

Similarity of XML Data

Autor diplomové práce se zabývá měřením podobnosti sady XML schémat popsanych pomocí jazyka DTD. Jazyk DTD byl zvolen především díky jeho jednoduchosti a rozšířenosti. Pro měření podobnosti autor kombinuje dva základní existující přístupy. Prvním přístupem je měření podobnosti elementů a atributů v XML dokumentech. Toto měření se skládá z měření editační vzdálenosti názvů a sémantické podobnosti názvů s využitím thesauru. Dále jsou srovnávány kardinality elementů. Druhým přístupem je měření editační vzdálenosti DTD stromů. Zde autor rozšiřuje existující editační operace vložení a odstranění daného podstromu, který existuje v obou porovnávaných XML schématech, do operací vložení a odstranění dvou podobných podstromů.

Autor v Kapitole 3 podrobně analyzuje některé existující přístupy v oblasti měření podobnosti XML dat a v Kapitole 4 zavádí vlastní přístup s odkazy do Kapitoly 3. Je tedy zřejmé, jaké existující metody autor využil a jak je rozšířil. Zvolená kombinace metod pro editační vzdálenost stromů a měření podobnosti na základě sémantiky je v této oblasti pravděpodobně originální. V Kapitole 6 potom ukazuje experimenty nad reálnými daty, které naznačují, že autor skutečně dosáhl vylepšení existujících metod.

Kromě drobnějších chyb v angličtině (především používání členů, občas nesrozumitelné větné konstrukce) lze autorovi vytknout především značnou stručnost textu. Práce by měla na začátku obsahovat rozsáhlejší motivační sekci s větším příkladem, na kterém by mohl autor ilustrovat jak existující metody, tak i metodu vlastní. Rovněž chybí více ilustrativní příklady pro vysvětlení editačních operací ve stromech popsanych v Kapitole 3.1.1 a jejich rozšíření v Kapitole 4.4. Příklady by byly zajímavé i pro popis algoritmů v Kapitole 4.5. Absence příkladů způsobuje, že je text práce místy velmi hutný. Např. vysvětlení algoritmu na Obrázku 4.8. není moc srozumitelné (trochu podivné větné konstrukce) a příklad by jej jistě ozřejmil.

I přes výše uvedené nedostatky lze ale konstatovat, že autor zadání práce splnil a proto **doporučuji práci uznat** jako diplomovou.

V Praze, 24.4.2008

