Systém pro odvozování XML schémat

(návrh SW projektu)

Vedoucí: Irena Mlýnková (irena.mlynkova@mff.cuni.cz)
Martin Nečaský (martin.necasky@mff.cuni.cz)

Předpokládaný počet řešitelů: 5

Motivace:
Většina metod pro zpracování XML dat je založena na existenci jejich přesného a konzistentního XML schématu, obvykle vyjádřeného v jazyce DTD nebo XML Schema. Ze statistických analýz ovšem plyne, že většina reálných kolekcí XML dat schéma nemá, nebo vůči němu dokumenty 100% validní nejsou. To je obvykle způsobeno tím, že XML schéma není považováno za přesný popis struktury XML dat, ale za (typicky neaktualizovanou) formu dokumentace.
Na druhou stranu problém odvození XML schématu pro danou kolekci XML dokumentů není triviální. Pro složitější data, jejichž přesnou strukturu a sémantiku navíc uživatel obvykle nezná, už to ani není v lidských silách. Je tedy třeba navrhnout a implementovat metody automatického nebo semi-automatického odvozování XML dat. Navíc, vzhledem k tomu, že vyjadřovací síla jazyka XML Schema je omezená především na popis struktury XML dat, dalším, z hlediska optimalizace zpracování XML dat zajímavým, výstupem takovýchto odvozovacích metod mohou být různá integritní omezení (IO), která v jazyce XML Schema vyjádřit nelze.

Cíle projektu:
Cílem projektu je implementace systému, který bude schopen pro daná vstupní data (tj. především XML dokumenty) odvodit odpovídající XML schéma. Pro tento účel by měl systém využívat maximální množství dostupných informací (ovšem na druhou stranu by na jejich existenci neměl být závislý). Jedná se především o:

- strukturu a sémantiku vstupních XML dokumentů,
- interakci s uživatelem,
- analýzu původního, již neplatného XML schématu a
- analýzu XML dotazů.

Z hlediska analýzy struktury vstupních XML dokumentů bude systém využívat a kombinovat ověřené přístupy z existujících metod odvozování XML schémat [3]. Ty mohou být dále rozšířeny např. o využití statistických analýz XML dokumentů nebo sémantiky XML dat vedoucí k přesnějším a reálnějším výsledkům.
Využití interakce s uživatelem by mělo být maximálně uživatelsky přívětivé. Cílem je, aby mohl uživatel proces odvozování XML schématu ovlivňovat a zpřesňovat, ale současně aby nebyl nucen ke složitým rozhodnutím, náročné analýze vstupních dat, zkoumání velkého množství variant apod.
Dalším možným vstupem může být původní XML schéma. Přestože už vstupní dokumenty nejsou vůči tomuto schématu validní, některé jejich části mohou stále být (a není je tedy nutné odvozovat znovu), zatímco jiné mohou napomoci při volbě vhodného zobecnění triviální schématu.
Nejzajímavější částí systému bude využití informací dostupných v XML dotazech, které mohou být poskytnuty jako vstupní informace spolu s daty. XML dotazy mohou

podat přesnější informace o výsledném schématu jako např. datových typech, prvcích, které se v datech nevyskytují, povinnosti výskytu, vícenásobném výskytu nebo dokonce funkčních závislostech nad daty.
Výstupní, odvozené schéma by mělo maximálně využívat všech konstruktů jazyka XML Schema jako je odvozování datových typů, substituce nebo integritní omezení. Mimo to může být výstupem také sada integritních omezení, které není možné v jazyce XML Schema vyjádřit, ale která ještě přesněji popisují strukturu XML dat a mohou pomoci při optimalizaci jejich zpracování.

Další požadavky na program:

- Program bude schopen zpracovávat rozsáhlejší kolekce XML dat, tj. velké XML dokumenty nebo velké množiny XML dokumentů.
- Program by měl být řešen jako freeware aplikace, jejíž instalace nebude vyžadovat složité úkony, bude pokud možno přenositelná atd. Cílem je zajistit, aby aplikaci využívalo co nejvíce uživatelů.
- 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
2. Podrobná specifikace konkrétních funkcí systému, architektury a rozhraní mezi jednotlivými moduly
3. Implementace projektu
4. Testy na reálných XML datech, ladění
5. 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] Metody odvozování XML schématu:
Ahonen, H. (1996), Generating Grammars for Structured Documents Using Grammatical Inference Methods, Report A-1996-4, Dept. of Computer Science, University of Helsinki.

Bex, G. J., Neven, F. & Vansummeren, S. (2007), Inferring XML Schema Definitions from XML Data, in 'VLDB'07', ACM, Vienna, Austria, pp. 998–1009.

Garofalakis, M., Gionis, A., Rastogi, R., Seshadri, S. & Shim, K. (2000), XTRACT: a System for Extracting Document Type Descriptors from XML Documents, in 'SIGMOD'00', ACM, New York, NY, USA, pp. 165– 176.

Moh, C.-H., Lim, E.-P. & Ng, W.-K. (2000), Reengineering Structures from Web Documents, in 'DL'00', ACM, New York, NY, USA, pp. 67–76.

Wong, R. K. & Sankey, J. (2003), On Structural Inference for XML Data, TechReport UNSW-CSE-TR-0313, School of Computer Science, University of NSW.