Posudek *vedoucího* diplomové práce

Jméno a příjmení autora posudku: Peter Vojtáš

Jméno a příjmení autora práce: Jakub Kýpeť

Název práce: Semantic annotation and querying RDF data

Práce má za cíl usnadnit (automatizovat) zpracování obsahu webu a to extrakcí a anotací obsahu web stránek bez asistence tvůrců stránek. Měla být „server" verzí DP práce D. Fišera (který naprogramoval a testoval „client" - plugin do prohlížeče).

Předem musím konstatovat, že jsem práci před odevzdáním neviděl. Práci nedoporučuji hlavně kvůli nedostatečné kvalitě textové části (i když mám výtky také k provedení a popisu experimentů). Celé řešení je většinou slovně popsané v ne-příliš kvalitní angličtině, formální úroveň práce je velmi nízká.

Terminologický zmatek: Item/objekt má atributy a ty hodnoty – na webu to může byt spojeno s tagy a labely – nebo jsou to typy? Str. 11 – prosím použít standardní terminologii DBMS a pokud se používá jiná tak prosím uvést odkud (html, …nebo vlastní terminologie a / nebo metoda vlastní aplikace? …). Prosím popsat, co jsou „labely" a k čemu slouží. Někdy se míchá terminologie např. template – web shop-u – versus moje schema? (jak je to pro single web shop a pro celou skupinu – 14_3?).

Nedostatečný popis algoritmů: Autor uvádí triviální algoritmus pro výpočet Levenshteinovy vzdálenosti a neuvádí žádný svůj, prosím podrobně a formálně popsat i s uvedením příkladů alespoň klíčové části řešení:

-algoritmus pro 19^11 „our first task", dále 19_11„use labels, that we have described earlier in this section" – žádné nebyly popsány, jak se určí Threshold value?

-pro 20^^8 pro skupiny a vzdálenosti a opět odkud je interval pro práh (testovat přesnost v závislosti na prahu / pro různé domény asi budou různé prahy, viz DP Maruščák).

-20_19 – vezmeme největší?(v jakém smyslu) nebo nejbližší skupinu a proč?

-algoritmus/y pro „getting values", která je klíčovou částí implementace.

Celkově schází ilustrativní příklady. Popis pravidel úplně schází.

Experimenty: Pro styl práce je příznačné, že experimenty nezmiňují metody aplikace, jsou popsány jen v přirozeném jazyce a ani obsah tabulek není dostatečně popsán (např. trénovací data při ruční anotaci). Z předešlého mi není jasné, v čem přesně je rozdíl mezi „using template" (bez „value classification") a "using labels and value classification" a nakonec použití všech dohromady. „Using" poukazuje na použité (lidmi vytvořené) znalosti? Který algoritmus/metoda byl-a použit-a není popsáno.

Aplikace funguje pouze lokálně na diplomandově notebooku, server instalace schází, není možno ověřit dotazování nad získanými daty.

Doporučuji podrobně přepracovat vzhledem k těmto připomínkám a KSI pravidlům.

PS. Schází citace na indago crawler.

Doporučení k obhajobě: Z výše uvedených důvodů práci *nedoporučuji* k obhajobě.

Vynikající práce vhodná pro soutěž studentských prací	**NE**

V Praze dne 9. 1. 2015

Podpis: