Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Rastislav Kadleček		
Název práce	Converting HTML product data to Linked Data		
Rok odevzdání	2018		
Studijní program	Informatika	**Studijní obor**	Softwarové systémy

Autor posudku	RNDr. Martin Svoboda, Ph.D.	**Role**	Oponent
Pracoviště	KSI		

Text posudku:

Cílem hodnocené diplomové práce byl návrh rozšíření existujícího frameworku Odelic (softwarový projekt na MFF), na základě kterého by bylo možné vylepšit semi-automatickou transformaci dat získaných z několika vybraných webových obchodů do RDF trojic, a to na základě principů Linked Data. Zadání práce bylo převážně splněno, některé části jako např. detekce tříd produktů v závislosti na detekovaných predikátech však nikoli.

Autor práce nejprve extrahoval vybraná data (názvy a další vlastnosti) o produktech konkrétně mobilních telefonů, tabletů, noteboků a stolních počítačů z webových stránek (HTML souborů) několika konkrétních obchodů a transformoval je do relačního modelu (CSV souborů). Na základě provedeného experimentálního srovnání několika existujících metod založených na principech strojového učení pak vybral a použil konkrétní klasifikátor, pomocí kterého je schopen detekovat význam (autorem zavedené ML třídy) jednotlivých hodnot a celých sloupců. Ty jsou následně uživatelem fixně mapovány na standardní ontologické třídy a vlastnosti.

Celkový rozsah práce je adekvátní, obsahuje všechny očekávané části. Pravdou však je, že poměrně velký prostor je věnován nejrůznějším přílohám a následně popisu existujících řešení, takže nejzajímavější část práce týkající se popisu vlastního navrženého řešení a jeho detailů (kapitola 7) odpovídá jen řádově pětině rozsahu. Text práce je psán relativně dobrou angličtinou, avšak s poměrně početnými chybami. Po stylistické stránce je však text kvalitní a převážně dobře formulovaný a plynulý. Počet citovaných odborných zdrojů je nadprůměrný, stejně jako počet zvažovaných existujících přístupů a technologií.

Zpracovávané téma není po formální stránce složité, většina používaných nebo navržených přístupů je relativně přímočará. Určité části textu jsou bohužel poměrně obtížně uchopitelné čtenáři, kteří s danou problematikou ještě nejsou detailně seznámeni. Práci by proto výrazně pomohlo, pokud by byly nejprve definovány základní pojmy a s těmi se následně pracovalo důsledně a konzistentně. Naopak některé uvedené definice (kapitola 1) jsou jen ilustračními příklady, nikoli skutečnými formálními definicemi. Textu by rovněž pomohlo, pokud by byl oddělen popis navrženého řešení (principy, model, algoritmy či vlastnosti) od popisu implementačních detailů, integrace do již zmíněného frameworku či uživatelského rozhraní. Nejzajímavějším částem práce mělo být věnováno více prostoru a měly být popsány detailněji.

Přestože autor navrhl vylepšení několika konkrétních součástí frameworku, hlavní přínos spočívá ve využívání již zmíněných klasifikátorů namísto dohledávání sémantických

informací ze zdrojů jako DBpedia apod. Ty se totiž ve zvolené doméně ukazují jako nepoužitelné, protože nové produkty vznikají rychleji než záznamy o nich v takový zdrojích. Zadání práce je tak dobře motivováno, avšak skutečné a hlubší přínosy navrženého řešení jsou minimálně diskutabilní.

Vybraný klasifikátor se totiž neobejde bez datové sady na učení, a tedy uživatel v konečném důsledku stejně většinu sémantické informace musí zadat zcela manuálně (jiné pořadí nebo názvy sloupců jsou jen technickou záležitostí). Navíc se pracuje s nerealistickými předpoklady, že každý sloupec může být asociován výhradně jen s jednou třídou, naopak každá třída nejvýše s jedním sloupcem. Na druhou stranu je potřeba vyzdvihnout skutečnost, že se autor musel seznámit s celou paletou nejrůznějších existujících technologií a přístupů, stejně jako integrovat vlastní vylepšení do cizího již existujícího softwarového řešení.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 27. srpna 2018 **Podpis**