Наставно-научном већу
Математичког факултета
Универзитета у Београду

Одлуком Наставно-научног већа Математичког факултета Универзитета у Београду донетом на 376. седници одржаној 20.11.2020. године именовани смо за чланове комисије за оцену докторске дисертације **„Моделовање упитних језика са применама у рефакторисању и оптимизацији кода”** кандидата Мирка Спасића, дипломираног математичара. После прегледа поднетог рукописа подносимо следећи

Извештај

Биографски подаци

Мирко Д. Спасић је рођен у Брусу, 01.03.1985. године, где је завршио основну школу и гимназију. Математички факултет у Београду, смер Рачунарство и информатика, завршио је 2009. године са просечном оценом 10,00. На истом факултету уписује и докторске студије, у оквиру којих је положио све испите предвиђене планом и програмом студија са максималним оценама.

Запослен је на Катедри за рачунарство и информатику Математичног факултета Универзитета у Београду у звању сарадника у настави од 2009. године, а у звању асистента од 2011. године. До сада је изводио вежбе из низа предмета на основним и мастер студијама:

- Програмирање 1 и 2,
- Лексичка анализа са применама,
- Компилација програмских језика,
- Превођење програмских језика,
- Конструкција компилатора,
- Анализа и дизајн алгоритама,
- Конструкција и анализа алгоритама 1 и 2.

У периоду од 2011. до 2013. године, ангажован је као спољни сарадник у Институту Михајло Пупин, а после тога у компанији Openlink Software (Велика Британија).

Област научног интересовања му је верификација софтвера, упитни језици и повезани подаци. Радио је на многобројним европским истраживачким пројектима управо у тим областима (LOD2, Cesar, GeoKnow, LDBC, Hobbit и Sage). Учесник је пројекта „Српски језик и његови ресурси” Министарства просвете, науке и технолошког развоја Републике Србије.

У току студија био је стипендиста Републичке фондације за развој научног и уметничког подмлатка, као и компаније „Дунав осигурање”. У току докторских студија, добио је главне награде за најефикаснији систем за управљање повезаним подацима на такмичењима MOCHA2017 и MOCHA2018 у оквиру конференција ESWC2017 (Хераклион, Грчка) и ESWC2018 (Порторож, Словенија), и похађао је Летњу школу повезаних отворених података ISSLOD2011 у Лајпцигу, Немачка.

Објављени научни радови и саопштења са научних скупова повезана са дисертацијом

Мирко Спасић има девет радова и саопштења са научних скупова повезаних са дисертацијом, од којих је један у часопису са SCI листе (M22), један је самостални, а у шест радова је Мирко Спасић први аутор:

1. M. Spasić, M. Vujošević Janičić. “Verification supported refactoring of embedded SQL”. Software Quality Journal (2020), pp. 1–37. IF 2.169 (M22), doi: 10.1007/s11219-020-09517-y.

2. M. Spasić, M. Vujošević Janičić. “SpeCS — SPARQL Query Containment Solver.” In: 2020 Zooming Innovation in Consumer Technologies Conference (ZINC). IEEE, 2020, doi: 10.1109/ZINC50678.2020.9161435.

3. Mirko Spasić. „Životni ciklus povezanih podataka kroz razvoj skupa podataka vozila u pokretu”. InfoM - Časopis za informacione tehnologije i multimedijalne sisteme, Broj 71 (2020), pp. 44-55. issn: 1451-4397.

4. M. Spasić, M. Vujošević Janičić, “First Steps towards Proving Functional Equivalence of Embedded SQL,” in 24th International Conference on Types for Proofs and Programs, TYPES2018, J. E. Santo and L. Pinto, Eds., Braga, Portugal: Centro de Matemática, University of Minho, Jun. 2018.

5. M. Spasić, M. Jovanovik. “MOCHA 2017 as a Challenge for Virtuoso.” In: Semantic Web Challenges - 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28 - June 1, 2017, Revised Selected Papers. Ed. by M. Dragoni, M. Solanki, and E. Blomqvist. Vol. 769. Communications in Computer and Information Science. Springer, 2017, pp. 21–32. doi: 10.1007/978-3-319-69146-6_3.

6. M. Spasić, M. Jovanovik, A. Prat-Pérez. “An RDF Dataset Generator for the Social Network Benchmark with Real-World Coherence.” In: Proceedings of the Workshop on Benchmarking Linked Data (BLINK 2016) co-located with the 15th International Semantic Web Conference (ISWC), Kobe, Japan, October 18, 2016. Ed. by I. Fundulaki, A. Krithara, A.-C. N. Ngomo, and V. Rentoumi. Vol. 1700. CEUR Workshop Proceedings. Oct. 2016. doi: 10.13140/RG.2.2.30490.64965/1.

7. R. Angles, J. B. Antal, A. Averbuch, P. Boncz, O. Erling, A. Gubichev, V. Haprian, M. Kaufmann, J. L. L. Pey, N. Martínez, J. Marton, M. Paradies, M.-D. Pham, A. Prat-Pérez, M. Spasić, B. A. Steer, G. Szárnyas, J. Waudby. The LDBC Social Network Benchmark. 2020. arXiv: 2001.02299 [cs.DB].

8. K. Georgala, M. Spasić, M. Jovanovik, V. Papakonstantinou, C. Stadler, M. Röder, A.-C. N. Ngomo. “MOCHA2018: The Mighty Storage Challenge at ESWC 2018.” In: Semantic Web Challenges. Ed. by D. Buscaldi, A. Gangemi, and D. Reforgiato Recupero. Springer, 2018, pp. 3–16. isbn: 978-3-030-00072-1.

9. M. Jovanovik, M. Spasić. “Benchmarking Virtuoso 8 at the Mighty Storage Challenge 2018: Challenge Results.” In: Semantic Web Challenges - 5th SemWebEval Challenge at ESWC 2018, Heraklion, Greece, June 3-7, 2018, Revised Selected Papers. Ed. by D. Buscaldi, A. Gangemi, and D. R. Recupero. Vol. 927. Communications in Computer and Information Science. Springer, 2018, pp. 24–35. doi: 10.1007/978-3-030-00072-1_3.

Предмет дисертације

Предмет истраживања ове докторске дисертације је моделовање упитних језика, тј. њихових основних конструката са циљем унапређења процеса рефакторисања и оптимизације кода који користи податке похрањене у бази података.

За приступ традиционалним релационим базама података, као упитни језик, стандардно се користи SQL, док су у данашње време све више заступљени и други модели података, као што су повезани подаци (енгл. Linked Data) у RDF формату, и SPARQL, као стандардни упитни језик у овом свету, па су управо ова два језика предмет истраживања ове дисертације.

Унапређивање квалитета кода без промене његове функционалности, у смислу рефакторисања или оптимизације, спада у свакодневну програмерску активност. Генерално прихваћена пракса у овој области подразумева да такве промене буду пропраћене провером да ли је очувано жељено понашање програма. Ова провера најчешће се врши тестирањем. Међутим, у таквом сценарију, то може трајати дуго и притом, тестирање не гарантује одсуство разлика у понашању између две верзије кода. С друге стране, коришћење алата за аутоматску верификацију еквивалентности кода може да пружи гаранцију задржавања понашања програма при рефакторисању или оптимизацији и као такво од суштинског је значаја за ефикасан и поуздан развој софтвера.

Проблем одређивања еквиваленције програма је NP-комплетан, и тренутно реална примена алата који га решавају у индустријском окружењу је ограничена. Разлози су ефикасност и једноставност коришћења алата, и пре свега њихова поузданост. Како би њихово коришћење у потпуности заменило тестирање у развојном процесу, алати морају бити сагласни и потпуни, што је врло јак захтев. У супротном, могу се користити као допуна процесу тестирања, пружајући већи степен поузданости и одсуство грешки.

У циљу развоја алата за аутоматску верификацију еквивалентности кода у апликацијама које користе упитне језике, потребно је прецизно моделовати семантику ових програмских језика. Предмет ове докторске дисертације је моделовање основних конструката упитних језика у теоријама логике првог реда, захваљујући ком је, употребом одговарајућих решавача, могуће доношење корисних закључака о упитима. У циљу рефакторисања или оптимизације, у складу са задатом семантиком, подржано је утврђивање еквивалентности и других односа између два упита.

Један део дисертације посвећен је проблему садржаности упита (енгл. query containment problem) у језику SPARQL. Овај проблем, али за упите над релационим базама, постављен је још 1977. године као проблем утврђивања да ли су сви резултати једног упита садржани у скупу резултата другог упита над било којом базом података. Ово је један од фундаменталних проблема у области база података и у општем случају је неодлучив. За неке класе упита овај проблем је NP-комплетан и представља велики изазов за истраживаче. Опис проблема природно се преноси на језик SPARQL, који је релативно нов, али већ има истраживача који се баве овим проблемом баш у овом језику. Са аспекта глобалне оптимизације у оквиру статичке анализе упита ово је најбитнији проблем, док се други значајни проблеми као што су еквивалентност упита и задовољивост упита, могу свести на проблем садржаности упита. Два упита су еквивалентна ако имају једнаке скупове одговора над било којом базом, што се своди на проверу задовољења релације садржаности упита у оба смера. Ако су два упита еквивалентна, оптимизатор упита може извршити један упит уместо другог, у случајевима када такав избор води ефикаснијем извршавању. Упит је задовољив ако постоји бар једно његово решење, па се проблем задовољивости може решити свођењем на утврђивање еквиваленције (или садржаности) упита са неким упитом за који је познато да нема решења. Ако упит није задовољив, и ако се незадовољивност може закључити статички, без његове евалуације, то може уштедети значајно време, посебно у дистрибуираним окружењима.

Како се упитни језици користе и у комбинацији са неким програмским језиком опште намене, резоновање о таквим програмима постаје далеко компликованије, јер укључује семантике и императивних и декларативних језика. Примена моделовања упитних језика је могућа и у овим сценаријима, и други део дисертације бави се управо

тиме, на примеру SQL језика уграђеног у језик C/C++ (енгл. embedded SQL). Оваквим приступом омогућује се аутоматско резоновање о функционалној еквивалентности C/C++ програма који садрже уграђене SQL упите за приступ подацима, у циљу једноставнијег рефакторисања и оптимизације такве врсте кода.

Приказ дисертације

Основни део рукописа дисертације садржи 156 страница које обухватају 5 поглавља, као и списак коришћене литературе од 206 библиографских јединица. У оквиру рукописа постоји 5 табела и 49 слика. Структура рукописа је следећа:

У уводном поглављу дат је шири контекст проблема који се решава у дисертацији, објашњен је његов значај и примене. Објашњена су два нивоа на којима се могу појавити оптимизације, ниво система за управљање базама података и ниво апликација, и за сваки од њих постоји посебно потпоглавље. У последњем потпоглављу, приказана је организација тезе и издвојени су главни доприноси које она доноси.

У Глави 2 дат је кратак преглед релевантних информација и општих појмова који се користе у наставку тезе, и приказује сродне приступе. Објашњена је функција аутоматског доказивања теорема у служби анализе кода, представљени су упитни језици који се користе у тези, SQL и SPARQL, и њихове семантике. Дефинисан је проблем садржаности упита у ова два језика, и дат је преглед алата који их решавају. Такође, глава даје преглед аутоматске анализе кода, императивног, упитног и њихове комбинације, као и еквивалентности програма, тј. регресионе верификације.

У Глави 3 формално је представљена синтакса и семантика језика SPARQL и дефинише проблеме садржаности и задовољивости упита у овом језику. Представљено је моделовање великог броја језичких конструката и моделовање проблема садржаности упита у теоријама логике првог реда. Овим је омогућено ефикасно аутоматско резоновање о овом проблему коришћењем одговарајућих SMT (енгл. Satisfiability Modulo Theory) решавача. Подржани су конјунктивни упити, неименовани чворови, пројектоване променљиве, оператори union, filter, optional, graph, подупити, изрази у select клаузули, уграђене функције, циклични упити и преименовање пројекција. Доказана је сагласност и потпуност предложеног моделовања и за релацију садржаности упита у својој стандардној форми и за релацију стапања, која је често коришћена форма релације садржаности. Подржано је разматрање релације садржаности узимајући у обзир RDF схему. На крају главе, приказани су и резултати евалуације имплементираног алата SpeCS у поређењу са савременим решавачима (SPARQL-Algebra, AFMU, TreeSolver и JSAG) на већ постојећим скуповима примера за тестирање (Query Containment Benchmark и SQCFramework). Евалуација је показала да је SpeCS ефикасан, бржи од осталих решавача истог проблема, уз бољу покривеност језичких констурката. Такође, SpeCS је открио више недостатака у постојећим скуповима примера за тестирање, како типографских грешака тако и суштинских проблема, који су адекватно отклоњени, и на тај начин је формиран нови поузданији скуп примера.

У Глави 4 представљена је још једна примена моделовања упитних језика у теоријама логике првог реда. Повезују се семантике императивних и упитних језика, и на тај начин омогућава аутоматско резоновање о функционалној еквивалентности императивних програма који садрже приступ релационим базама података кроз уграђене SQL упите. На тај начин, унапређује се процес рефакторисања оваквог типа кода, пружајући додатну сигурност програмерима при овој свакодневној активности. Апликације за приступ базама података имају специфичне могућности за рефакторисање, јер обично укључују бар два различита програмска језика: упитни језик (најчешће SQL) и неки програмски језик опште намене, тј. матични језик. Унапређење кода у апликацијама за приступ базама података може подразумевати рефакторисање SQL упита, рефакторисање наредби матичног језика, и рефакторисање интеракције између њих, тј. симултане измене у оба дела кода које чувају глобалну еквивалентност (без чувања еквиваленције ових двају делова посматраних одвојено). Док постоје бројна истраживања

на тему провере еквивалентности два SQL упита и провере еквивалентности императивних програма, провера еквиваленције која покрива интеракцију између SQL упита и матичног језика до сада је скоро непокривена у литератури, иако нуди значајне могућности. У поглављу дат је један предвиђени случај коришћења имплементираног алата SQLav, описано је моделовање програма са уграђеним SQL кодом, и представљена је конструкција услова еквивалентности функција. У последњем потпоглављу, дате су информације о имплементацији и евалуацији присупа на специјално конструисаном корпусу примера за ову сврху. Такође, описана је и интеграција алата у GitHub акције, којом се олакшава свакодневна употреба алата.

У Глави 5 изложени су главни теоријски и практични закључци тезе, и представљени су могући правци даљих истраживања који би се ослањали на резултате приказане у овој дисертацији.

Научни допринос дисертације

- Оригинално моделовање основних конструката упитних језика SQL и SPARQL у теоријама логике првог реда, које се користи у различитим сценаријима, и то при:
    - моделовању проблема садржаности, еквиваленције и задовољивости упита у језику SPARQL, за које је доказана сагласност и потпуност, и које се може користити за проверу еквивалентности упита насталих презаписивањем у оквиру оптимизатора упита,
    - моделовању проблема функционалне еквивалентности сегмената императивног кода који садржи приступ релационим базама података користећи уграђене SQL упите, које се може користити за утврђивање еквивалентности оригиналног и рефакторисаног кода.
- Имплементација јавно доступних алата отвореног кода за резоновање о овако моделованим проблемима:
    - алат SPECS, сагласан и потпун решавач проблема садржаности упита у језику SPARQL, надмашује остале савремене приступе у решавању овог проблема, како у погледу ефикасности, тако и са аспекта подржаности различитих језичких конструката,
    - алат SQLAV, користан у процесу рефакторисања императивног кода са уграђеним SQL упитама, проверава еквиваленцију две верзије кода или указује на могуће пропусте и грешке.
- Унапређивање постојећих корпуса за проверу коректности и развијање нових, који су расположиви и корисни свим истраживачима који се баве овом проблематиком:
    - откривање и отклањање проблема у постојећим скуповима тестова садржаности SPARQL упита,
    - корпус са више стотина примера рефакторисања C/C++ функција са уграђеним SQL упитима, изграђен за потребе евалуације развијеног алата SQLAV, и у будућности сличних алата.

Закључак

У рукопису „Моделовање упитних језика са применама у рефакторисању и оптимизацији кода” кандидат Мирко Спасић показао је широко и систематично познавање области семантика програмских језика, упитних језика, аутоматског доказивања теорема и моделовања проблема. Повезавши знања из ових области, кандидат је успео да да значајан допринос решавању проблема садржаности упита у новом и модернијем контексту језика SPARQL. Овај проблем је један од фундаменталних проблема у базама података, дефинисан још 1977. године. Проблем је NP-комплетан и веома важан са теоријског и са практичног аспекта. Решаван је у последњих десетак година на различите начине од стране више истраживачких група. Предложени приступ је у поређењу са релевантним савременим решењима напреднији и бољи, како са аспекта ефикасности, тако и у смислу поузданости, јер је математички доказана његова сагласност и потпуност. У тези је приказана још једна битна примена оваквог моделовања, практично применљива у свакодневним програмерским активностима рефакторисања кода.

Кандидат Мирко Спасић је кроз научно-истраживачки рад приказан у овој дисертацији дао значајан допринос решавању битних проблема у области база података, како теоријских, тако и у домену примена. Има девет радова и саопштења са научних скупова повезаних са дисертацијом, од којих је један у часопису са SCI листе (M22), а један је самостални. Стога, са задовољством предлажемо Наставно-научном већу Математичког факултета да рукопис „Моделовање упитних језика са применама у рефакторисању и оптимизацији кода” кандидата Мирка Спасића прихвати као докторску дисертацију и одреди комисију за јавну одбрану.

Београд, 10. децембар 2020.

Чланови комисије:

_____

др Ненад Митић, редовни професор,

Математички факултет, Универзитет у Београду

_____

др Филип Марић, ванредни професор,

Математички факултет, Универзитет у Београду

_____

др Силвиа Гилезан, редовни професор,

Факултет техничких наука, Универзитет у Новом Саду