Uživatelsky řízené mapování XML dat do relací

(návrh SW projektu)

Vedoucí: Irena Mlýnková (irena.mlynkova@mff.cuni.cz)
Počet členů týmu: 5

Motivace:
Jedním ze způsobů jak implementovat správu XML dat je využití existujících relačních nebo objektově-relačních databázových systémů [3]. V dnešní době již sice existují efektivnější přístupy – tzv. nativní XML databáze – ale využití (objektově-) relačních systémů je stále v praxi nejpoužívanější technikou především proto, že existují jejich robustní, léty prověřené a tudíž spolehlivé aplikace, kterým v tomto ohledu žádná nativní XML databáze (zatím) konkurovat nemůže.
Základní myšlenka tohoto přístupu je následující: XML data jsou uložena do relací zvoleného databázového systému. XML dotazy nad uloženými daty jsou transformovány na odpovídající SQL dotazy a z výsledných relačních dat jsou původní XML data zpětně zrekonstruována. Hlavním problémem je tedy určení nejvhodnějšího relačního schématu, tj. způsobu jakým jsou XML data do databáze uložena – tzv. *mapování* XML dat do relací. Pravděpodobně žádná z existujících mapovacích metod totiž není univerzálně použitelná pro libovolnou aplikaci. Každá se hodí pouze pro určitý typ zpracování XML dat, zatímco pro jiné může být značně neefektivní. Proto je třeba při procesu mapování využít maximální množství informací o budoucí aplikaci, popř. přímo interakci s uživatelem.

Cíle projektu:
Cílem projektu je implementace systému, který bude pro správu a dotazování XML dat v (objektově-)relační databázi využívat tzv. *uživatelsky řízené* mapování XML dat do relací [4], tedy mapování které je ovlivněno interakcí s uživatelem, popř. dalšími pomocnými informacemi. V prvé řadě bude systém podporovat netriviální množinu obecných mapovacích metod (např. [7]), které bude možné vzájemně kombinovat. Uživatel tedy bude moci nejen určit jakým způsobem má být mapováno celé schéma, ale prostřednictvím anotací také změnit mapování u zvolených podčástí schématu. Podobně bude moci specifikovat množinu typických XML dotazů (popř. jiných operací) nad podčástmi schématu, celým schématem nebo více schématy. A další vhodné informace může poskytnout množina ukázkových XML dokumentů nebo určení vztahů mezi prvky různých schémat.
Informace budou využity pro definici schématu např. následujícím způsobem:

- Specifikované mapovací metody budou přímo aplikovány na určené podčásti, stejně jako v existujících systémech [4].
- Pomocí některé z metod pro určení podobnosti XML dat [5] (nebo vlastní metody) systém nalezne (strukturálně a/nebo sémanticky) podobné fragmenty XML schématu i ve zbytku schématu a nabídne uživateli jejich uložení stejným způsobem.
- Pro specifikované typické dotazy a ukázková data systém nabídne vhodné způsoby mapování. Pro tyto účely je možné využít např. některou z existujících (tzv. *adaptivních*) metod [6], které zkoušejí více možných mapování a hledají to nejefektivnější, nebo navrhnout vlastní.

Na základě konečné volby uživatele (nebo defaultního nastavení) systém vytvoří výsledné relační schéma, do něhož bude možné načíst (validní) XML data a nad nimi

se dotazovat prostřednictvím zvoleného XML dotazovacího jazyka (např. XPath). Vzhledem k tomu, že hlavní důraz projektu bude kladen na hledání vhodného mapování, v případě XML dotazů se předpokládá podpora pouze podmnožiny zvoleného jazyka, která umožní demonstrovat vlastnosti systému.

Další požadavky na program:

- Systém bude schopen zpracovat rozsáhlé kolekce XML dat, tj. velké XML dokumenty nebo velké množiny XML dokumentů.
- Podporovaná množina obecných mapování bude rozšiřitelná.
- Proces mapování bude využívat maximální možnou interakci s uživatelem a odpovídající implicitní nastavení, veškeré parametry bude možné nastavovat a při procesu mapování měnit.
- Pro práci se systémem bude implementováno vhodné rozhraní.
- Výsledné schéma (a související metadata) bude možné exportovat a přenést do jiného databázového systému. V ideálním případě bude systém pracovat pouze s metadaty popisujícími cílové schéma, která bude možné exportovat do SQL příkazů.
- Součástí projektu bude ukázka efektivity mapování pro několik zvolených situací, tj. předvedení smysluplnosti a použitelnosti programu v praxi.
- Program by měl být řešen jako freeware aplikace, pokud možno přenositelná.
- Veškerá dokumentace bude v angličtině, k projektu vznikne odpovídající webová stránka, která jej bude detailně popisovat.

Předpoklady:

Řešitelé projektu by měli mít absolvovanou přednášku *Technologie XML* (PRG036) nebo alespoň nastudované znalosti v rozsahu skript [1]. V průběhu implementace se předpokládá získání potřebných znalostí v rozsahu [2].

Předpokládaný průběh práce:

1. Analýza existujících implementací a přístupů v jednotlivých oblastech
2. Podrobná specifikace konkrétních funkcí systému, architektury a rozhraní mezi jednotlivými moduly
3. Implementace projektu
4. Testy, ladění
5. Analýza efektivity výsledných mapování pro několik zvolených aplikací
6. Dokumentace (programátorská, uživatelská, instalační)

Poznámka:

Problematiku řešenou v rámci implementace projektu je možné rozšířit do diplomových prací.

Doporučená literatura:

[1] *Mlýnková, I. – Pokorný, J. – Richta, K. – Toman, K. – Toman, V.: Technologie XML. Univerzita Karlova v Praze, Česká republika, září 2006. Vydalo nakladatelství Karolinum, ISBN 80-246-1272-0.*

[2] *W3C Technical Reports and Publications: http://www.w3.org/TR/*

[3] Databázové systémy:
Oracle Database: http://www.oracle.com/database/index.html

IBM DB2: http://www-306.ibm.com/software/data/db2/
MS SQL Server: http://www.microsoft.com/sql/default.mspx

[4] Uživatelsky řízené mapování:
Balmin, A. – Papakonstantinou, Y.: Storing and Querying XML Data Using Denormalized Relational Databases, The VLDB Journal, 2005, 14(1), pages 30–49.

Du, F. – Amer-Yahia, S. – Freire, J.: ShreX: Managing XML Documents in Relational Databases. In VLDB '04: Proc. of the 30th Int. Conf. on Very Large Data Bases, Morgan Kaufmann Publishers Inc., Toronto, ON, Canada, 2004, pages 1297–1300.

[5] Podobnost XML dat:
Do, H. H. – Rahm, E.: COMA – A System for Flexible Combination of Schema Matching Approaches. In VLDB'02: Proc. of the 28th Int. Conf. on Very Large Data Bases, pages 610–621, Hong Kong, China, 2002. Morgan Kaufmann Publishers Inc.

Madhavan, J. – Bernstein, P. A. – Rahm, E.: Generic Schema Matching with Cupid. In VLDB'01: Proc. of the 27th Int. Conf. On Very Large Data Bases, pages 49–58, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[6] Adaptivní metody:
Klettke, M. – Meyer, H.: XML and Object-Relational Database Systems – Enhancing Structural Mappings Based on Statistics. In Lecture Notes in Computer Science, volume 1997, pages 151–170, 2000.

Ramanath, M. – Freire, J. – Haritsa, J. – Roy, P.: Searching for Efficient XML-to-Relational Mappings. In XSym'03: Proc. of the 1st Int. XML Database Symposium, volume 2824, pages 19–36, Berlin, Germany, 2003. Springer.

Xiao-ling, W. – Jin-feng, L. – Yi-sheng, D.: An Adaptable and Adjustable Mapping from XML Data to Tables in RDB. In Proc. of the VLDB'02 Workshop EEXTT and CAiSE'02 Workshop DTWeb, pages 117–130, London, UK, 2003. Springer-Verlag.

Zheng, S. – Wen, J. – Lu, H.: Cost-Driven Storage Schema Selection for XML. In DASFAA'03: Proc. of the 8th Int. Conf. on Database Systems for Advanced Applications, pages 337–344, Kyoto, Japan, 2003. IEEE Computer Society.

[7] Obecné fixní mapovací metody:
Shanmugasundaram, J. – Tufte, K. – Zhang, C. – He, G. – DeWitt, D. J. – Naughton, J. F.: Relational Databases for Querying XML Documents: Limitations and Opportunities In VLDB '99: Proc. of the 25th Int. Conf. on Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pages 302–314.

Florescu, D. – Kossmann, D.: Storing and Querying XML Data Using an RDMBS, IEEE Data Eng. Bull. 22(3), 1999, pages 27–34.

Yoshikawa, M. – Amagasa, T. – Shimura, T. – Uemura, S.: XRel: A Path-Based Approach to Storage and Retrieval of XML Documents Using Relational Databases, ACM Trans. Inter. Tech. 1(1), 2001, pages 110–141.