Posudek diplomové práce Tomáše Kohana

User-Defined XML-to-Relational Mapping

Předkládaná diplomová práce se zabývá problematikou uživatelsky definovaného mapování XML dat do relací, jehož hlavní rysem je umožnit uživateli ovlivnit proces mapování. Namísto použití fixní metody, která nezohledňuje aktuální aplikaci, metody uživatelsky definovaného mapování využívají požadavky specifikované uživatelem prostřednictvím anotací XML schématu ukládaných XML dat. Cílem práce bylo analyzovat tyto metody a na základě jejich nevýhod navrhnout vlastní řešení a jeho vlastnosti ověřit prostřednictvím prototypové implementace.

Hlavní klady práce:

- Práce má strukturu jakou by měla mít (tj. úvod, přehled použitých technologií, analýzu existujících přístupů, vlastní návrh, popis implementace a experimentů, závěr), dostatečné množství referencí, méně podstatné části jsou korektně odsunuty až do příloh.
- Nápad na doplnění anotací obsahujících typické dotazy nad XML daty je velmi zajímavý a bezpochyby originální.

Přestože struktura práce je korektní, z hlediska obsahu požadavky na diplomovou práci rozhodně splněny nejsou.

Hlavní připomínky k práci:

- Úroveň angličtiny je opravdu nízká – v textu je velké množství chyb (viz níže), a to nejen v použitých pojmech, ale i v základních věcech jako je tvorba přídavných jmen, jednotné a množné číslo, členy apod. Při čtení je sice možné pochopit co chtěl autor větou říci, ale rozhodně se nejedná o anglický text. Práce působí dojmem, že ji po sobě autor ani jednou nečetl.
- V textu práce se několikrát vyskytuji kapitoly, které byly převzaty z cizích zdrojů, což by nevadilo, pokud by bylo důsledně zmíněno, že se jedná o převzatý text. Vzhledem k tomu, že v těchto pasážích je angličtina velmi dobrá, je navíc rozdíl opravdu markantní.
- Autor používá pojmy, které se v dané oblasti nepoužívají, což působí dojmem, že existujícím pracím nevěnoval dostatečnou pozornost.
- Kapitoly neobsahují to, co by obsahovat měly (viz níže) – např. z úvodu není jasné čím se bude práce zabývat, v použitých technologiích není vysvětleno vše, co je dále používáno apod.
- Hlavní problém práce spočívá v tom, že její hlavní přínos je opravdu odbytý, a to z hlediska popisu i provedení. Popis sice opět vypadá zajímavě – obsahuje velké množství obrázků a schémat, ale při bližším ohledání je text velmi obecný, což budí představu, že toho asi mnoho realizováno nebylo. To zajímavé, co je přínosem práce a čím by se měla tudíž detailně zabývat, je popsáno na několika málo řádcích. Z popisu není jasné jak lze přesně nové typy anotací používat, jak funguje algoritmus, který je zpracovává, jaký efekt budou mít specifikované dotazy na výsledné schéma, v čem jsou výhody tohoto přínosu oproti atributům používaným v existujících pracích atd.

Připomínky k jednotlivým kapitolám:

- Abstrakt je příliš „abstraktní", měl by blíže popisovat co konkrétně je v práci řešeno a jak, nejen, že je něco řešeno lépe než v existujících pracích. I v krátkém českém abstraktu je navíc velké množství hrubek a překlepů.
- Kapitola *Introduction* neobsahuje jasné vysvětlení toho, čím se práce zabývá a proč. Tvrzení, že XML je podmnožinou HTML, je navíc chybné.
- Kapitola *Goals* popisuje velmi neurčitě, ze práce popisuje nějaké metody a cílem je implementovat nějakou jinou metodu, která má něco nového. Ale netušíme co.
- Kapitola *Comparison with DTD* na str. 9 je nic neříkající.
- Kapitola 2.3 neobsahuje jedinou referenci a přitom se evidentně nejedná o původní text. Stejně tak neobsahuje žádný příklad a pochopení textu je tudíž velmi obtížné.
- V kapitolách 3.1 a 3.2 jsou uvedena tvrzení jaké dotazy jsou typické nad jednotlivými typy dokumentů. Toto tvrzení není nijak podloženo (např. referencí) ani blíže zdůvodněno/vysvětleno. Příklady zmíněných typů dotazů by pochopení textu také prospěly. Příklad 3.1 není příliš vhodný – pro názvy elementů jsou použity jakési zkratky, u nichž čtenář jen stěží odhaduje význam. Příklad by měl být samovysvětlující nebo by zkratky měly být v textu vysvětleny.
- Informace z kapitoly 3.3 jsou zjevně převzaty z určitého zdroje, který není zmíněn.
- V kapitole 3.4.1 je poměrně matoucím způsobem popsáno současně KFO mapování a Edge mapping, což jsou dvě odlišné metody.
- V kapitole 4 je popsán pouze jediný existující přístup uživatelsky definovaného mapování. Popsané systémy MXM i ShreX jsou totiž díla stejných autorů a jedná se o postupná vylepšení stejného nápadu. Naopak zde vůbec není zmíněn systém XCacheDB (*Balmin, A., Papakonstantinou, Y. (2005), 'Storing and Querying XML Data Using Denormalized Relational Databases', The VLDB Journal 14(1), 30–49*), který je opravdu dalším významným zástupcem těchto přístupů.
- Vzhledem k výrazně dobré úrovni angličtiny v kapitole 5 je zřejmé, že obsah byl opět převzat z nějakého zdroje, pravděpodobně ze specifikací příslušných databází.
- V kapitole 7 se dozvíme omezení implementace, které by nevadilo, pokud by neimplementované součásti byly vyřešeny alespoň teoreticky v kapitole 6. To ale nejsou.
- Experimenty v kapitole 8 obsahují pouze dva příklady anotování schématu a jaký počet tabulek vznikne. Vzhledem k tomu, že z obecného popisu navrženého řešení netušíme proč je počet tabulek zredukován, nelze toto považovat za experiment.

Připomínky k anglickému textu:

- V textu schází obrovské množství určitých i neurčitých členů a jejich použití není vždy správné – např. str. 7 „a XML document", str. 23 „In the case DTD is used", str. 44 „as an result".
- Nesprávné použití různých časů je poměrně matoucí a stěžuje pochopení textu.
- V mnoha případech chybí na konci slovesa ve třetí osobě jednotného čísla „s".
- Poměrně velké množství chyb se vyskytuje v použití jednotného a množného čísla – např. str. 4 „to handle this problems", „parenthesis" vs. parentheses, str. 5 „every XML documents", „a given schemata", str. 14 „HL7 documents illustrates", str. 19 „each nodes id contain", str. 23 „input document are described".
- Volba některých anglických termínů je nestandardní nebo nevhodná – např. str. 6 „the definition looks like", str. 8 „lineal descendant", str. 8 „improvements against DTD",

str. 16 „documents observing the same XML scheme“, „XML-to-relations mapping“, str. 64 „benefit of this thesis“.

- Chyby se také vyskytují při vytváření přídavných jmen – např. str. iv „present work“ namísto „presented work“, str. 1 „easy-to-learn“ nikoli „easy to learn“, str. 42 „this performance is platform dependence“.
- Za „on the other hand“, „however“ apod. se obvykle píše čárka.
- Kromě členů, předložek a spojek se v angličtině v nadpisech píší všechna slova velkými písmeny.
- Pokud jsou v anglickém textu vyjmenovány alespoň 3 položky, píše se před „and“ nebo „or“ čárka.

Další připomínky:

- Definovaný pojem není často dodržen – např. „start tag“ vs. „opening tag“, „schema“ vs. „schemata“ vs. „scheme“.
- Některé použité pojmy nejsou definovány/vysvětleny vůbec – např. „relvar“, „XPath axes“, „to inline an element“.
- Namísto „regular“ a „mixed“ XML dokumentů by bylo vhodné použít standardní pojmy „data-centric“ a „document-centric“. Pojem „mixed“ se používá ve smyslu smíšeného obsahu elementu, což je něco jiného. Namísto „schema-obvious“ by bylo vhodné použít klasický termín „generic“, namísto „schema-aware“ standardní „schema-driven“.
- „Database vendor“ znamená firma, která databázi prodává, nikoli součást databáze, která má určitou funkci.
- Ne všechny zmiňované technologie a pojmy jsou popsány v kapitole o použitých technologiích, popř. při prvním použití referencovány – např. jazyky XPath a XQuery, Edge mapping, B+ strom, R strom, Basic, Shared a Hybrid mapování.
- Mezi referencí a předchozím slovem se píše mezera.
- V textu by nemělo být používáno oslovení čtenáře – např. str. 34 „they offer you“, „you can use“.

Vzhledem k výše uvedeným připomínkám práce Tomáše Kohana podle mého názoru nesplňuje podmínky na diplomovou práci kladené. Práci lze považovat za velmi slibný začátek kvalitní diplomové práce, ovšem velké množství připomínek, je každopádně nutné zohlednit a především je nutné nápad zmíněný v práci patřičně rozvést, popsat, implementovat a otestovat, což bylo cílem práce. Proto ji k obhajobě v tomto stadiu rozhodně **nedoporučuji**.

V Praze, 24.8. 2007

[redacted]