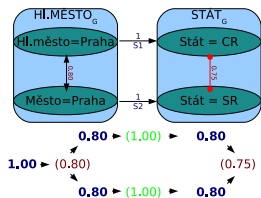


Nekonzistence a vážení pravidel

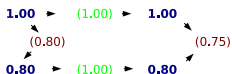


$$\nu = y_j y_{j'} \psi_{jj'} = 0.80 \cdot 0.80 \cdot 0.75 = 0.48$$

Nekonzistence a vážení pravidel

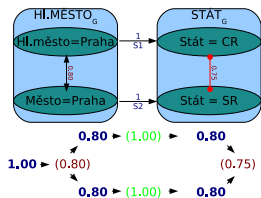


$$\nu = y_j y_{j'} \psi_{jj'} = 0.80 \cdot 0.80 \cdot 0.75 = 0.48$$

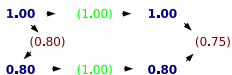


$$\nu = y_j y_{j'} \psi_{jj'} = 1.00 \cdot 0.80 \cdot 0.75 = 0.60$$

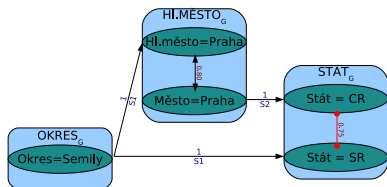
Nekonzistence a vážení pravidel



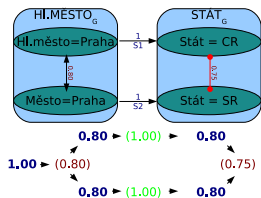
$$\nu = y_j y_{j'} \psi_{jj'} = 0.80 \cdot 0.80 \cdot 0.75 = 0.48$$



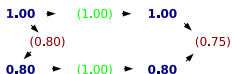
$$\nu = y_j y_{j'} \psi_{jj'} = 1.00 \cdot 0.80 \cdot 0.75 = 0.60$$



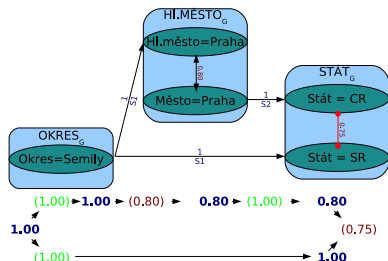
Nekonzistence a vážení pravidel



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



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



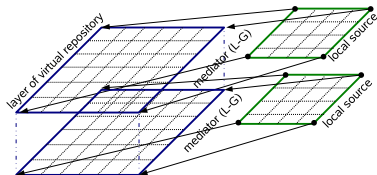
Reputace zdrojů

$$\rho_{S_i}^k = a(S_i^k) \cdot (1 + b(S_i^k)) \cdot c(S_i^k) \cdot (1 - d(S_i^k))$$

- $a(S_i^k)$.. podíl instancí funkčních závislostí pokrytých více zdroji
- $b(S_i^k)$.. podíl instancí funkčních závislostí následně potvrzených
- $c(S_i^k)$.. podíl dotazů na zdroj
- $d(S_i^k)$.. podíl nekonzistencí

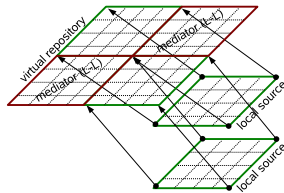
Centralizované a decentralizované řešení

Centralizované



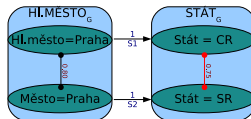
$$\Gamma'_{S_l} = \rho_l \Gamma_{S_l}$$

Decentralizované



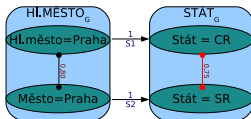
$$\Psi'_{kl} = \rho_{k/l} \Psi_{kl} = \rho_{k/l} \Gamma_k^T \Gamma_l$$

Váhy a reputace zdrojů



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

Váhy a reputace zdrojů



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



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

Závěr

- Vize platformy sémantického webu:
 - Důvěryhodné prostředí datových zdrojů
 - Decentralizované, aktivní, sebereflexivní (zpětná vazba)
- Orientace na přínos laického uživatele
- Možné (nepřesné/odhadnuté) propojení současného a sémantického webu:
 - použít metody extrakce dat
 - odhad struktury dat (sémantický kontext)
 - odhad integračních pravidel
 - odhad kvality (reputace) zdrojů