Posudek vedoucího na diplomovou práci
Dušana Maruščáka

Dolovanie a mapovanie ontológií

Předmětem práce je praktický příspěvek k dolování a mapování ontologií. Místo abstraktních teorií jsme se vydali praktickým směrem. Práce na konkrétních příkladech a vlastním software věnuje dolování ontologií (v mém chápání inovativním způsobem) přes extrakci dat z webu a obohacováním OWL ontologie (která na začátku celého procesu může obsahovat nějaké rudimentální zarodky ontologie, např klíčová slova z uživatelova dotazu). Mapovaní je prováděno identifikaci konceptů zdroje a (do těď vydolované) ontologie. Osobně si myslím že dobré dolování a mapování ani jinak nepůjde, než přes instance v reálných webovských zdrojích.

Po úvodní, teoretičtejší partii o Ontologiích se autor věnuje popisu jazyka OWL. Povšimnutí hodným příspěvkem je použití Stanfordem navrhované rozšíření jazyka OWL (není ještě standardizováno) o uživatelem definované datové typy. Diplomand ho šikovně vužil pro záměr, který autoři ani původně neměli v úmyslu (naštěstí Stanfordem vyvíjený systém Protégé ho již obsahuje). Vytvořil něco co jsme pracovně nazývali „extrakční ontologie", která kromě konceptualizace domény obsahuje i informatice o (možné) struktuře dat.

Hlavní kapitola práce obsahuje popis softwareového řešení. Jednou z prvních alternativ která došla do experimentální podoby bylo použití OCR systému. Hlavní myšlenkou bylo, že web stránky jsou v současnosti tvořené hlavně pro lidi a některé vizuální aspekty se dají odhalit jen OCR technologií. Později jsme ale tento krok (i když uchazeč do něj investoval nemalé usilí) zavrhli hlavně kvůli nepřístupnosti OCR systému s API, které by umožnilo automaticky extrahovat většímnožství stránek zabudováním do řešení. Po této odbočce autor iplementoval řešení založené na přespracovaném DOM modelu stránky, ve které se ve třech fázích najde najprve datová oblast, pak konkrétní záznam a nakonec (i s pomocí exttrakční ontologie) hodnoty jednotlivých atributů. Řešení bylo experimentálně ověřeno na větším množství stránek z domény prodeje notebooků a několika stránek z domény prodeje aut. Řešení i experimenty jsou předmětem publikace zaslané do recezního řízení.

Autor vytvořil originální Java software obohacující prostředí Mozilla Firefox (i díky zkušenostem z dílny L. Galambose pod kterého vedením pracoval v týmu softwareového projektu). Systém je vhodný na integraci do balíků software už v minulosti vybudovaném.

Po popisu software následuje srovnání z jinými systémy. Poloautomatické systémy dosahují větší přesnosti ale za předpokladu že je někdo natrénoval a fungují pak jen na

ten typ stránek (pro jisté aplikace, např. sledování pevně zvolených stránek které nemění strukturu ale zato obsah, je to nejlepší řešení). Tradeoff mezi lidskou intervencí a přesností pro sledování libovolných stránek (na kterých je větší množství opakujících se záznamů) se více hodi automatické systémy. Tady je náš systém podle našich znalostí lepší i ve funkčnosti i v přesnosti.

Touto prací autor obohatil myšlenku semantického webu. Ten si totiž za hlavní cíl klade vytvořit z webu univerzální prostředí pro výměnu dat strojově zpracovatelnych (i software vyvinutým nezávisle).

Schází mi tady experimenty, které by popisovali ladění některých parametrů vyhladávání v závoslosti na domeně.

Závěr: Práce je vysoce kvalitní a obsahuje i publikovatelné výsledky. Práci doporučuji k obhajobě.

V Praze dne 1. 9. 2007

Prof. RNDr. P. Vojtáš, DrSc.
KSI MFF UK