Systém pro analýzu XML dat

(návrh SW projektu)

Vedoucí: Irena Mlýnková (irena.mlynkova@mff.cuni.cz)

Členové týmu: Jakub Stárka
Jan Sochna
Martin Svoboda
Jiří Schejbal

Motivace:

Většina existujících metod pro zpracování XML dat obvykle není schopna efektivně podporovat všechny konstrukty, které povolují specifikace W3C [2]. Je tedy přirozené, že autoři využívají předpoklad, že některé konstrukty nejsou obvykle v reálných datech využívány a tudíž jejich neefektivní zpracování není zásadním problémem. Otázkou ovšem je, které konstrukty to opravdu jsou a naopak na které by se měly nové metody zaměřovat.

Odpověď na tuto otázku se snaží podat nejrůznější analýzy reálných XML dat [3]. Ale i když se toto řešení zdá být snadné, při jeho realizaci narážíme na množství problémů. Přestože Internet je dobrým a typickým zdrojem informací, už sběr reálných dat je poměrně obtížný. Náhodným stahováním obvykle získáme množinu dat, která obsahuje velké množství triviálních XML dokumentů (např. dokumentů s jediným elementem), které pak zkreslují výsledné analýzy. Navíc přibližně 50% XML dokumentů není buďto správně strukturovaných nebo validních vůči svému schématu. Další problém je, že schéma XML dokumentů se už obvykle nenachází v uvedeném místě a je třeba jej dohledat ručně. A hlavním problémem zůstává fakt, že většina autorů existujících analýz implementovala systém jako jednoduchou pomocnou aplikaci, jejíž rozšiřitelnost či opakované použití je téměř nemožné. Navíc, jelikož se autoři zaměřují především na analýzu dat, je nutné výstupy programů do vhodné grafické nebo tabulkové podoby zpracovat ručně.

Cíle projektu:

Cílem projektu je implementace systému, který bude schopen analyzovat strukturu a složitost reálných XML dokumentů, jejich XML schémat a operací nad nimi (tj. dotazů, transformací apod.).

Analýzy nad daty budou sledovat obecné charakteristiky jako např. velikost dokumentu, počet elementů a atributů, hloubka dokumentu, fan-out (tj. počet podelementů a atributů), zastoupení textového obsahu, smíšeného obsahu nebo rekurze, jejich složitost apod. Podobné charakteristiky bude možné sledovat i pro XML schémata a výsledky vzájemně srovnat. Pro schémata budou dále sledovány jejich specifické vlastnosti (např. délka cyklů, typ rekurze, fan-in apod.) nebo výskyty různých operátorů a speciálních prvků (např. neuspořádané množiny dat, dědičnost apod.). Rozsah sledovaných konstruktů by měl pokud možno pokrývat existující práce [3]. Další funkcí programu bude hledání fragmentů XML dokumentů a schémat odpovídajících nebo podobných zadaným vzorům nebo vyhovujících zadané specifikaci. Pro tyto účely je možné využít některý z existujících přístupů [4] nebo navrhnout vlastní.

Analýzy nad operacemi s daty se zaměří na jejich složitost (např. délky cest, využití poddotazů apod.), využití různých konstruktů (např. os, predikátů, funkcí, konstruktorů apod.), kontext (např. XPath dotazy v XSLT a v XML Schema) apod.

Uživatel bude moci dále zvolit, které charakteristiky mají být analyzovány a v jakém rozsahu (např. v rámci celé množiny vstupních dat, uživatelem zvolených kategorií vstupních dat, jednotlivých dokumentů, jednotlivých úrovní dokumentů, jednotlivých uzlů, vybraných jazyků apod.).
Vstupní data budou buď poskytnuta uživatelem, nebo je systém automaticky stáhne z Internetu. Jelikož hlavním cílem programu bude analýza dat, pro potřeby automatického stahování se předpokládá využití/modifikace vhodné existující aplikace a/nebo implementace pouze klíčových částí potřebných pro systém. Např. pro XML dokumenty by měl být systém schopen dohledat jejich XML schéma, pro XML schémata rozdělená do více souborů všechny potřebné součásti apod.
Vstupní data bude dále možné filtrovat/kategorizovat dle zadaných parametrů, jako např. velikost množiny dat, velikost/složitost dat, zdroj dat, typ (např. schémata v jazyce DTD, dotazy v jazyce XQuery apod.) apod.
Dále bude systém schopen detekovat korektnost dat, tj. správnou strukturovanost XML dokumentů a jejich validitu vůči schématu. S nekorektními daty by se měl v maximální možné umět vyrovnat, např. prostřednictvím vlastního rozhraní pro načítání XML dat, které bude schopné zpracovávat i nekorektní data, automatickými opravami jednodušších případů (např. doplnění XML deklarací, doplnění chybějících koncových tagů apod.), interakcí s uživatelem apod. Cílem je aby bylo možné XML data opravit co nejjednodušeji a nejrychleji.
Nejjednodušším výstupem analýzy budou výsledné hodnoty (např. hloubky, počty výskytů, procentuální zastoupení konstruktů apod.). Nad výstupy bude systém podporovat vhodné matematické operace (např. minimum, maximum, počet, průměr, medián apod.). Zvolená výstupní data bude možné zobrazit ve formě grafů a tabulek nebo exportovat ve vhodných, dále zpracovatelných formátech (např. XSL).

Další požadavky na program:

- Program bude schopen zpracovávat rozsáhlé kolekce XML dat, tj. velké XML dokumenty nebo velké množiny XML dokumentů.
- Výsledky analýz budou průběžně ukládány, aby nebylo nutné opakovat výpočty při přidání nových vstupních dat.
- Program bude podporovat schémata v jazyce DTD i XML Schema, zpracovávat XML schémata rozmístěná ve více dokumentech, schémata umístěná přímo v XML dokumentech apod.
- Z hlediska operací nad daty se program zaměří především na jazyk XPath a jeho využití v dalších XML technologiích jako je XQuery, XSLT nebo jazyk XML Schema.
- Množina analyzovaných konstruktů bude rozsáhlá a snadno rozšiřitelná.
- Program by měl být řešen jako freeware aplikace, jejíž instalace nebude vyžadovat složité úkony (např. instalaci a parametrizaci drahého nebo složitého databázového systému apod.), 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 v jednotlivých oblastech
2. Podrobná specifikace konkrétních funkcí systému, architektury a rozhraní mezi jednotlivými moduly
3. Implementace projektu
4. Testy, ladění
5. Analýza netriviální množiny reálných dat – náhodně vybraných i standardních kolekcí (např. Inex, Shakespeare v XML, Bible v XML apod.) – a srovnání s výsledky existujících prací [3].
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] Existující analýzy:
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

Sahuguet, A.: Everything You Ever Wanted to Know About DTDs, But Were Afraid to Ask (Extended Abstract). In Selected papers from the 3rd International Workshop WebDB 2000 on The World Wide Web and Databases, pages 171–183, London, UK, 2001. Springer-Verlag.

Choi, B.: What are real DTDs like? In WebDB '02, Proceedings of the 5th International Workshop on the Web and Databases, pages 43–48, Madison, Wisconsin, USA, 2002. ACM Press.

Bex, G. J. – Neven, F. – Van den Bussche, J.: DTDs versus XML Schema: a Practical Study. In WebDB '04, Proceedings of the 7th International Workshop on the Web and Databases, pages 79–84, New York, NY, USA, 2004. ACM Press.

McDowell, A. – Schmidt, C. – Yue, K.: Analysis and Metrics of XML Schema. In SERP '04, Proceedings of the International Conference on Software Engineering Research and Practice, pages 538–544. CSREA Press, 2004.

Mignet, L. – Barbosa, D. – Veltri, P.: The XML Web: a First Study. In WWW '03, Proceedings of the 12th international conference on World Wide Web, Volume 2, pages 500–510, New York, NY, USA, 2003. ACM Press.

[4] 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.