

ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ

ИЗВЕШТАЈ О ОЦЕНИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

I ПОДАЦИ О КОМИСИЈИ
<ol style="list-style-type: none"> 1. Датум и орган који је именовao комисију 15. 10. 2020, Наставно-научно веће Природно-математичког факултета у Новом Саду 2. Састав комисије са назнаком имена и презимена сваког члана, звања, назива уже научне области за коју је изабран у звање, датума избора у звање и назив факултета, установе у којој је члан комисије запослен: <ol style="list-style-type: none"> 1. Др Мирјана Ивановић, редовни професор, рачунарске науке, 29. 4. 2002, Природно-математички факултет у Новом Саду, председник комисије 2. Др Милош Радовановић, ванредни професор, рачунарске науке, 12. 12. 2016, Природно-математички факултет у Новом Саду, члан комисије 3. Др Мајкл Хоул (Michael Houle), гостујући професор, рачунарске науке, 1. 2. 2004, Национални институт за информатику у Токију (National Institute of Informatics, Tokyo), члан комисије 4. Др Золтан Гелер, доцент, информатика у друштвеним и хуманистичким профилима, 21. 4. 2016, Филозофски факултет у Новом Саду, члан комисије 5. Др Владимир Курбалија, ванредни професор, рачунарске науке, 1. 2. 2020, Природно-математички факултет у Новом Саду, ментор
II ПОДАЦИ О КАНДИДАТУ
<ol style="list-style-type: none"> 1. Име, име једног родитеља, презиме: Бранкица (Зоран) Братић 2. Датум рођења, општина, држава: 12. 12. 1989, Кикинда, Србија 3. Назив факултета, назив студијског програма дипломских академских студија – мастер и стечени стручни назив Природно-математички факултет у Новом Саду, информатика – информационе технологије, мастер информатичар 4. Година уписа на докторске студије и назив студијског програма докторских студија 2014, студијски програм: информатика 5. Назив факултета, назив магистарске тезе, научна област и датум одбране: нема
<ol style="list-style-type: none"> 6. Научна област из које је стечено академско звање магистра наука: нема
III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:
<p>Approximation algorithms for k-NN graph construction (Апроксимативни алгоритми за генерисање k-NN графа)</p>

IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Докторска дисертација је написана на енглеском језику и има 165 страна. Дисертација је подељена у 7 поглавља: Introduction (Увод), Background (Преглед литературе), Existing algorithms for k -NN graph construction (Постојећи алгоритми за генерисање k -NN графа), NN -Descent on high dimensional data (NN -Descent над високодимензионалним подацима), Proposed methods for improving NN -Descent (Предложени методи за побољшавање NN -Descent алгоритма), NN -Descent-based approximation algorithms for k -NN graph updates (Апроксимативни алгоритми, базирани на NN -Descent'-у, за ажурирање k -NN графова), Conclusions (Закључак). Поред тога, у дисертацији се налазе и 2 апендикса: Console application for algorithms related to k -NN graph, Results of the experiments related to k -NN graph update algorithms. Рад садржи 26 слика, 18 табела и 61 литературних навода. На почетку дисертације се налазе апстракт на енглеском и српском, као и предговор, док се на крају тезе налазе: проширени извод на српском, литература, листа слика, листа табела, листа коришћених скраћеница и кратка биографија кандидата.

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Прво поглавље представља увод. У његовој првој секцији је дат преглед примена k -NN графа, које уједно представљају мотивацију за истраживања представљена у тези. У другој секцији је дат кратак преглед главних резултата тезе.

У другом поглављу кандидат уводи основне концепте и дефиниције које се користе у дисертацији. Кандидат ово поглавље даље дели на четири одељка. У првом одељку су дате дефиниције општих концепата, попут графа, временских серија, метричких простора, итд. У другом одељку је дата дефиниција k -NN графа и уведене су мере које служе за евалуацију његових апроксимација. Трећи одељак је посвећен феномену који се зове хабнес (hubness), а који је такође врло значајан за резултате дисертације. На крају поглавља, унутар четвртог одељка, описани су проблеми и примене k -NN графова.

Треће поглавље садржи преглед ранијих резултата из литературе на које се истраживање из дисертације непосредно надовезује. У оквиру прве секције овог поглавља, представљени су разни алгоритми за генерисање k -NN графа. Сви они оптимизују генерисање ових графова, што се постиже на један од следећа три начина: 1) генерисањем k -NN графова тачно одређених особина, 2) паралелизацијом постојећих алгоритама за генерисање k -NN графова и 3) развојем апроксимативних алгоритама за генерисање k -NN графова. Како је фокус ове дисертације управо трећа ставка, тј. апроксимативни алгоритми за генерисање k -NN графова, други одељак овог поглавља је посвећен управо једном таквом алгоритму, који се зове NN -Descent. Унутар овог одељка детаљно је представљен поменути алгоритам, заједно са његовим главним манама, од којих је једна посебно од интереса за ову тезу: NN -Descent алгоритам не даје тачне апроксимације над високодимензионалним подацима. Најзад, последњи одељак је посвећен темпоралним k -NN графовима, који представљају k -NN графове који су генерисани над подацима који се мењају у протоку времена.

Четврто, пето и шесто поглавље су у потпуности посвећени оригиналним резултатима дисертације.

Унутар четвртог поглавља, објашњено је зашто NN -Descent алгоритам не генерише тачне апроксимације графа онда када је пуштен над високодимензионалним подацима. Испоставља се да је главни разлог за овакво понашање алгоритма феномен звани хабнес, који се генерално јавља у високодимензионалним подацима и који негативно утиче на многе алгоритме машинског учења.

У петом поглављу је представљено пет нових модификација NN -Descent алгоритма, које за циљ имају ублажавање поменутог проблема, тј. побољшавање рада овог алгоритма над високодимензионалним подацима. Прва модификација је заснована на чињеници да чворови високих улазних степена имају добре апроксимације најближих суседа, док су главни проблем заправо чворови ниских улазних степена, чије су апроксимације суседства врло лоше. У складу са тим, алгоритам за циљ има да усмери рачунарске ресурсе управо на одређивање суседа чворова ниских улазних степена. Друга модификација узима у обзир чињеницу да је тачност NN -Descent алгоритма већа за веће k вредности. Идеја овог приступа је да се генерише k -NN граф за већу k вредност, а да се потом редукује на жељено k . Трећа и четврта модификација омогућавају да се

количина утрошених рачунарских ресурса подешава на нивоу појединачних чворова, што онда може даље да се искористи тако да се више ресурса додели чворовима који имају лоша суседства. Пета, а уједно и последња, модификација, заснована је на чињеници да позиција чвора у иницијалном случајном графу игра велику улогу у тачности финалне апроксимације његовог суседства. Стога ова модификација спроводи додатна насумична поређења тачки које су у погрешном суседству у оквиру иницијалног случајног графа. Све модификације основног алгоритма су евалуиране експерименталном анализом над два синтетичка и четири реална скупа података. Резултати ове анализе сугеришу да сваки од предложених алгорита повећава тачност финалне апроксимације алгорита над подацима високе димензионалности.

Унутар шестог поглавља представљени су алгоритми за ажурирање k -NN графа. Наиме, подаци често имају тенденцију да се мењају временом, због чега постоји потреба за алгоритмима који би ефикасно ажурирали k -NN граф након што се подскуп његових чворова измени. Унутар дисертације, представљена су два апроксимативна алгорита за ажурирање k -NN графа. Унутар оба алгорита се спроводе кратке шетње које почињу од чворова који су измењени. Потом се рачунају раздаљине између почетних и крајњих чворова ових шетњи, а на основу тих раздаљина се онда и k -NN граф ажурира. Предложени алгоритми су евалуирани обимним експериментима над временским серијама. Експерименти заправо представљају симулације реалног сценарија, у ком временске серије добијају нове вредности, и на тај начин се мењају.

У последњем, седмом, поглављу су сумирани сви резултати и доприноси ове дисертације, а поред тога су описани и правци за даљи рад.

Према правилима Универзитета, текст дисертације тестиран је на плагијаризам коришћењем софтвера iThenticate. Софтвер је пријавио укупан индекс сличности од 15%. Притом је 10% преклапања текста дисертације везано за текст рада:

Brankica Bratić, Michael E. Houle, Vladimir Kurbalija, Vincent Oria, Miloš Radovanović. "The Influence of Hubness on NN-Descent", International Journal on Artificial Intelligence Tools, 2019

(што је очекивано, будући да је ово кандидаткињин коауторски рад на основу ког је стекла услове за предају дисертације), док је за све остале изворе преклапање мање од 1%.

VI СПИСАК НАУЧНИХ И СТРУЧНИХ РАДОВА КОЈИ СУ ОБЈАВЉЕНИ ИЛИ ПРИХВАЋЕНИ ЗА ОБЈАВЉИВАЊЕ НА ОСНОВУ РЕЗУЛТАТА ИСТРАЖИВАЊА У ОКВИРУ РАДА НА ДОКТОРСКОЈ ДИСЕРТАЦИЈИ

1. (M23) Brankica Bratić, Michael E. Houle, Vladimir Kurbalija, Vincent Oria, Miloš Radovanović, The Influence of Hubness on NN-Descent, *International Journal on Artificial Intelligence Tools*, 28(06), 2019, doi: 10.1142/S0218213019600029
2. (M33) Brankica Bratić, Michael E. Houle, Vladimir Kurbalija, Vincent Oria, Miloš Radovanović, NN-Descent on High-Dimensional Data, *In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, 2018, doi: 10.1145/3227609.3227643

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Први допринос ове дисертације јесте откривање разлога због којих *NN-Descent* алгорита не даје добре резултате над тачкама простора великих димензија. До сада је у литератури навођено да је разлог за то општепознати феномен машинског учења назван клетвом димензионалности (*curse of dimensionality*), но објашњење које је дубље од тога није постојало. Унутар дисертације објашњено је да је феномен звани хабнес заправо главни разлог за лош рад овог алгорита над високодимензионалним подацима. Све описано је и верификовано опсежним експерименталним анализама.

Други допринос истраживања односи се на превазилажење поменутог проблема *NN-Descent* алгорита. Наиме, предложено је пет модификација алгорита које га чине погоднијим за високодимензионалне податке. Све предложене модификације алгорита су евалуиране над шест високодимензионалних скупова података (четири реална и два синтетичка). Крајњи закључак ових експеримената јесте да предложене варијанте *NN-Descent* алгорита заиста увећавају тачност финалне апроксимације.

Унутар последње целине тезе, предложена су два апроксимативна алгоритма за ажурирање k -NN графа након што се измене подаци над којима је граф генерисан. У литератури ова тема није довољно истражена – велики број радова се бави генерисањем, али врло мали број се бави ажурирањем k -NN графова. С обзиром на то да се подаци врло често мењају у току времена, предложени алгоритми имају велику употребну вредност. Унутар експеримената, нови алгоритми су упоређени са методом „грубе силе“ (brute-force) и са NN -Descent алгоритмом, и показало се да се постижу значајна убрзања.

VIII ОЦЕНА НАЧИНА ПРИКАЗА И ТУМАЧЕЊА РЕЗУЛТАТА ИСТРАЖИВАЊА

Докторска дисертација је написана у складу са свим принципима писања овакве врсте рада. Веома јасно и прегледно су представљени сви проблеми, алгоритми и анализе. Добијени резултати су такође изложени врло јасно и систематично. Веза између резултата из дисертације и резултата истраживања других аутора из ове тематике је јасно успостављена.

IX КОНАЧНА ОЦЕНА ДОКТОРСKE ДИСЕРТАЦИЈЕ:

1. Да ли је дисертација написана у складу са образложењем наведеним у пријави теме

Дисертација је у потпуности написана у складу са образложењем наведеним у пријави теме.

2. Да ли дисертација садржи све битне елементе

Дисертација је написана у складу са свим принципима научног рада и садржи све битне елементе. На почетку су дате дефиниције свих појмова који се користе у дисертацији. Потом је дат преглед литературе и описани су досадашњи резултати. Након тога следе поглавља која садрже у потпуности оригиналне резултате. У оквиру њих су проблеми јасно представљени, а сви експерименти и њихови параметри прецизно описани. На крају су, унутар закључка, сумирани сви резултати тезе, и предложени су правци за будућа истраживања. Поред тога, програмски код експеримената је такође објављен и јавно доступан, а унутар додатака тезе је дато упутство за његово коришћење, те се сви експерименти из тезе могу лако репродуковати.

3. По чему је дисертација оригиналан допринос науци

Иако NN -Descent алгоритам ради врло добро у већини случајева, над високодимензионалним подацима производи апроксимације које су веома лоше. Аутори тог алгоритма су истакли да је клетва димензионалности разлог за такво понашање, али нико није дао детаљније и дубље објашњење. Унутар ове дисертације и пропратних кандидатских објављених радова, први пут су објашњени разлози за лоше функционисање овог познатог алгоритма.

Додатно, врло мали број радова се бави унапређивањем NN -Descent алгоритма над високодимензионалним подацима. Унутар ове дисертације, предложено је пет модификација овог алгоритма, који побољшавају његове перформансе у описаним условима.

Последњи битан допринос тиче се алгоритама за ажурирање k -NN графова. Наиме, постоји веома мали број истраживања који се баве том темом. Међутим, подаци из разних домена имају тенденцију да се мењају временом, те уколико је k -NN граф генерисан над таквим подацима, потребно га је ажурирати сваки пут кад се подаци измене. Уколико се измене догађају често, алгоритам је такође потребно често извршавати, те је у том случају брзина извршавања веома важан фактор. У оквиру ове дисертације предложена су два апроксимативна алгоритма за ажурирање k -NN графова. На основу експерименталних анализа, утврђено је да предложени алгоритми многоструко убрзавају процес ажурирања графова, што је свакако битан научни допринос.

4. Недостаци дисертације и њихов утицај на резултат истраживања

Дисертација нема значајних недостатака.

X ПРЕДЛОГ:
На основу укупне оцене дисертације, комисија предлаже:
- да се докторска дисертација Approximation algorithms for k-NN graph construction (Апроксимативни алгоритми за генерисање k -NN графа) прихвати, а кандидату Бранкици Братић одобри јавна одбрана.

ПОТПИСИ ЧЛАНОВА КОМИСИЈЕ

Др Мирјана Ивановић,
ред. проф. ПМФ-а,
председник комисије

Др Милош Радовановић,
ванр. проф. ПМФ-а,
члан комисије

Др Мајкл Хоул,
гостујући проф. Националног института информатике у Токију,
члан комисије

Др Золтан Гелер,
доцент Филозофског факултета у Новом Саду,
члан комисије

Др Владимир Курбалија,
ванр. проф. ПМФ-а,
ментор

НАПОМЕНА: Члан комисије који не жели да потпише извештај јер се не слаже са мишљењем већине чланова комисије, дужан је да унесе у извештај образложење односно разлоге због којих не жели да потпише извештај.

**UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES**

**REPORT ON THE ASSESSMENT OF DOCTORAL DISSERTATION OF
BRANKICA BRATIĆ**

I INFORMATION ABOUT THE DEFENSE BOARD
<p>1. Date and body that appointed the Defense Board: 15. 10. 2020, Teaching and Science Council of Faculty of Sciences, Faculty of Sciences, Novi Sad</p> <p>2. The members of the Defense Board with their names and surnames, titles, scientific fields, dates of obtaining the title, and the institution where the member is currently employed.</p> <p>1. Mirjana Ivanović, Ph.D., full professor, computer science, 29. 4. 2002, Faculty of Sciences, Novi Sad, president</p> <p>2. Miloš Radovanović, Ph.D., associate professor, computer science, 12. 12. 2016., Faculty of Sciences, Novi Sad, member</p> <p>3. Michael Houle, Ph.D., visiting professor, computer science, 1. 2. 2004, National Institute of Informatics, Tokyo, member</p> <p>4. Zoltan Geler, Ph.D., assistant professor, informatics in social and humanistic profiles, 21. 4. 2016, Faculty of Philosophy, Novi Sad, member</p> <p>5. Vladimir Kurbalija, Ph.D., associate professor, computer science, 1. 2. 2020, Faculty of Sciences, Novi Sad, advisor</p>
II INFORMATION ABOUT THE CANDIDATE
<p>1. Name, parent's name, surname: Brankica (Zoran) Bratić</p> <p>2. Date of birth, municipality, country: 12. 12. 1989, Kikinda, Serbia</p> <p>3. Name of faculty, title of study program of graduate academic studies - master and professional title Faculty of Sciences, University of Novi Sad, informatics – information technology, master in informatics</p> <p>4. Year of enrolment in doctoral studies and the title of the doctoral study program 2014, informatics</p> <p>5. Name of faculty, title of final work, scientific field and date of defense: none</p> <p>6. Scientific area from which the academic title is obtained: none</p>
III TITLE OF THE DOCTORAL DISSERTATION:
Approximation algorithms for k -NN graph construction (Апроксимативни алгоритми за генерисање k -NN графа)

IV OVERVIEW OF THE DOCTORAL DISSERTATION:

The thesis is written in English, and has 165 pages in total. It is divided in 7 chapters: Introduction, Background, Existing algorithms for k -NN graph construction, *NN-Descent* on high dimensional data, Proposed methods for improving *NN-Descent*, *NN-Descent*-based approximation algorithms for k -NN graph updates, Conclusions. In addition to that, the thesis has 2 appendices: Console application for algorithms related to k -NN graphs, and Detailed results of the experiments related to k -NN graph update algorithms. There are 26 figures, 18 tables and 61 literature entries. The thesis begins with abstract in English and Serbian, followed by preface, while it ends with extended abstract in Serbian, literature, list of figures, list of tables, abbreviations and short biography of the candidate.

V EVALUATION OF THE INDIVIDUAL PARTS OF THE DOCTORAL DISSERTATION:

The first chapter is the introductory chapter. Its first section gives an overview of the k -NN graph applications, which motivate the research presented in the thesis. In the second section, a brief overview of the main results is given.

The second chapter gives an overview of the basic concepts and definitions that are used in the thesis. This chapter is further divided in four sections. In the first section the definitions of the well-known, general concepts are given, e.g. definitions of graph, time series, metric spaces, etc. The definitions of k -NN graph and the measures for evaluation of k -NN graph approximations are given in the second section. The third section is dedicated to the phenomenon called *hubness*, which is very important for the results of this thesis.

The third chapter is dedicated to the existing results that are relevant to this thesis. Inside the first section of this chapter, various algorithms for k -NN graph construction are introduced. All of these algorithms optimize the construction of k -NN graphs, which can be achieved in one of the following three ways: 1) by constructing k -NN graphs with certain properties that can be used for optimization, 2) by parallelizing existing algorithms for k -NN graph construction, and 3) by developing approximation algorithms for k -NN graph construction. Since the focus of this thesis is the third way of optimization, i.e. approximation algorithms for k -NN graph construction, the second section of this chapter is dedicated to one such algorithm, *NN-Descent*. Inside that section, the algorithm is thoroughly presented, together with its main flaws, one of which is especially interesting for this thesis: *NN-Descent* does not construct correct approximation on high-dimensional datasets. Finally, the last section is focused on temporal k -NN graphs, which are k -NN graphs that are constructed on data that change over time.

The fourth, fifth and sixth chapters are dedicated to the original results of this thesis.

Inside the fourth chapter the candidate explained why *NN-Descent* algorithm does not construct correct graph approximations on high-dimensional datasets. It turned out that the main reason for such a behavior is the phenomenon called *hubness*, which is generally present in high-dimensional datasets, and which negatively influences many machine learning algorithms.

Five new *NN-Descent* variants are proposed in the fifth chapter. These variants are designed to alleviate the described *NN-Descent*'s flaw, i.e. they aim to improve the performance of this algorithm on the high-dimensional datasets. The first variant is based on the fact that nodes with high in-degrees have good approximations of their neighborhoods, while the nodes with low in-degrees are actually the major problem, since their neighborhood approximations are very incorrect. In respect with that, the algorithm tries to use more computational power for calculation of the neighborhoods of low in-degree nodes. The second *NN-Descent*'s variant takes into account the fact that the accuracy of the algorithm is higher for higher k values. The idea is to construct k -NN graph for higher k , and then to reduce it to the desired value of k . Third and fourth variants enable configuration of the spent computational power on the level of single nodes, which can then be used to assign more resources to the nodes with incorrect neighborhoods. The last, fifth, variant makes use of the fact that a node's position in an initial random graph highly affects the accuracy of the node's neighborhood in the final k -NN graph approximation. Therefore, this modification conducts additional random comparisons with the nodes that are in the wrong neighborhoods in the initial random graph. All of these variants were evaluated by the experimental analysis upon two synthetic and four real datasets. The results of this analysis suggest that each of the proposed *NN-Descent* variants increases the accuracy of the final k -NN graph approximation, when the underlying data is highly dimensional.

The sixth chapter deals with algorithms for k -NN graph updates. Namely, data often tend to change over time, raising the need for algorithms that would efficiently update k -NN graph after its underlying data have changed. In this thesis, candidate proposed two approximation algorithms for k -NN graph update. Both algorithms are based on short walks that start from the changed nodes. The distances between start and end nodes of these walks are calculated, and the k -NN graph is updated accordingly. Proposed algorithms are evaluated with extensive experiments on time series data. Experiments are based on simulations of a real-world scenario in which time series change by getting new values.

In the last, seventh, chapter, all the results and contributions of this thesis are summarized and the directions for future work are proposed.

According to the rules of University of Novi Sad, the thesis' text is verified against plagiarism with the software iThenticate. The software reported overall similarity index of 15%. It detected 10% overlap with the paper:

Brankica Bratić, Michael E. Houle, Vladimir Kurbalija, Vincent Oria, Miloš Radovanović. "The Influence of Hubness on NN-Descent", International Journal on Artificial Intelligence Tools, 2019

(which is expected since that paper is authored by the candidate of the thesis), while for all the other papers the reported overlap was less than 1%.

VI LIST OF PUBLISHED (OR ACCEPTED FOR PUBLICATION) SCIENTIFIC PAPERS THAT ARE BASED ON RESEARCH RESULTS OF THE DOCTORAL DISERTATION

1. Brankica Bratić, Michael E. Houle, Vladimir Kurbalija, Vincent Oria, Miloš Radovanović, The Influence of Hubness on NN-Descent, *International Journal on Artificial Intelligence Tools*, 28(06), 2019, doi: 10.1142/S0218213019600029
2. Brankica Bratić, Michael E. Houle, Vladimir Kurbalija, Vincent Oria, Miloš Radovanović, NN-Descent on High-Dimensional Data, *In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, 2018, doi: 10.1145/3227609.3227643

VII CONCLUSIONS AND SPECIFIC RESULTS OF THE RESEARCH

The first contribution of this thesis is analysis which revealed the reasons for misbehavior of *NN-Descent* on high-dimensional datasets. It has been suggested in the literature that the reason for this *NN-Descent* misbehavior is a well-known machine learning phenomenon called curse of dimensionality, however, the deeper explanations have not existed. In this thesis it is stated that the main reason for the misbehavior of *NN-Descent* is a phenomenon known as 'hubness'. This statement is verified with experimental analyses.

The second contribution of the research is related to the alleviation of the abovementioned flaws in *NN-Descent*. Namely, five new *NN-Descent*'s variants are proposed in the thesis. These five variants make *NN-Descent* more suitable for high-dimensional datasets. All the variants are evaluated on six high-dimensional datasets (four real and two synthetic). The conclusion of these experiments is that all the *NN-Descent*'s variants indeed increase the accuracy of the resulting approximations.

In the last part of the thesis, two approximation algorithms for k -NN graph update are proposed – these algorithms update a k -NN graph after its underlying data change. The problem of k -NN graph update is not properly covered in the literature – the majority of papers are dedicated to construction, but very few of them to the update of k -NN graphs. Given that data very often change over time, the proposed algorithms are greatly beneficial. These algorithms are experimentally evaluated and compared to the brute force approach and to the *NN-Descent* algorithm. The experiments suggest that these algorithms are capable of major speedup.

VIII ASSESSMENT OF RESEARCH RESULTS PRESENTATION AND INTERPRETATION QUALITY

The doctoral thesis is written in accordance with all the principles that hold for this kind of scientific artefact. All the problems, algorithms and analyses are clearly and systematically presented. Obtained results are also presented very clearly and thoroughly. Connection between the results of this thesis and results of the other authors' research is well established.

IX FINAL ASSESSMENT OF THE DOCTORAL DISERTATION:

1. Is the dissertation written in accordance with the statements given in the dissertation proposal?

The thesis is completely written in accordance with the statements given in the dissertation proposal.

2. Does the dissertation contain all essential elements?

The thesis is written in accordance with all principles of a scientific artefact, and it contains all essential elements. The definitions of all relevant concepts are given at the beginning of the thesis. After that, the overview of the literature is given, and the results in the field are presented. This is followed by the chapters that contain original results. Inside these chapters, the problems are well described, and all the experiments, together with their parameters, are precisely presented. At the end of the thesis, in the conclusions, all the results are summarized, and directions for future work are given. Additionally, the code of all the experiments is released and publicly available, and its usage specifications are given inside a thesis's appendix, making all the experiments easily reproducible.

3. What is the dissertation's original contribution to science?

Even that *NN-Descent* algorithm behaves pretty well in most of the cases, on high-dimensional datasets it produces very inaccurate approximations. The authors of the algorithm pointed out that the reason for that misbehavior is related to the curse of dimensionality, however, nobody has given a more detailed explanation. Inside this thesis and inside the candidate's accompanying papers, the reasons for the *NN-Descent*'s poor behavior on high-dimensional data is explained for the first time.

Additionally, exceedingly small number of papers deal with improvements of *NN-Descent* on high-dimensional data. In this thesis five modifications of *NN-Descent* are proposed, and all of them improve its performance in the mentioned settings.

The last important contribution is related to the algorithms for k -NN graph update. Namely, there has not been much research that deals with this problem. However, data, regardless of the domain, tend to change over time, so if a k -NN graph was built upon such data, it would require updating each time the data changed. If the changes happen very often, the algorithm that updates the graph must also be executed very often. Therefore, it might be very important that one such algorithm is fast. Inside this thesis, two approximation algorithms for k -NN graph update are proposed. Experimental analyses suggest that the proposed algorithms speed up the process of k -NN graph update, which is definitely an important contribution to science.

4. Disadvantages of the dissertation and their impact on the research results

The thesis does not have any major disadvantage.

X THE PROPOSAL:
Based on the overall dissertation assessment, the commission proposes:
- that the thesis Approximation algorithms for k-NN graph construction (Апроксимативни алгоритми за генерисање k-NN графа) should be accepted, and that the defense should be approved to the candidate Brankica Bratić .

THE SIGNATURES OF THE MEMBERS OF THE DEFENSE BOARD

Mirjana Ivanović, Ph.D.
full professor, Faculty of Sciences, Novi Sad
president

Miloš Radovanović, Ph.D.
associate professor, Faculty of Sciences, Novi Sad
member

Michael Houle, Ph.D.
visiting professor, National Institute of Informatics, Tokyo,
member

Zoltan Geler, Ph.D.
assistant professor, Faculty of Philosophy, Novi Sad,
member

Vladimir Kurbalija, Ph.D.
associate professor, Faculty of Sciences, Novi Sad,
advisor