Systém pro generování umělých testovacích XML dat

(návrh SW projektu)

Vedoucí: Irena Mlýnková (irena.mlynkova@mff.cuni.cz)
Počet členů týmu: 4

Motivace:
V současné době existuje obrovské množství metod pro efektivní zpracování XML dat a stále přibývají nové. Každý autor pochopitelně svoji metodu nějakým způsobem experimentálně otestuje a popíše, v čem jsou její výhody. Pokud by se ale chtěl konkrétní uživatel rozhodnout, která metoda je pro jeho účely nejvhodnější, z popisu jednotlivých metod to lze obvykle velmi obtížně. Metody jsou totiž typicky testovány na různých datech pocházejících z různých zdrojů, které již nemusejí existovat, mohly být vytvořeny pouze pro testovací účely, se speciálními charakteristikami určité aplikace apod. V dnešní době sice existuje několik projektů, které jsou buďto pevnou (nebo postupně rozšiřovanou) databází reálných testovacích XML dat, popř. odpovídajících dotazů (např. projekt Inex), nebo umožňují umělá testovací data generovat (např. projekt XMark). Testovací množina fixních dat je ovšem poměrně hodně omezující, zatímco v případě generátorů umělých dat jsou možné parametry v existujících systémech pouze triviální.

Cíle projektu:
Prvním cílem projektu je implementace systému, který bude schopen na základě parametrů zadaných uživatelem generovat umělá testovací XML data. Vstupem programu bude množina charakteristik [3] jako např. počet XML dokumentů a jejich velikost (popř. přípustný rozsah velikostí), hloubka dokumentů, přípustný fan-out elementů (tj. počty podelementů různých typů, počty atributů apod.), procentuální zastoupení různých prvků (např. smíšeného obsahu, atributů apod.), rozmístění různých prvků v rámci úrovní dokumentu, zvolené rozdělení, které má daná charakteristika (např. hloubka) na výstupu mít, schéma, kterému musejí XML dokumenty vyhovovat apod. Systém zkontroluje přípustnost zvoleného nastavení (neboť ne všechny požadované kombinace charakteristik bude možné splnit současně) a vygeneruje odpovídající množinu XML dokumentů. Tato část projektu by měla být řešena minimálně v rozsahu existujících systémů [4].
Implementovaný systém bude dále schopen k vytvořeným datům vygenerovat odpovídající XML schéma, jehož charakteristiky bude opět specifikovat uživatel. Příkladem charakteristik [3] může být cílový jazyk (DTD, XML Schema apod.), přípustné konstrukty jazyka (např. použití rekurzivity, dědičnosti apod.), procentuální zastoupení a rozmístění povolených konstruktů (podobně jako v předchozím případě), volba konstruktů v případě, že je možné danou situaci popsat více způsoby (např. dědičnost vs. globální prvky jazyka XML Schema), přípustná míra zobecnění regulárních výrazů apod. Pro generování je možné použít (modifikovat) některou z existujících metod [5] nebo navrhnout vlastní.
Třetím typem výstupu bude sada XML dotazů nad požadovanými daty opět vygenerovaná dle zadaných charakteristik. Kromě možnosti zadat požadovanou složitost dotazu (např. využití různých os jazyka XPath apod.) nebo jeho typ [6], by jednou z hlavních charakteristik mohla být úspěšnost dotazů nad požadovanými daty, tj. specifikace složitosti výstupu, který by měly dotazy vracet (např. podstromy určité hloubky, % vstupních dat, které mají být na výstupu dotazu apod.)

Posledním cílem projektu bude schopnost poskytovat vygenerovaná XML data nejen ve formě textových souborů, které si budoucí uživatel načte prostřednictvím libovolného rozhraní pro XML dokumenty, ale i ve formě vhodného programátorského rozhraní.

Další požadavky na program:

- Program bude schopen generovat rozsáhlé kolekce XML dat, tj. velké XML dokumenty nebo velké množiny XML dokumentů.
- Generování dat bude možné zopakovat, tj. zadáním stejných parametrů a stejného náhodného semínka může libovolný uživatel získat totožná data. Cílem je, aby uživatel popsal množinu testovacích dat prostřednictvím sady parametrů, kterou může libovolný jiný uživatel využít pro vygenerování totožných dat a tudíž totožného vstupu pro vlastní metodu.
- Systém bude umožňovat generování dat s postupně se měnícími parametry, např. s hloubkou rostoucí dle zadaného přírůstku apod. Cílem je, aby mohl uživatel měřit efektivitu dané metody vzhledem k měnícím se charakteristikám dat.
- V programu bude možné předdefinovat nastavení sady charakteristik pro různé standardní typy dat – např. datově-orientované / dokumentově-orientované XML dokumenty apod. Součástí výsledného programu budou i tyto sady a experimentální srovnání několika vybraných existujících implementací (např. nativních XML databází vs. relačních databází s podporou XML) nad těmito sadami dokumentů.
- Program by měl být řešen jako freeware aplikace, jejíž instalace nebude vyžadovat složité úkony (např. instalaci a parametrizaci drahého nebo složitého databázového systému apod.), nebude mít vysoké nároky na systém, bude pokud možno přenositelná atd. Cílem je zajistit, aby aplikaci využívalo co nejvíce uživatelů. Je třeba uvážit i možnost webového rozhraní.
- 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* 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. Návrh konkrétních funkcí systému, návrh rozhraní mezi jednotlivými moduly
3. Implementace projektu
4. Testy, ladění
5. Návrh předdefinovaných typů dat a odpovídající analýza chování vybraných metod
6. Dokumentace

Poznámka:

Problematiku řešenou v rámci implementace projektu je možné rozšířit do (až čtyř) 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] *Mlynkova, I. - Toman, K. - Pokorny, J.: Statistical Analysis of Real XML Data Collections. Technical report 2006/5. Charles University, Prague, Czech Republic, June 2006, 43 pages.*
http://kocour.ms.mff.cuni.cz/~mlynkova/doc/tr2006-5.pdf

- Technická zpráva MFF UK popisující statistickou analýzu reálných XML dat. Obsahuje velké množství definic různých konstruktů jazyka XML Schema i typů XML dat.

[4] Existující projekty podobného zaměření:
W3C Test Suites - http://www.w3.org/XML/Test/
XMark - http://monetdb.cwi.nl/xml/
XOO7 Benchmark - http://www.comp.nus.edu.sg/~ebh/XOO7.html
XBench - http://se.uwaterloo.ca/~ddbms/projects/xbench/
The Michigan Benchmark - http://www.eecs.umich.edu/db/mbench/
XMach-1 - http://dbs.uni-leipzig.de/en/projekte/XML/XmlBenchmarking.html
ToXgene - http://www.alphaworks.ibm.com/tech/toxgene

[5] Metody generování schématu pro množinu XML dokumentů:
Vošta, O.: Automatická konstrukce schématu pro množinu XML dokumentů. Diplomová práce, MFF UK, 2005.
http://kocour.ms.mff.cuni.cz/~mlynkova/dp/Vosta.ps

Moh, C.-H. - Lim, E.-P. - Ng, W.-K.: Re-engineering Structures from Web Documents. In DL '00: Proc. of the 5th ACM Conf. on Digital Libraries, pages 67-76, New York, NY, USA, 2000. ACM Press.

Garofalakis, M. - Gionis, A. - Rastogi, R. - Seshadri, S. - Shim K.: XTRACT: a System for Extracting Document Type Descriptors from XML Documents. In SIGMOD '00: Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, pages 165-176, New York, NY, USA, 2000. ACM Press.

Ahonen, H.: Generating Grammars for Structured Documents Using Grammatical Inference Methods. Report A-1996-4, Department of Computer Science, University of Helsinki, 1996.

[6] Typy dotazů:
Specifikace jazyka XPath, popř. XQuery:
http://www.w3.org/TR/2007/REC-xpath-datamodel-20070123/
http://www.w3.org/TR/2007/REC-xpath-functions-20070123/
http://www.w3.org/TR/2007/REC-xquery-20070123/
http://www.w3.org/TR/2007/REC-xquery-semantics-20070123/

XML Query Use Cases: http://www.w3.org/TR/xquery-use-cases/