Posudek oponenta závěrečné práce

České vysoké učení technické v Praze **Fakulta informačních technologií**

Student: Bc. Jana Čabaiová
Oponent práce: RNDr. Jakub Klímek, Ph.D.
Název práce: Summarizing Linked Open Data Datasets
Obor: Webové a softwarové inženýrství

Datum vytvoření: 16. 1. 2017

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
1. Náročnost a další komentář k zadání	*1=mimořádně náročné zadání,* *2=náročnější zadání,* ***3=průměrně náročné zadání,*** *4=lehčí, ale ještě dostatečně náročné zadání,* *5=nedostatečně náročné zadání*

Popis kritéria:
Podrobněji charakterizujte diplomovou (bakalářskou) práci a její případné návaznosti na předchozí nebo běžící projekty. Dále posuďte, čím je zadání této ZP náročné. (U obtížnější ZP lze dále tolerovat některé nedostatky, které by u ZP standardní obtížnosti tolerovány nebyly; a naopak u jednoduché ZP mohou být zjištěné nedostatky hodnoceny přísněji.)

Komentář:
Zadání hodnotím jako lehčí až průměrně náročné, přičemž záleží na jeho interpretaci, zejména bod 3) zadání implementace se dá splnit jednoduše (jako v této práci) nebo složitě. V zadání je pak překlep, místo HDF má být HDT.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
2. Splnění zadání	*1=zadání splněno,* *2=zadání splněno s menšími výhradami,* ***3=zadání splněno s většími výhradami,*** *4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP splňuje zadání. V komentáři uveďte body zadání, které nebyly zcela splněny, případně rozšíření ZP oproti původnímu zadání. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Komentář:
Část zadání "micro visualization - completeness of information for a given entity type" splněna nebyla, navíc o termínech macro a micro visualization v práci není jediná zmínka. Dále zadání zahrnuje vývoj metody (předpokládám že nikoliv v terminologii OOP) pro sumarizaci datasetů, což je v textu odbyto tím, že je to nějakých 5 SPARQL dotazů, které ale v práci nejsou ani napsány, natož aby zde byla diskuze k tomu, jak byly vybrány atd.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
3. Rozsah písemné zprávy	*1=splňuje požadavky,* ***2=splňuje požadavky s menšími výhradami,*** *3=splňuje požadavky s většími výhradami,* *4=nesplňuje požadavky*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části.

Komentář:
152 stran působí jako úctyhodné číslo, nicméně samotná práce začíná na straně 19 a závěr je na straně 120, na samotnou práci tak připadá přiměřených cca 100 stran.
První třetina práce se věnuje přehledu technologií sémantického webu, přehled existujících podobných nástrojů by ale mohl být komplexnější, toto je v současnosti žhavé téma a nástrojů je více. Zbytek práce se věnuje standardnímu softwarově inženýrskému popisu analýzy, návrhu, implementace a testování standardní webové aplikace. Co mi zde chybí je hlubší návaznost na technologie sémantického webu a na to, co je pro tuto konkrétní aplikaci specifické.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*40 (F)*

Popis kritéria:
Posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře.

Komentář:
Díky tomu, že práce obsahuje velmi velké množství různých překlepů a zároveň je úroveň angličtiny nízká, někdy není jasné, zda nalezené nedostatky patří mezi věcné nebo formální, tedy zda jde o nedostatečné porozumění tématu, uspěchanost, nebo o nedostatečnou schopnost se přesně vyjadřovat. Pár příkladů:
- Strana 14 dole: "RDF je framework, jehož hlavní funkcí je reprezentace Linked Data" - nemůže být pravdivé už jen kvůli tomu, že RDF existuje déle než principy Linked Data
- Tabulky 1.1 a 1.2 mají stejné záhlaví, u tabulky 1.1 je to chybné, místo Web of Data by to mělo být Web 2.0. Termín Web of Data je ale pro dané téma klíčový
- Figure 1.1 - Věcné chyby. Jendak v pravé, RDF části, by se serverem měla komunikovat aplikace, nikoliv Web browser jako v levé části, jednak IRI na které se přistupuje přes HTTP je v obrázku http://dbpedia.org/page/Roger_Federer, což je IRI stránky o Rogerovi Federerovi, nikoliv IRI Rogera Federera, jak je uvedeno v popisku a které je ve skutečnosti http://dbpedia.org/resource/Roger_Federer
- Pojmosloví je nekonzistentní, stejným věcem se říká několika názvy (vocabularies, dictionaries), různé věci se naopak skrývají za stejným pojmem
- Datový model RDF v sekci 1.4.1.1 je popsán nesystematicky, se zbytečným důrazem na nedůležité věci (blank nodes) a s nedostatečným důrazem na důležité věci. Jednou z vlastností RDF je že je tzv. "schema-less", ale hned v prvním odstavci 1.4.1 se mluví o různících se schématech, bez dalšího vysvětlení.
- IRI jsou case-sensitive, ale v práci toto není dodržováno, například v odstavci na straně 19 (DCAT, VoID and DCMI) dcat:Keyword by naznačovalo třídu, existuje ale (a byla myšlena) vlastnost dcat:keyword. Stejně se dále v práci špatně vyskytuje owl:SameAs (owl:sameAs, strana 39).
- DCMI je vysvětleno jako slovník a rozepsáno jako Dublin Core Metadata Information, přitom jde o organizaci, jmenuje se Dublin Core Metadata Initiative a slovník je jejím výtvorem.

Toto je výčet příkladů do strany 19, podobných nepřesností je ale v práci mnoho. Každý zvlášť by byl snadno přehlédnutelný, ale jejich množství je zarážející.
Značná část pojmů zavedených v sekci 1 je poté použita pouze povrchově.

Omezení na vstup v serializaci N-Triples nebo HDT nedává smysl už jen proto, že je použit framework Apache Jena, který obsahuje potřebné implementace i ostatních serializací, a není tedy jasné, proč není umožněno je použít.

Tabulky 7.1 a 7.2 v experimentech na konci obsahují nulové řádky, což je zajímavé, možná nečekané, ale nikde k tomu není žádná diskuze. Není uvedeno, v jakém prostředí byly experimenty prováděny a jak dlouho trvaly.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*30 (F)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 12/2014, článek 3.

Komentář:
Jazyková úroveň, zejména v úvodních kapitolách práce, je velmi nízká a práce se pak těžko čte. Kromě špatné angličtiny (měla by proběhnout korektura, například rodilým mluvčím) je zde velké množství překlepů (odhalitelných spell-checkerem). Občas jsou chyby až úsměvné, například "lo-fi a wi-fi" místo "lo-fi a hi-fi" (4.5).
Řada obrázků obsahujících diagramy a text by zasloužila vektorovou podobu místo bitmapové (například v úvodu Figure 1.1, 1.2, 1.3, 1.4., 1.5, 1.6).
Někdy se o práci mluví jako o "paper" (například v závěru nebo v obsahu CD), což vyvolává otázku, zda byly výsledky někde publikovány jako článek a některé části byly kopírovány, nebo zda jde jen o nesprávný termín.
Nekonzistentní je například i výskyt některých zkratek, jednou SPARQL, jednou sparql apod.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*60 (D)*

Popis kritéria:
Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení ZP. Charakterizujte výběr studijních pramenů. Posuďte, zda student využil všechny relevantní zdroje nebo zda se pokoušel řešit již vyřešené problémy. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Komentář:
Některé zdroje by mohly být více autoritativní, například místo reference [2] - slidy k předmětu MI-SWE by bylo vhodnější citovat přímo specifikace, publikace, originální zdroje na webu apod. U některých referencí není jasný jejich rozsah, například v sekci 1.2 je celý výčet 5 bodů převzatý z webu [5], daná reference je ale uvedena na konci páté položky.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*60 (D)*

Popis kritéria:
Vyjádřete se k úrovni dosažených hlavních výsledků ZP, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, apod. Případně také zhodnoťte, zda software nebo zdrojové texty, které nevytvořil sám student, byly v ZP použity v souladu s licenčními podmínkami a autorským právem. Popište případnou publikační činnost a získaná ocenění související s řešením této ZP.

Komentář:
Práce popisuje implementaci webové aplikace počítající statistiky nad RDF daty. Novou teorii neobsahuje. Výsledná aplikace je více méně funkční, v době testování nefugoval jen export výsledků do souboru a zobrazování výsledků větších analýz trvalo nepřiměřeně dlouho.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
8. Komentář o využitelnosti výsledků	

Popis kritéria:
Uveďte, zda hlavní výsledky ZP rozšiřují již publikované známé výsledky a/nebo přinášející zcela nové poznatky. Uveďte možnosti využití výsledků ZP v praxi.

Komentář:
Vzhledem k nutnosti ruční definice domén, což jsou vlastně pojmenované skupiny RDF tříd, je praktická využitelnost nástroje diskutabilní. Otázkou totiž je, k čemu je využitelná informace, že daná datová sada obsahuje x% entit z domény Meteorologie a y% entit z domény Jídlo a pití. Pokud jsem doménu nedefinoval, nebo nerozumím RDF, tak nevím co konkrétně toto pro mě znamená. Pokud jsem doménu musel definovat a tedy RDF rozumím, pak už je pro mě zajímavější podíl instancí konkrétních tříd, což ale zjistím jedním SPARQL dotazem i bez aplikace. Navíc na běžném notebooku už vzorová analýza DBpedie se načítá pomalu, řádově minuty, což by měl být jeden z řešených problémů v diplomové práci.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
9. Otázky k obhajobě	

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:
- Na jakém HW (nebo konfiguraci virtuálního stroje) byly provedeny experimenty?
- Jak byly vybrány použité SPARQL dotazy?

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*40 (F)*

Popis kritéria:
Shrňte stránky ZP studenta, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení **nemusí** být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích 1 až 9.

Text hodnocení:
Práce sice obsahuje všechny očekávané části, ale působí VELMI uspěchaným a nedotaženým dojmem. Obsahuje obrovské množství překlepů, formálních i věcných nepřesností, nedotažených popisů a používá nekonzistentní pojmosloví. Jazyková úroveň je taktéž nízká. Výsledná aplikace rovněž obsahuje překlepy. Celkově tedy práce působí dojmem, že je ve stádiu návrhu a je třeba ji ještě měsíc či dva dodělávat, aby byla v pořádku. Většina nedostatků jsou drobnosti, nicméně jejich množství způsobuje, že mnohdy není jasné, zda jde ještě o sadu překlepů, nebo skutečně o jiný význam věty a celkový dojem z práce je pak negativní.

Samotná aplikace bez části věnující se počítání statistik nad RDF datasety je na úrovni standardní bakalářské práce. Bohužel ta stěžejní část (dle zadání), která by ji posunula na úroveň práce diplomové, je sice implementována, ale není dostatečně popsána - jedná se o sadu SPARQL dotazů, které v práci nejsou popsány - a je vyřešena celkem jednoduše - sada použitých dotazů je triviální [1] a téma by šlo zpracovat do větší hloubky odpovídající diplomové práci. V experimentech mi navíc chybí údaj, na jak výkoném stroji byly provedeny, což je vzhledem k velikosti datasetů, což je nejběžnější problém, klíčové.

[1]
https://github.com/jcabaiova/RDFDataAnalyser/blob/master/src/main/java/cz/cvut/fit/cabaijan/calculationAnalysis/Queries.java

Podpis oponenta práce: