Oponentský posudek diplomové práce

Název DP: **Mapování a dolování ontolofií s uživatelskou preferencí**
Diplomant: **Dušan Maruščák**

Obsah práce:

Předmětem diplomové práce je dolování dat z webových stránek na základě zadané ontologie. V rámci práce byla vytvořena ontologie pro oblast notebooků (a v minimalistické verzi take aut) a systém pro dolování dat na základě této ontologie z webových stránek. Systém těží z podobnosti webových dokumentů reprezentující nabídku notebooků jednotlivých firem.

Úvodní kapitoly seznamují s pojmem ontologie, dolování ontologií a s jazyky pro popis ontologií založených na XML, zvláště pak OWL, který je použit v rámci aplikace, která vznikla jako výstup práce. V dalších kapitole je popsán algoritmus pro dolování ontologií, který je použit v rámci této aplikace. Další kapitola shrnuje výsledky aplikace navrženého algoritmu na reálných datech. Předposlední kapitola popisuje architekturu samotné aplikace a jako poslední je srovnání algoritmu aplikace a jejích výsledků s existujícími systémy pro extrakci ontologií. Jako samostatná sekce diplomové práce je pak uživatelská dokumentace s popisem instalace a nastavení aplikace. Tato sekce, stejně jako jádro diplomové práce, je součástí přiloženého CD, kde najdeme i samotnou aplikaci a vzorové ontologie (notebooky a auta).

Hodnocení:

Úvod popisující základní pojmy a jazyky pro práci s ontologiemi je napsaný velice kvalitně a matematicky korektně i s praktickými příklady na kterých jsou uvedené matematické definice ilustrovány a to relativně přehledně.

Při testování postupů pro extrakci dat byly zkoumány možnosti použití OCR nástrojů a možnosti extrakce dat na základě syntaktické analýzy HTML stránek. OCR nástroje se neukázaly jako schůdné řešení kvůli nedokonalosti rozpoznání fontů a své pomalosti. Osobně bych uvítal alespoň krátkou sekci popisující možnosti a návrhy spojení níže uvedené technologie s OCR nástroji (rozeznání bloků, kde se nacházejí nabídky a toto použít při extrakci z HTML, ...).

K algoritmu extrakce dat z HTML na základě ontologie, který je jádrem celé práce nemám připomínky a jeho prezentace v práci ja dobře strukturovaná a pochopitelná.

Hlavní testování bylo provedeno na doméně notebooků a testy podle přiložené tabulky dopadly velice dobře (z většiny stránek byla naprostá většina informací vydolována korektně). Dále pak byly provedeny testy na doméně automobilů, nicméně vzhledem k faktu, že byly extrahovány pouze 3 vlastnosti automobilů, nelze výsledky považovat za průkazné (nehledě na fakt, že všechny 3 vlastnosti nebyly správně extrahovány ani v polovině případů).

Srovnání s ostatními produkty postrádá jakákoli čísla, což v důsledku ústí ve formulaci "Věříme, že v porovnání s wrappery dosahujeme..." (důvody pro tuto doměnku jsou uvedeny a jsou pádné, nicméně chybí exaktní porovnání).

Velice kladně hodnotím architekturu samotné aplikace, která je rozdělena na klientskou a serverovou část, stejně jako volbu zavedení klientského rozhraní do webového prohlížeče (Mozilla Firefox). Mám ovšem výhrady k uživatelské přívětivosti aplikace, kdy některé chyby jsou viditelné pouze v serverové části programu, kterou uživatel teoreticky vůbec nemusí mít na svém počítači. Dále pak by stálo za úvahu zvážení použití komponenty na stavění DOM stromu z HTML dokumentu, kterážto část práce extrakce při mých testech stála nejvíce času a přitom naprostou většinu času vytěžuje procesor maximálně ze 2 procent (možná je to věc nastavení). S tím pak souvisí také vypisování hlášek o stavu uživatelovi, který v průběhu zpracování netuší, co se děje (nemá-li na svém počítači i serverouvou část).

Při testování se také projevila chyba, kdy na serverovou část je posílána pouze adresa načtené stránky a ne přímo stránka, což má za důsledek, že serverové proměnné vztahující se k aktuální session nejsou deklarovány a tudíž se může stat, že je zpracovávána jiná stránka, než jaká byla požadována (stalo se např. u Czech Computer). Také se stalo, že při druhém načtení byl výsledkem nekorektní výstup.

Klady:

1) Názorná analýza problému.
2) Navržení algoritmu pro extrakci dat na základě definovaných ontologií.
3) Korektní návrh a implementace architektury.
4) Uživatelské rozhraní jako plugin webového prohlížeče.

Zápory:

1) Syntaktické chyby v práci (slovosled, interpunkce).
2) Uživatelsky nepřívětivé ošetření chybových stavů.
3) Místy nejsou ošetřeny chyby vedoucí k nekorektním výsledkům.

Závěr:

Práce splnila zadání (za jádro práce lze jistě považovat navržený algoritmus, který je v pořádku). Práci doporučuji k obhajobě.

V Praze dne 7. září 2007