

Podobnost XML

Maticový model jako řešení pro DIS v prostředí XML

Vladimír Rejlek



Podobnost XML

Obsah

- jazyk XML
- pojem podobnosti v oblasti XML dokumentů
 - kategorizace přístupů
- přístup DIS s indexací
 - Maticový model

2

Podobnost XML

XML – ukázka dat

```

<knihovna>
  <knihka rok="2000">
    <nazev> XML pro každého </nazev>
    <autor>
      <jmeno> Jiří </jmeno>
      <prijmeni> Kosek </prijmeni>
    </autor>
    <vydavatel> Grada Publishing s.r.o. </vydavatel>
  </knihka>

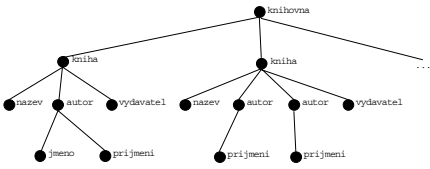
  <knihka rok="1998">
    <nazev> Microsoft Word pro pokročilé </nazev>
    <autor>
      <prijmeni> Šimek </prijmeni>
    </autor>
    <autor>
      <prijmeni> Vacek </prijmeni>
    </autor>
    <vydavatel> Computer Press </vydavatel>
  </knihka>
  ...

```

3

Podobnost XML

XML – stromová struktura



4

Podobnost XML

Přístupy k podobnosti XML

- 1) Klasické XML dotazovací jazyky rozšířené o operátor podobnosti
- 2) Závislost odpovědí na dotazy nad XML daty na granularitě těchto dat
- 3) Podobnost mezi XML dokumenty a DTD
- 4) **Přístup DIS s indexací**
- 5) Přístup DIS bez indexace

5

Podobnost XML

1. XML dotazovací jazyky rozšířené o operátor podobnosti

- vychází z již navržených XML dotazovacích jazyků (XQL, XML-QL)
- přidání operátoru podobnosti (~)
- dvojí použití
 - porovnání na konstantu
 - porovnání dvou částí dat mezi sebou

6

Podobnost XML

1. XML dotazovací jazyky rozšířené o operátor podobnosti

- dotaz v jazyku XXL

```

SELECT H, S
FROM cd01.xml, cd02.xml
WHERE ~cd AS C
AND C.#.interpret AS I
AND I = "Gustav Brom se svým orchestrem"
AND C.#. (~skladba)? AS S
AND S.~hudebnik AS H
AND H.# ~ "barytonsaxofon"

```

7

Podobnost XML

2. Závislost odpovědí na dotazy na granularitě dat

- orientace na "text-rich" dokumenty
- v čase konstrukce dotazu neznáme přesný tvar odpovědi
- chceme nalézt co nejrelevantnější kontext pro hledané termíny
- dva způsoby řešení:
 - přídavné informace
 - speciální operátory

8

Podobnost XML

2. Závislost odpovědí na dotazy na granularitě dat

- jazyk XIRQL přidává kontextové uzly

9

Podobnost XML

3. Podobnost mezi XML dokumenty a DTD

- zkoumá XML dokumenty, pro něž neznáme DTD
- pro XML dokument hledáme v množině DTD to nejpodobnější
- podobnost DTD mezi sebou

10

Podobnost XML

4. Přístup DIS s indexací

11

Podobnost XML

4. Přístup DIS s indexací

- dotazy typu:
 - "najdi všechny dokumenty z kolekce s co největší relevancí k zadanému popisu"
- klasické DIS vůbec nepočítají s vnitřní strukturou dokumentů
- potřeba rozšíření indexu o tyto informace

12

Podobnost XML

5. Přístup DIS bez indexace

- výstupem není množina relevantních dokumentů
- ale množina relevantních *podstromů dokumentů*

13

Podobnost XML

5. Přístup DIS bez indexace

- přibližné vnořování stromů (jazyk *ApproXQL*)

14

Podobnost XML

Maticový model pro XML DIS

- přístup DIS s indexací
- přímo vychází z vektorového modelu pro DIS
- dokument je v indexu reprezentován *maticí* namísto vektorem
- přidává nový prvek: *Malice převodu cest*

15

Podobnost XML

Reprezentace dokumentu

- Reprezentací dokumentu d_i v rámci kolekce c v maticovém modelu rozumíme matici D_i o rozměrech $m \times k$, kde m je počet měřených termů a k je počet cest v XML struktuře kolekce c . Hodnota $d_{i,j,s} \in \langle 0,1 \rangle$ udává váhu termu t_j na cestě s v dokumentu d_i .

$$D_i = \begin{bmatrix} d_{i,1,1} & d_{i,1,2} & \dots & d_{i,1,k} \\ d_{i,2,1} & d_{i,2,2} & \dots & d_{i,2,k} \\ \dots & \dots & \dots & \dots \\ d_{i,m,1} & d_{i,m,2} & \dots & d_{i,m,k} \end{bmatrix} \in \langle 0,1 \rangle^{m,k}$$

16

Podobnost XML

Příklad – strom kolekce

cesty:

- kniha->autor
- kniha->autor->jméno
- kniha->autor->příjmení

17

Podobnost XML

Příklad – matice dokumentů

	a	b	c	a	b	c	a	b	c
D_1 :	0	1	0	0	0	1	0	0	0
D_2 :	0.5	0	0	0	0	0	0.5	0	0
D_3 :	0.5	0	0	0.5	0	0	0	0	0

"david" "morrell" "caine"

cesty:

- kniha->autor
- kniha->autor->jméno
- kniha->autor->příjmení

18

Podobnost XML

Definice podobnosti

- podobnost

$$Sim_1(D_i, Q) = \sum_{l=1}^m \sum_{j=1}^k d_{i,j} * q_{l,j}$$

$$Sim_2(D_i, Q) = \sum_{l=1}^m \frac{\sum_{j=1}^k d_{i,j} * q_{l,j}}{\sqrt{\sum_{j=1}^k (d_{i,j})^2 * \sum_{j=1}^k (q_{l,j})^2}}$$

19

Podobnost XML

Matice převodu cest

- matice převodu cest
 - čtvercová reálná matice A o rozměrech $k \times k$, kde k je počet cest v kolekci; $a_{i,j} \in <0, 1>$ a $a_{i,i} = 1$
 - pro každou cestu vektor, který vyjádří vztah této cesty ke všem ostatním

20

Podobnost XML

Matice převodu cest

- Jednokrokový převod
 - Mějme matici dokumentu D o rozměrech $m \times k$ a matici převodu cest A o rozměrech $k \times k$, kde $a_{i,j} \in <0, 1>$ a $a_{i,i} = 1$. Pak *jednokrokovým převodem* rozumíme funkci $JP(D,A)=UD$, kde UD je opět matice o rozměrech $m \times k$ a platí, že:

$$UD = \left(\max \left(d_{i,j}, \max_{j=1}^k (a_{j,i} * d_{i,j}) \right) \right)_{i,j}$$

21

Podobnost XML

Matice převodu cest

- Převod
 - převodem matice dokumentu D podle matice převodu cest A rozumíme *tranzitivní uzávěr* funkce $JP(D,A)$
- zjednodušeně:
 - váha termu se distribuuje po cestách podle matice převodu cest pomocí funkce maximum

22

Podobnost XML

Příklad – matice převodu cest

matice převodu cest

a	b	c	a) kniha→autor	
a	1	0.2	0.2	b) kniha→autor→jméno
b	0.5	1	0	c) kniha→autor→prijmeni
c	0.5	0	1	

matice dokumentů po převodu

	a	b	c	a	b	c	a	b	c										
UD_1 :	(0.5	,	1	,	0.1	,	0.5	,	0.1	,	1	,	0	,	0	,	0)
UD_2 :	(0.5	,	0.1	,	0.1	,	0	,	0	,	0	,	0.5	,	0.1	,	0.1)
UD_3 :	(0.5	,	0.1	,	0.1	,	0.5	,	0.1	,	0.1	,	0	,	0	,	0)

23

Podobnost XML

Maticový model pro XML DIS

- využití matice převodu cest
 - každá matice dokumentu je před uložením do indexu upravena převodní maticí
 - jednotlivé cesty (elementy) se tak dostávají do vztahů
 - dva dokumenty se stejným termem na různých cestách si budou (mohou) více či méně podobné

24

