Přehled a možná vylepšení technik pro zpracování XML dokumentů v (O)RDBMS

Irena Mlýnková

irena.mlynkova@mff.cuni.cz

Univerzita Karlova
Matematicko-fyzikální fakulta
Katedra softwarového inženýrství
Malostranské nám. 25
118 00 Praha 1

Úvod (1)

- **XML = dnes jeden z nejlepších standardů pro reprezentaci dat**
  - **Roste využití XML technologií**
  - **Rostou požadavky na efektivní správu XML dokumentů a dotazování XML dat**
  - **Množina standardů vs. efektivní implementace**

⇒ **Přirozená myšlenka: Uložit a zpracovávat data pomocí (O)RDBMS**

⇒ **Výhoda: Využití osvědčených databázových mechanizmů (tj. indexy, transakce apod.)**

Úvod (2)

- Existuje množství různých technologií
  - Společné rysy => klasifikace
  - Výhody i nevýhody (žádná není ideální)
- Hlavní námitka: Databázové zpracování XML dat může být velmi pomalé
- Důležité otázky:
  - Je možné DB zpracování XML dat zefektivnit?
  - Kde jsou hranice těchto vylepšení?
  - Stojí výhody použití databází za to?

Obsah

1. Přehled existujících technologií
2. Přehled mapovacích metod
3. Možná vylepšení, otevřené problémy

Existující technologie

- Základní dělení: Určeny pro datově / dokumentově-orientované XML dokumenty
- Dokumentově-orientované dokumenty:
  - Obecně nepravidelná struktura
  - Pořadí sousedních elementů je významné
  - Obsahují elementy se smíšeným obsahem, komentáře, sekce CDATA…
- Datově-orientované dokumenty:
  - Opačné vlastnosti

Dokumentově-orientované technologie

- Zpracování vyžaduje uchovat strukturu
  - Někdy včetně detailů jako jsou bílé znaky
  - Tzv. „round tripping“
- Techniky:
  - BLOB / CLOB sloupec
    - Rychlé, ale nelze se efektivně dotazovat
  - Nativní XML databáze (varianty: PDOM, CMS…)
    - XML dotazovací jazyky, DOM, SAX…
    - Určitá daná strategie pro uložení dat

Datově-orientované technologie

- Společná myšlenka: Data ukládána a zpracovávána v DB
  - Mapovací metoda = přenos mezi XML a OR strukturami
- Přenos provádí jiný SW (middleware) nebo databáze sama (XML-enabled database)
- Úroveň round trippingu stačí nízká
  - Elementy, atributy, hierarchie, data samotná
- XML data binding – speciální případ

Obsah

1. Přehled existujících technologií
2. **Přehled mapovacích metod**
3. Možná vylepšení, otevřené problémy

Mapovací metody

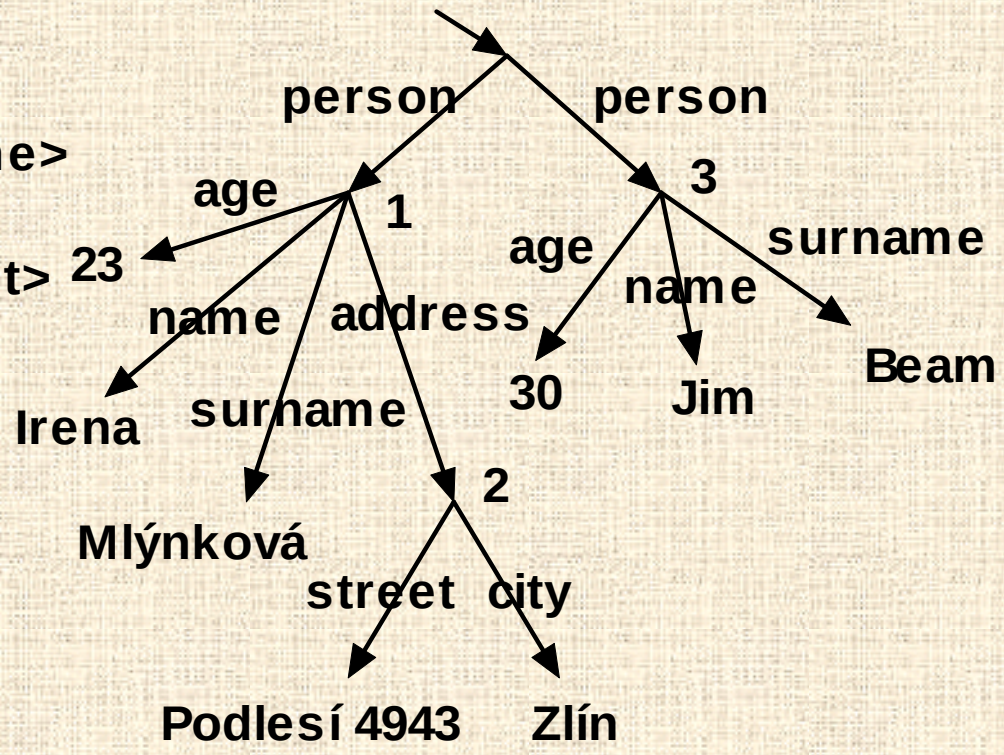
- Metody pro přenos dat mezi XML dokumenty a (O)R strukturami
- Základní klasifikace:
  - Generické – nevyužívají schéma ukládaných dokumentů
  - Schématem řízené – založeny na využití schématu ukládaných dokumentů
  - Uživatelsky definované – cílové schéma i mapování definuje uživatel
- Podrobný popis viz. [1]

Generické metody

- Pro ukládání nevyužívají schéma XML dat
- Dva přístupy:
  - Vytvořit obecné (O)R schéma pro uložení lib. XML dokumentu
  - Vytvořit schéma pouze pro určitou množinu XML dokumentů s danou strukturou
- Pohlíží na XML dokument jako na strom => problém uložení stromové struktury do (O)R struktur

Př. Generic-tree mapping

```
<person id=1 age=23>
   <name>Irena</name>
   <surname>Mlýnková</surname>
   <address id=2>
      <street>Podlesí 4943</street>
      <city>Zlín</city>
   </address>
</person>
<person id=3 age=30>
   <name>Jim</name>
   <surname>Beam</surname>
</person>
...
```



Schématem řízené metody (1)

- Založené na existujícím XML schématu
  - XML schéma je mapováno na (O)R DB schéma
  - Do jeho relací jsou ukládána data z validních XML dokumentů
- Cíl: Vytvořit optimální DB schéma = s „rozumným“ množstvím relací
- Vylepšování základní myšlenky:
  - Pro každý element vytvořit relaci složenou z jeho atributů
  - Vztahy element-podelement mapovat pomocí klíčů a cizích klíčů

Schématem řízené metody (2)

- Zdrojové XML schéma: DTD / XML Schema
- Cílové DB schéma: Relační / objektově-relační
- Další klasifikace:
  - Fixní metody – nevyužívají jiné informace než XML schéma samotné
  - Flexibilní metody – využívají další informace (např. ukázky XML dokumentů, XML dotazů apod.)
    - Snaha vytvořit optimální schéma pro určitou aplikaci

Vlastní metoda

- Schématem řízená fixní metoda
- Mapování struktur jazyka XML Schema na OR schéma
  - Zaměření na OO rysy a integritní omezení jazyka XML Schema
  - Využití OR prvků SQL (UDT, reference, typové tabulky, hnízdění)
- Implementace části jazyka XPath
- Podrobný popis viz. [2]

Obsah

1. Přehled existujících technologií
2. Přehled mapovacích metod
3. **Možná vylepšení, otevřené problémy**

Možná vylepšení?

- Flexibilní metody
  - Prozatím dva poměrně odlišné přístupy ([3], [4])
  - Zefektivnění pouze pro určitou aplikaci
- Kombinace generických a schématem řízených metod
  - Využití metod pro generování XML schématu pro množinu „podobných“ dokumentů
- Modifikace již na úrovni XML schématu
- Uložení pomocných informací pro pokrytí všech os XPath

Otevřené problémy

- Flexibilní metody vylepšují schéma pouze pro danou aplikaci
  - Co když se typická množina dotazů / dat změní?
  - Má smysl řešit průběžnou modifikaci schématu?
- Je možné vygenerovat „lepší“ XML schéma pro množinu XML dokumentů automaticky?
  - Nutno zavést vhodnou metriku pro hodnocení XML schémat (na způsob NF u relací)
- Do jaké míry je možné takto implementovat XML dotazovací jazyky? Jak efektivně?

Konec...

Reference

[1] Mlýnková, I., Pokorný, J.: XML in the World of (Object-) Relational Database Systems. Sborník z XIII. mezinárodní konference ISD 2004, Vilnius, Litva, září 2004.

[2] Mlýnková, I., Pokorný, J.: From XML Schema to Object-Relational Database – an XML Schema-Driven Mapping Algorithm. Sborník z mezinárodní konference IADIS WWW/Internet 2004, Madrid, Španělsko, říjen 2004.

[3] Bohannon, P., Freire, J., Roy, P., Siméon, J.: From XML Schema to Relations: A Cost-Based Approach to XML Storage. Sborník z mezinárodní konference ICDE 2002.

[4] Klettke, M., Meyer, H.: XML and Object-Relational Database Systems – Enhancing Structural Mappings Based on Statistics. Neformální sborník z WorkShopu WebDB 2000.